Using Multiple Imputation of Latent Classes to construct population census tables with data from multiple sources

Laura Boeschoten, Sander Scholtus, Jacco Daalmans, Jeroen K. Vermunt and Ton de Waal[1]

Abstract

The Multiple Imputation of Latent Classes (MILC) method combines multiple imputation and latent class analysis to correct for misclassification in combined datasets. Furthermore, MILC generates a multiply imputed dataset which can be used to estimate different statistics in a straightforward manner, ensuring that uncertainty due to misclassification is incorporated when estimating the total variance. In this paper, it is investigated how the MILC method can be adjusted to be applied for census purposes. More specifically, it is investigated how the MILC method deals with a finite and complete population register, how the MILC method can simultaneously correct misclassification in multiple latent variables and how multiple edit restrictions can be incorporated. A simulation study shows that the MILC method is in general able to reproduce cell frequencies in both low- and high-dimensional tables with low amounts of bias. In addition, variance can also be estimated appropriately, although variance is overestimated when cell frequencies are small.

Key Words: Combined survey-register data; Population census; Misclassification; Multiple imputation; Latent Class analysis.

1. Introduction

Official Statistics are increasingly often compiled from a combination of data sources, including surveys and administrative registers. The use of different sources poses multiple challenges. Different sources can be overlapping, meaning that more than one observation is obtained for the same person and variable. Often, it is observed that data sources are contaminated by errors and missing values. Therefore it can happen that two data sources provide two different values for the same unit and variable. Most of the data collected by statistical agencies have to be corrected or processed somehow to obtain consistent and publishable results. Several strategies are available to deal with multiple, overlapping data sources that are each contaminated by erroneous and missing values, see e.g. Pankowska, Pavlopoulos, Bakker and Oberski (2020). A first, and in practice often chosen strategy, is to ignore inconsistencies between data sources. This happens for instance if one data source is chosen that is believed to have the highest quality (de Waal, van Delden and Scholtus, 2020). When such strategies are chosen, the information in all available sources is not fully exploited.

A second strategy is to apply weighting techniques (Särndal, Swensson and Wretman, 2003). When weighting is used, survey records are calibrated towards the totals from a register source. Differences between data sources are fully explained from the selection effects of the sample. This approach ignores

[1] Laura Boeschoten, Tilburg University Tilburg School of Social and Behavioral Sciences - Methodology and Statistics Warandelaan 2 Tilburg, Tilburg, Noord-Brabant 5000 LE Netherlands and Centraal Bureau voor de Statistiek - Methodology Henri Faasdreef 312, Den Haag 2490 HA Netherlands. E-mail: lauraboeschoten@gmail.com; Sander Scholtus, Centraal Bureau voor de Statistiek - Methodology Den Haag, Zuid-Holland Netherlands; Jacco Daalmans, Centraal Bureau voor de Statistiek - Methodology Den Haag, Zuid-Holland Netherlands; Jeroen K. Vermunt, Tilburg University Tilburg School of Social and Behavioral Sciences - Methodology & Statistics Tilburg, Noord-Brabant Netherlands; Ton De Waal, Centraal Bureau voor de Statistiek - Methodology Den Haag, Zuid-Holland Netherlands and Tilburg University Tilburg School of Social and Behavioral Sciences - Methodology & Statistics Tilburg, Noord-Brabant Netherlands.

the fact that the register totals, as well as the sample surveys, might be subject to measurement error. An additional complication is that weighting does not always lead to fully consistent output, as it only achieves consistency with regard to the variables that are incorporated in the weighting model. The number of variables that can be included in a weighting model is however limited.

A third strategy to resolve inconsistencies between multiple sources is macro-integration, an approach that reconciles statistical output at aggregate level. This approach usually consists of two steps. First, differences with a known cause are resolved (i.e. bias). The remaining, mostly smaller, discrepancies that usually arise due to noise are corrected in a second step. Several mathematical methods have been developed for this purpose, e.g. Bikker, Daalmans and Mushkudiani (2013), Daalmans (2019), Di Fonzo and Martini (2003), Magnus, van Tongeren and de Vos (2000), Sefton and Weale (1995) and Stone, Champernowne and Meade (1942). A first drawback of macro integration is that the connection between the micro-data and the published results gets lost. The macro-integrated results cannot be computed by aggregation of the micro data. A second drawback is that the detailed micro data might not be fully exploited, as the corrections are made at the macro level.

Many of the issues arising when one of the previously discussed strategies is used can be circumvented by Multiple Imputation of Latent Class analysis (MILC) by Boeschoten, Oberski and de Waal (2017). This method combines multiple measures from different sources (population register and sample survey) at micro level. The different observations are considered indicators of a Latent Class (LC) model. The MILC-model corrects for misclassification while also taking edit restrictions into account. These are rules that identify logically impossible combinations of scores (e.g. pregnant men). After the LC model has been estimated, multiple imputed versions of the target variable are created, that are corrected for the estimated misclassification. Differences between imputed values reflect the uncertainty due to missing and conflicting values. The total variance can be estimated based on these differences. The method can be considered a model-based imputation method that requires the Missing At Random (MAR) assumption. A simulation study on the performance of this method showed that its performance is strongly related to the entropy R^2 value of the LC model; a measure which indicates how well the LC model can predict class membership based on the observed variables, or how well classes are separated.

After MILC was introduced, multiple studies have extended the method to broaden its scope of applicability. Boeschoten, de Waal and Vermunt (2019) extended the method to impute values that are missing by design, for example because they were not present in the sample, using a quasi-latent variable. More specifically, a quasi-latent variable is a latent variable that is restricted to have a perfect relationship with an observed variable that contains missing values. In that way, the relationship between the quasi-latent variable and all other variables specified in the model can be used to estimate the missing values. In addition, they investigated the performance of the method when two combined sources follow different missingness mechanisms. Furthermore, Boeschoten, Filipponi and Varriale (2021) investigated how the method can be extended for longitudinal situations and how unit missingness can be imputed in a situation of combined survey and register data.

Although these previous studies investigated a number of relevant issues, there are still cases for which it is unclear how the MILC-method can be applied. The aim of this paper is to further enhance the possibilities of MILC in terms of application and, with that, to further increase the capabilities of producing multi-source statistics.

Currently, the application of MILC has been limited to univariate problems. In practice, however, there is often a need to estimate multiple variables at once. The first important extension in this paper is to allow the simultaneous imputation of multiple latent variables. As population registers can contain misclassification, it is worthwhile to correct for the misclassification if possible. For multivariate problems, corrections should be performed simultaneously, which is more difficult than for one variable only.

Second, statistical agencies generally consider finite target populations (e.g. containing all registered inhabitants of a country). It is unclear if the MILC method can be applied directly to a finite population, or that adaptations to the method should be made.

The usefulness of the extensions in this paper is illustrated by an application to the Dutch virtual census; an application that would otherwise not be possible. For the census, a large number of tables have to be estimated from a population register and a sample survey. To the best of our knowledge, this is the first time that MILC has been applied to such a large estimation problem. Theoretically, it is already known that edit restrictions can be incorporated in an LC model to prevent the occurrence of logically impossible combinations of scores (Boeschoten et al., 2017). However, it is not trivial how the MILC method performs if edit restrictions are incorporated in such a way that they affect multiple cells in a population census table.

In Section 2, a description of the MILC method is given, tailored to handle the specific extensions discussed. In Section 3, a description of the simulation study is given. Simulation results are shown in Section 4 and Section 5 provides a discussion.

2. Methodology

When applying the MILC method, the starting point is a unit-linked combined dataset, which can consists of combinations of administrative population registries and survey samples. In order to account for uncertainty regarding the parameters of the LC model estimated at a later step in MILC, a non-parametric bootstrap procedure is applied on this dataset first (step 1). This involves creating M bootstrap samples by drawing observations from the observed dataset with replacement. Subsequently, for each bootstrap sample, the LC model of interest is estimated (step 2) using Latent GOLD software (Vermunt and Magidson, 2013a). Here, model parameters are estimated by Maximum Likelihood using a combination of the Expectation-Maximization and Newton-Raphson algorithms. Note that here, by explicitly stating which cells should be restricted, constrained estimation is used. Next, M imputations are created using the M sets of parameter values obtained from the M latent class models (step 3). If imputations would be created based on the maximum-likelihood estimates obtained directly using the original observed data, sampling uncertainty regarding the estimated parameters of the latent class model would be ignored.

In the following subsections, we explain each of the steps of MILC in more detail and present the extension for the estimation of multiple latent variables for a finite population from register and sample survey data.

2.1 Step 1: Creating bootstrap samples

We propose to use the "classical" bootstrap procedure here, which consists of repeatedly drawing samples with replacement from the original dataset, of the same size as the original dataset. A motivation for using this classical with-replacement bootstrap here, as opposed to an adapted bootstrap procedure for a finite population, is provided in Section 2.5 below.

The bootstrap should be applied to the dataset that is used to estimate the LC models. When register data and survey data are combined, the indicator variables from the survey will typically be missing for a large part (e.g., 90% or more) of the population. The LC models could then be estimated by two different approaches:

- using only the subset of persons observed in both the survey and the register (complete cases);
- using all available data, including cases with missing indicators.

Under the second approach, full information maximum likelihood can be used to handle missing values when estimating the LC models. This has the advantage of using all available information. Since this amounts to estimating the LC model on M datasets with the size of the target population, a practical drawback of this approach is that it may be computationally demanding in terms of time and memory. Therefore, the first approach may be more attractive, in particular when the associations among the covariates and target variables are relatively weak. In the latter approach, the cases with missing survey data will contain relatively little information about the parameters of the LC model. Note that under both approaches, the estimated LC models are used to impute predictions of the latent classes throughout the population. Depending on which approach is chosen to estimate the LC models, bootstrapping is applied either to the subset of complete cases or to the target population. In the simulation study in this paper, the complete-case approach will be used.

2.2 Step 2: Estimating the latent class model

The second step performed is the estimation of the LC model. It is explained below how this is done for multiple latent variables. As described in the previous section, the LC model is typically estimated M times using the M bootstrapped datasets. In the situation under evaluation in this paper, the LC model is estimated M times on M subsets of complete observations coming from the M bootstrap samples. An extensive discussion of the model and the assumptions made when using the model to correct for measurement error can be found in Boeschoten et al. (2017). Multiple latent variables can be estimated simultaneously in one model, which yields the following model structure for the joint probability of the response variables given covariate values, denoted by $\mathrm{P}(\mathbf{Y}=\mathbf{y} \mid \mathbf{Q}=\mathbf{q})$. The number of latent variables is denoted as v and K_h is the number of classes of latent variable X_h (scalar), where $(h=1,\ldots,v)$. Furthermore, $\mathbf{Y}$ are the observed target variables, i.e. the indicator variables, L_h is the number of indicator variables for X_h and $\mathbf{Q}$ are the (also observed) covariate variables:

$$\begin{aligned}\mathrm{P}(\mathbf{Y}=\mathbf{y} \mid \mathbf{Q}=\mathbf{q}) = \sum_{x_1=1}^{K_1} \cdots \sum_{x_v=1}^{K_v} & \mathrm{P}(X_1=x_1,\ldots,X_v=x_v \mid \mathbf{Q}=\mathbf{q}) \\ & \prod_{l_1=1}^{L_1} \mathrm{P}(Y_{l_1,1}=y_{l_1,1} \mid X_1=x_1) \\ & \cdots \\ & \prod_{l_v=1}^{L_v} \mathrm{P}(Y_{l_v,v}=y_{l_v,v} \mid X_v=x_v). \end{aligned} \tag{2.1}$$

Here, local independence is assumed as well as independence of covariates.

Constrained parameter estimation is used when certain cells within $\mathrm{P}(X_1=x_1,\ldots,X_v=x_v \mid \mathbf{Q}=\mathbf{q})$ are restricted. This can be used to specify that certain combinations of scores between covariates and latent variables are logically impossible, or when a “quasi-latent” variable is used to create imputations for missing values in a variable (Vermunt and Magidson, 2013b).

2.3 Step 3: Multiple imputation

To be able to create multiple imputations, joint posterior membership probabilities are calculated for every person in the original dataset. They represent the probability that a unit is part of a combination of latent classes from the different latent variables, given its combination of scores on the indicators and covariates used in the LC model. These probabilities can be used to create multiple imputations of the latent variables which contain their “true scores”.

The joint posterior membership probabilities can be calculated by applying Bayes’ rule to the conditional response probabilities obtained from the M LC models:

$$\mathrm{P}(X_1=x_1,\ldots,X_v=x_v \mid \mathbf{Y}=\mathbf{y},\mathbf{Q}=\mathbf{q}) = \frac{\mathrm{P}(X_1=x_1,\ldots,X_v=x_v,\mathbf{Y}=\mathbf{y} \mid \mathbf{Q}=\mathbf{q})}{\mathrm{P}(\mathbf{Y}=\mathbf{y} \mid \mathbf{Q}=\mathbf{q})}, \tag{2.2}$$

where

$$\begin{aligned}\mathrm{P}(X_1=x_1,\ldots,X_v=x_v,\mathbf{Y}=\mathbf{y} \mid \mathbf{Q}=\mathbf{q}) = {} & \mathrm{P}(X_1=x_1,\ldots,X_v=x_v \mid \mathbf{Q}=\mathbf{q}) \\ & \prod_{l_1=1}^{L_1} \mathrm{P}(Y_{l_1,1}=y_{l_1,1} \mid X_1=x_1) \\ & \cdots \\ & \prod_{l_v=1}^{L_v} \mathrm{P}(Y_{l_v,v}=y_{l_v,v} \mid X_v=x_v) \end{aligned} \tag{2.3}$$

and $\mathrm{P}(\mathbf{Y}=\mathbf{y} \mid \mathbf{Q}=\mathbf{q})$ is defined in equation 2.1. For one profile (so one set of scores on all indicator and covariate variables), the joint posterior membership probabilities sum up to one.

To be able to include parameter uncertainty in our variance estimates, we perform the model estimation on M bootstrap samples of the dataset, resulting in M different LC models. We generate imputations in the original dataset accounting for the parameter uncertainty by using the resulting M sets of bootstrap parameter estimates. More specifically, with each of these M parameter sets we compute the posterior class membership probabilities for the original sample, and use these to generate the imputations. In other words, the M imputations are based of M different sets of posterior probabilities.

2.4 Step 4: Pooling

The next step is to obtain estimates of interest for every imputation, and to pool them using Rubin's Rules (Rubin, 1987, page 76). For this research, the main interest is producing a frequency table. Therefore, the frequency table of interest is obtained for the M imputations and they are pooled, which means taking the average over the imputations for every cell in the frequency table:

$$\hat{\theta}_j = \frac{1}{M}\sum_{i=1}^{M}\hat{\theta}_{ij}, \tag{2.4}$$

where j refers to a specific cell in the frequency table.

Next, an estimate of the uncertainty around these frequencies is of interest. In general, the variance of the pooled estimate j can be estimated by Rubin's total variance formula for multiple imputation (Rubin, 1987, page 76):

$$\text{VAR}_{\text{total}_j} = \overline{\text{VAR}}_{\text{within}_j} + \text{VAR}_{\text{between}_j} + \frac{\text{VAR}_{\text{between}_j}}{M}. \tag{2.5}$$

Here, $\text{VAR}_{\text{between}_j}$ can be estimated as

$$\text{VAR}_{\text{between}_j} = \frac{1}{M-1}\sum_{i=1}^{M}\left(\hat{\theta}_{ij}-\hat{\theta}_j\right)\left(\hat{\theta}_{ij}-\hat{\theta}_j\right)'. \tag{2.6}$$

The within variance $\text{VAR}_{\text{within}_j}$ reflects the average sampling variance of ij when the imputed values are treated as observed. In our application, as the population is finite and imputations are generated for the complete population, this within variance component is zero and can be mitigated (Vink and van Buuren, 2014). Note that this is a property of multiple imputation and is due to the fact that the complete population is imputed. This should not be confused with the decision to only use a sample for LC model estimation. Hence, formula (2.5) is reduced in this case to:

$$\text{VAR}_{\text{total}_j} = \text{VAR}_{\text{between}_j} + \frac{\text{VAR}_{\text{between}_j}}{M}. \tag{2.7}$$

2.5 A note on bootstrapping for multiple imputation in finite populations

The aim of a census is to estimate certain target parameters of a finite population (e.g., all persons currently living in the Netherlands). Hence, a natural idea might be to apply a finite-population bootstrap procedure in this context; see Mashreghi, Haziza and Léger (2016) for an overview of bootstrap methods for finite populations. However, when determining the appropriate bootstrap approach, it should be noted that the bootstrap in MILC is specifically implemented to account for the between imputation variance component of formula (2.5) in Section 2.4. In general, variability in the target parameters due to the fact that a sample was drawn from a finite population is incorporated in the within variance component of formula (2.5). As we use mass imputation here, the within variance component in fact reduces to zero; cf. formula (2.7). More generally, this component would be estimated separately from the bootstrap method at hand; see Boeschoten et al. (2017) for an example.

Furthermore, the reason for incorporating the bootstrap in the MILC approach is to account for uncertainty in the estimated parameters of the latent class model. Note that these parameters are not associated with a finite population, but with a model. Even if we had observed the entire finite population, there would still be uncertainty about the true parameter values of the latent class model. This uncertainty can be considered as drawing from an infinite distribution. Therefore, we select the classical with-replacement bootstrap. We argue that bootstrap methods for finite populations should not be used in this context. For large samples, such methods would result in a substantial underestimation of the variance when combined with the usual approach to multiple imputation. We also checked this empirically in the simulation study to be discussed in Section 3. As an example, when a pseudo-population bootstrap method for finite populations was used, the resulting se/sd ratios in Table 4.7 for the condition MAR, $M = 5$ were 0.7217, 0.7887, 0.7536 and 0.8607, respectively, all pointing to a non-negligible underestimation of the true variance.

In the simulation study in this paper, we will restrict attention to surveys based on simple random sampling and stratified simple random sampling. For more complex survey designs, e.g. involving cluster sampling or sampling with unequal probabilities, it is unclear whether the proposed bootstrap approach is always appropriate. It is possible that in some cases such complex design features could indirectly affect the uncertainty of estimated parameters of the latent class model and therefore become relevant for variance estimation. We will return to this point in the discussion section.

3. Simulation study

In this section, we describe a simulation study that is performed to evaluate the extensions of the MILC method in Section 2. The topic of this study is the estimation of a table from the Dutch Population and Housing Census.

3.1 The Dutch Census

Population and housing censuses provide a picture about the socio-demographic and socio-economic situation of a country and it is ubiquitous that a census should cover the entire population of people and dwellings that are present in a country. Every ten years the United Nations Economic and Social Council (ECOSOC) adopts a resolution, urging Member States to carry out a population and housing census and to disseminate census results as an essential source of information, see e.g. The Economic and Social Council (2005). In the EU, explicit agreements have been made about which variables should be listed in the census, and also which cross-tables should be produced (European Commission, 2008, 2009 and 2010).

The vast majority of countries produce census data by conducting a traditional census, which entails interviewing inhabitants in a complete enumeration, reaching every single household. An increasing number of countries however have adopted a different, innovative approach, in the form of a so-called virtual census. With a virtual census, census tables are compiled using data sources that are already available at the statistical agency. These are data sources that have not been primary collected for the census, but for other purposes. Statistics Netherlands can rely on population registers as the main source for most census tables. These registers are of relatively good quality, including a very broad coverage (Geerdinck, Goedhuys-van der Linden, Hoogbruin, De Rijk, Sluiter and Verkleij, 2014). All register variables are available from Statistics Netherlands' system of social statistical data-sets (Bakker, Van Rooijen and Van Toor, 2014). The backbone is the Central Population Register which combines the population registers from municipalities. The population registers are supplemented with variables originating from sample surveys, because not all variables that are necessary according to the EU regulations can be found in the population registers.

For the 2001 and 2011 Dutch censuses, only two variables could not be measured from registers: Occupation and Educational Attainment (Schulte Nordholt, Van Zeijl and Hoeksma, 2014). These two variables were observed from combined Labour Force Surveys (LFSs). To obtain the required cross-tables for the 2011 Dutch census, a procedure was used where all data sources were matched on the unit level. Then, a micro-integration process was carried out. Micro-integration brings together records from different micro-datasets and subsequently resolves data inconsistencies. The goal is to improve the quality, compatibility and scope of the data sets. The techniques that are used in micro integration are: completing, harmonising and correcting for measurement errors. Completing means that corrections are made for an under- or overcoverage of a target population. Harmonisation refers to transformations such that data sets fit to the concept that is supposed to be measured. Measurement correction means that inconsistencies between sources are resolved (Bakker, 2011; van Rooijen, Bloemendal and Krol, 2016). Also, inconsistencies between sources are removed, by using formal rules that make clear what happens in case of inconsistencies, e.g. which source is used (Bakker, 2010; de Waal, Pannekoek and Scholtus, 2011).

After micro-integration, two combined data sources were obtained: one based on a combination of registers and the other one based on a combination of sample surveys. All census tables that do not contain occupation and educational attainment were entirely compiled from the combined registers. The values in the cells of these tables were obtained by counting the occurrence of the categories in the matched registers. The other census tables, those with educational attainment and/or occupation, were estimated from the combined sample surveys. To establish consistent results, a procedure was applied based on weighting followed by macro integration (Daalmans, 2018; Schulte Nordholt et al., 2014). In the first step, weights were derived, such that the marginal totals of the weighted survey data comply with the known totals from the registers. The different tables that are obtained in this way are not necessarily consistent with each other, because different weighting schemes apply to each table. To resolve this problem, macro-integration is used. This step starts with initial estimates for each census table, derived from the weighted survey data or from the integrally counted register data. These initial estimates are adjusted, to arrive at fully consistent census tables, that comply with the known register totals.

MILC has a couple of advantages over the current estimation method. First, the assumption is often made that the population registers are free of error. If a variable is measured both in the population register and in a sample survey and the scores on these variables contradict each other, the register score usually overrides the survey score because of this assumption. In other words, sample survey data are ignored for the part that is also observed in a register. Second, for the current procedure, it is not easy to compute uncertainty measures that capture all steps of the estimation process, including the uncertainty due to the missing and conflicting values in the linked data-sets. For MILC on the other hand it is well-established how variances can be properly estimated. Third, the data processing procedure that is currently used contains a specific sequence of steps, where decisions made at one step are influenced by decisions made at previous steps. For instance, if there are two conflicting values for the same person, then one of these is chosen in the "micro-integration" step. In the subsequent weighting and macro integration steps only one value is used. Thus, the availability of the different values is ignored in the final estimation of the census tables. Basically, MILC exploits information provided by all observed values in contrast to the current procedure.

3.2 The census table under investigation

The starting point of this simulation study is an existing census table, which can be downloaded from Census Hub (Census Hub, 2017). This table comprises 2,691,477 persons who where living in the region "Noord-Holland" in the Netherlands in 2011. This census table is a cross-table between the following six variables:

1. Age in 21 categories: under 5 years; 5 to 9 years; 10 to 14 years; 15 to 19 years; 20 to 24 years; 25 to 29 years; 30 to 34 years; 35 to 39 years; 40 to 44 years; 45 to 49 years; 50 to 54 years; 55 to 59 years; 60 to 64 years; 65 to 69 years; 70 to 74 years; 75 to 79 years; 80 to 84 years; 85 to 89 years; 90 to 94 years; 95 to 99 years; 100 years and over.
2. Marital status in eight categories: never married; married; widowed; divorced; registered partnership; widow of registered partner; divorced from registered partner; not stated.
3. Gender in two categories: male; female.
4. Place of birth in five categories: the Netherlands; a country within the European Union; a country outside the European Union; other; not stated.
5. Type of family nucleus in which a person lives in five categories: partners; lone parents; sons/daughters; not stated; not applicable.

6. Country of citizenship in five categories: Dutch citizen; citizen of a country within the European Union; citizen of a country outside the European Union; stateless; not stated.

Thus, the census table consists of 42,000 cells.

3.3 Simulation setup

The goal of this simulation study is to replicate the frequencies of the 42,000 cells in the cross-table using multiple indicators contaminated with misclassification and missing values. Therefore, this misclassification should be induced first.

We generate two indicator variables for three different latent variables, all containing 5% random misclassification, which can be considered a very high amount, especially for Dutch population registers. The indicator variables are generated for the variables “Gender”, “Type of family nucleus” and “Country of citizenship”. Misclassification is generated in such a way that first, 5% of the cases are randomly selected. Second, their original score is identified and third, a different score is assigned by sampling from the observed frequency distribution of the other categories.

For the register indicators $Y_{l_v,1}$, misclassification is generated only once, as these indicator variables represent register variables for the complete and finite population, there should not be any variability in misclassification between replications in the simulation study for these variables. For the survey indicators $Y_{l_v,2}$, misclassification is newly generated for every replication in the simulation study, followed by generating missing values using either a Missing Completely At Random (MCAR) or Missing At Random (MAR) missingness mechanism with approximately 90% missingness for both situations. With a MCAR mechanism, the response probabilities for the respondents and non-respondents is equal. With a MAR mechanism, the response probabilities are related to other observed values (Rubin, 1976). These $Y_{l_v,2}$ indicators represent survey variables for a sample of the population.

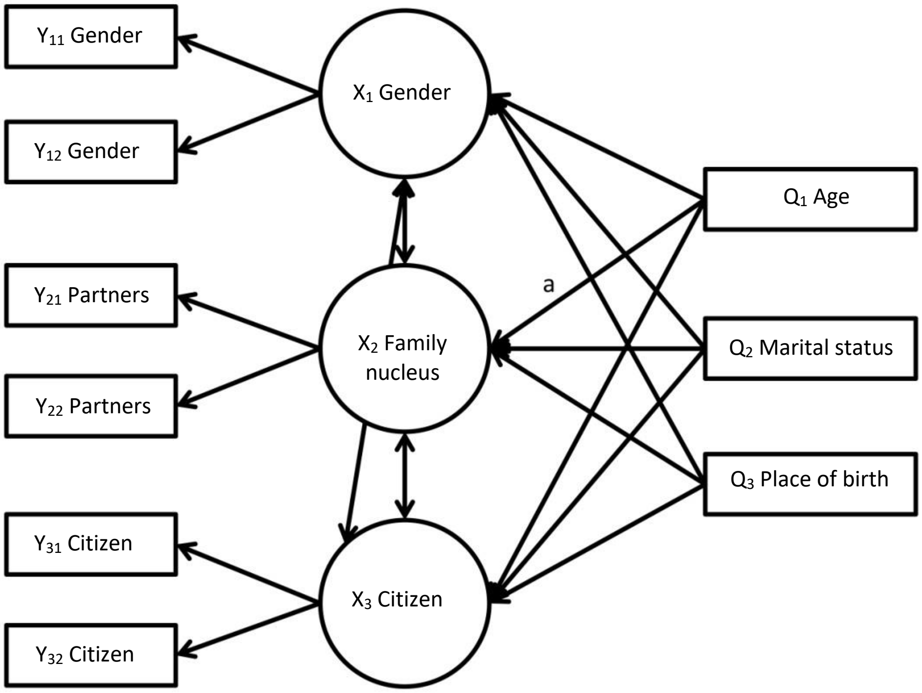
Missingness is generated in such a way that it mimics a situation that 10% of the population is included in the survey. Missingness is generated under MCAR and MAR. Under MCAR, the probability of being missing (i.e. not being included in the survey) is 90% and equal for every person in the population. Under MAR, the probability of being missing depends on a persons’ age and decreases as a person gets older. More specifically, the probability of being missing is lowest for persons in the age category “100 years and older”, and is 80%. This percentage gradually increases with the highest being 94% for the persons in the age category “under five years”. To summarize, for each of the 500 iterations in the simulation study, a simple random or stratified sample of the combined data-set is obtained that contains approximately 269,147 persons (10% of the population), on which the LC model is estimated.

3.4 Applying the MILC method

As discussed in Section 2, M bootstrap samples are generated from the combined dataset, and in this study the LC model is estimated only on the complete set of observations of each bootstrap sample. Results are obtained using $M = 5$, $M = 10$ and $M = 20$.

In Figure 3.1, the graphical overview of the latent class model can be found. Here, it can be seen that the latent variables X_1 "Gender", X_2 "Family nucleus" and X_3 "Citizen" are all measured by two indicators. The restriction on the relationship between Q_1 "Age" and X_2 "Family nucleus" is denoted by "a" in Figure 3.1. Here, we restricted that if someone is of age category "under 5 years", "5 to 9 years" or "10 to 14 years", it is impossible to be assigned to the latent classes "partners" or "lone parents" for the latent variable "Family nucleus".

Figure 3.1 Graphical overview of the LC model specified. Note that edit restrictions are applied between the variables "Type of family nucleus" and "age" (denoted in the model by "a").



To specify the LC model for response pattern $\mathrm{P}(\mathbf{Y}=\mathbf{y} \mid \mathbf{Q}=\mathbf{q})$ we can fill in at equation 2.1 that $v=3$, $K_1=2$, $K_2=4$, $K_3=4$, $L_1=2$, $L_2=2$ and $L_3=2$. Note that X_2 here only has four latent classes, while the variable "Family Nucleus" in the population census table has five categories. Therefore, it would have made sense for X_2 to also have five latent classes. However, there were no observations for the category "not applicable", so therefore we didn't have to include a latent class for this category. The same holds for the category "stateless" of X_3.

Next, multiple imputations can be created and estimates of interest can be pooled as described in Sections 2.3 and 2.4. As the cells of the frequency-tables of interest can become very small, a log-transformation is used to ensure appropriate confidence intervals around these small cells. Therefore, $\mathrm{VAR}_{\text{between}_j}$ is not estimated as the variance of $\hat{\theta}_{ij}$, as in equation 2.7, but as the variance of $\log(\hat{\theta}_{ij})$, where $\hat{\theta}_{ij}$ refers to the number of units in cell j in imputation i.

3.5 Evaluation

To evaluate the performance of the MILC method when trying to construct the census table initially used to create the misclassified data, it is useful to make comparisons to results obtained when the variable observed in the register is used directly to create cross-tables. We refer to these results as obtained using $Y_{v,1}$. These results are equal over the 500 simulation iterations and the bias here directly reflects the misclassification in this indicator, which becomes more severe as the categories are more imbalanced in size due to the misclassification mechanism. Furthermore, it would be difficult to draw general conclusions from results obtained by only evaluating every single of the 42,000 cells of the complete census table. Therefore, we investigate some specific characteristics of this table separately. First, we investigate whether the method is able to reconstruct the univariate marginal cell frequencies of the latent variables specified. Second, we investigate if the method is able to reconstruct the joint distribution of the three latent variables. Third, we investigate if the method correctly incorporates edit restrictions. At last, we investigate some features of the complete census table.

First, we evaluate the cell-proportions of the previously discussed cross-tables in terms of bias, by evaluating the average absolute bias and the root mean squared error (RMSE) over the $N_{it} = 500$ replications in the simulation study. More specifically, the bias of a cell frequency θ_j is calculated as

$$\text{bias}_{\theta_j} = \frac{\sum_{it=1}^{N_{it}} \left(\theta_j - \hat{\theta}_{j_{it}}\right)}{N_{it}}. \tag{3.1}$$

Furthermore, the RMSE is calculated as

$$\text{RMSE} = \frac{\sqrt{\sum_{it=1}^{N_{it}} \left(\theta_j - \hat{\theta}_{j_{it}}\right)^2}}{N_{it}}. \tag{3.2}$$

Second, results are evaluated in terms of variance. Here, it is of interest to evaluate whether MILC correctly reflects uncertainty due to missing and conflicting values in between imputation variance for both univariate and multivariate cross-tables. Therefore, we investigate if the average of the estimated standard errors is approximately equal to the standard deviation over the 500 estimates obtained from the 500 simulation replications by evaluating its ratio, which is calculated by

$$\frac{\left[\sum_{it=1}^{N_{it}} \text{SE}\left(\hat{\theta}_{j_{it}}\right) \big/ N_{it}\right]}{\text{SD}\left(\hat{\theta}_{j_{it}}\right)}, \tag{3.3}$$

where SE is the square root of the estimate of the total variance obtained after applying pooling rules (Rubin, 1976) and $\text{SD}\left(\hat{\theta}_{j_{it}}\right)$ is calculated as

$$\text{SD}\left(\hat{\theta}_{j_{it}}\right) = \sqrt{\frac{\sum_{it=1}^{N_{it}} \left(\hat{\theta}_{j_{it}} - \bar{\theta}_{j_{it}}\right)^2}{N_{it}}}. \tag{3.4}$$

To account for small cell frequencies, $\hat{\theta}_{j_{it}}$ and $\bar{\theta}_{j_{it}}$ are considered on a log scale in equations 3.2, 3.3 and 3.4. To summarize, we denote the specific conditions evaluated in this simulation study as $Y_{v,1}$, MILC-MCAR-5, MILC-MCAR-10, MILC-MCAR-20, MILC-MAR-5, MILC-MAR-10 and MILC-MAR-20.

4. Simulation results

First, cell-proportions of univariate and multivariate cross-tables are evaluated in terms of bias and root mean squared error (RMSE) over the 500 simulation replications. Second, these cell-proportions are evaluated in terms of variance by investigating the average of the estimated standard error divided by the standard deviation over the 500 estimates obtained from the 500 simulation replications (SESD). Due to the log-transformations we made in equations 3.2, 3.3 and 3.4 to account for small cell frequencies, the RMSE and SESD are reported on a log scale.

4.1 Results in terms of bias

4.1.1 Univariate marginal frequencies of imputed variables

In Table 4.1, the simulation results can be found that cover the univariate marginal frequencies of the imputed latent variable “Gender” in terms of bias and RMSE. Results from all simulation conditions are shown. Here, it can be seen that a smaller amount of bias is obtained if $Y_{1,1}$ is used, compared to results obtained using MILC under all conditions. In addition, it can be seen that the RMSE is also smaller if $Y_{1,1}$ is used instead of the MILC method. Furthermore, it can be seen that both bias and RMSE slightly decrease as M increases, and that the quality of the results appears to be unrelated to the missingness mechanism.

Table 4.1
Results in terms of bias and root mean squared error for the two categories of the imputed latent variable “Gender”

				MCAR			MAR		
	Gender	Frequency	$Y_{1,1}$	$M=5$	$M=10$	$M=20$	$M=5$	$M=10$	$M=20$
Bias	F.	1,367,167	-2,126	3,386	3,308	3,325	3,231	3,153	3,109
	M.	1,324,310	2,126	-3,386	-3,308	-3,325	-3,231	-3,153	-3,109
RMSE	F.	1,367,167	2,154	6,008	5,888	5,760	5,914	5,637	5,512
	M.	1,324,310	2,154	6,008	5,888	5,760	5,914	5,637	5,512

Note: “F.” is “Female” and “M.” is “Male”.

In Table 4.2, the simulation results can be found that cover the univariate marginal frequencies of the imputed latent variable “Type of family nucleus” in terms of bias and RMSE. Here, the results are very different from the results we found for “Gender”, the bias obtained for $Y_{2,1}$ is much higher compared to the bias obtained using MILC under all conditions and the same holds for RMSE. In addition, whether the results for the MILC method depend on the missingness mechanism differ per category. In terms of bias and RMSE, this is the case for the categories “N.A.” and “Partners”.

Table 4.2
Results in terms of bias and root mean squared error for the four observed categories of the imputed latent variable "Type of family nucleus"

			$Y_{2,1}$	MCAR			MAR		
	Type of family nucleus	**Frequency**		$M = 5$	$M = 10$	$M = 20$	$M = 5$	$M = 10$	$M = 20$
Bias	Lone parents	97,360	2,670	185	182	176	224	226	220
	N.A.	604,032	8,985	-957	-975	-989	-1,601	-1,612	-1,611
	Partners	1,272,339	-19,686	401	411	427	932	935	932
	Sons/daughters	717,746	8,030	371	381	386	446	451	459
RMSE	Lone parents	97,360	2,672	425	408	395	426	421	414
	N.A.	604,032	8,989	1,337	1,318	1,312	1,837	1,833	1,818
	Partners	1,272,339	19,688	954	914	904	1,256	1,235	1,218
	Sons/daughters	717,746	8,034	630	624	617	715	692	688

Note: "N.A." means "Not applicable". Note that the category "Not stated" is mitigated as it contained zero observations.

In Table 4.3, the simulation results can be found that cover the univariate marginal frequencies of the imputed latent variable "Citizen" in terms of bias and RMSE. Here, the results are comparable to the results we found for "Type of family nucleus", as the bias obtained when only $Y_{3,1}$ is used is again much higher compared to the bias obtained using MILC method and the same holds for RMSE. As was also the case for "Type of family nucleus", whether the results for the MILC method depend on the missingness mechanism differ per category, although this is more the case for the bias here, and not so much in terms of RMSE.

Table 4.3
Results in terms of bias and root mean squared error for the four observed categories of the imputed latent variable "Citizen"

			$Y_{3,1}$	MCAR			MAR		
	Citizen	**Frequency**		$M = 5$	$M = 10$	$M = 20$	$M = 5$	$M = 10$	$M = 20$
Bias	EU	79,212	51,365	-5	-7	-12	-199	-211	-216
	NL	2,511,214	-116,899	-555	-546	-545	117	124	107
	not EU	89,592	58,085	512	502	507	62	69	89
	Not stated	11,459	7,448	49	51	49	21	18	20
RMSE	EU	79,212	51,365	410	398	388	488	486	475
	NL	2,511,214	116,899	925	894	883	767	756	720
	not EU	89,592	58,086	800	770	767	618	611	590
	Not stated	11,459	7,449	201	197	190	204	205	198

Note: "N.S." means "Not stated". Note that the category "Stateless" is mitigated as it contained zero observations.

Boeschoten et al. (2017) concluded that the quality of the output when MILC is applied related to how well the latent class model is able to make classifications based on the observed data, which is summarized in the entropy R^2. The entropy R^2 values for "Gender", "Type of family nucleus" and "Citizen" are approximately 0.7352, 0.9191, and 0.8571 respectively under MCAR. So this corresponds to the quality of the results for the latent variables in terms of bias and RMSE. An additional explanation for "Gender" is that the two categories are of comparable size and the amount of misclassification in both categories is approximately equal and behaves symmetrical in our simulation study. This causes that the marginal distribution of $Y_{1,1}$ is very similar to the marginal distribution of X_1 and not so much affected by misclassification.

4.1.2 Joint frequencies of imputed variables

In Table 4.4, the simulation results can be found that cover the joint marginal frequencies of the three imputed latent variables in terms of bias and RMSE. Again, it can be seen here that if only $Y_{v,1}$ is used, severe bias is present in all cells of the joint frequency table. The results obtained when the MILC method is applied show much lower amounts of bias and RMSE. Here, the differences between different numbers for M or different missingness mechanism are much smaller compared to the differences between MILC and $Y_{v,1}$. Furthermore, the differences in the amount of bias for particular cells after applying the MILC method seem to be related to imbalances in cell frequencies within particular variables. More specifically, the variable "Citizen" knows substantive differences in cell frequencies and within Table 4.4, it can be seen that particular the category "not EU" is affected in terms of bias by this imbalance.

Table 4.4
Results in terms of bias and root mean squared error for the 32 observed categories of the joint distribution of the three imputed latent variables "Gender", "Type of family nucleus" and "Citizen"

	Gender × Type of family nucleus × Citizen			Frequency	$Y_{v,1}$	MCAR			MAR		
	Gender	Family nucleus	Citizen			$M = 5$	$M = 10$	$M = 20$	$M = 5$	$M = 10$	$M = 20$
Bias	F.	Lone parents	EU	2,091	1,434	8	7	7	1	0	0
	F.	Lone parents	NL	76,131	-6,620	652	650	646	240	241	234
	F.	Lone parents	not EU	3,120	1,513	33	32	32	39	39	38
	F.	Lone parents	N.S.	646	154	-5	-5	-6	-13	-13	-13
	F.	N.A.	EU	12,436	5,971	433	432	432	431	427	427
	F.	N.A.	NL	293,960	-11,998	-595	-618	-623	905	891	880
	F.	N.A.	not EU	9,509	7,317	1,032	1,031	1,032	1,069	1,069	1,071
	F.	N.A.	N.S.	1,221	982	182	182	182	198	197	197
	F.	Partners	EU	20,443	11,185	237	236	235	24	19	21
	F.	Partners	NL	584,547	-34,001	294	262	279	-564	-599	-624
	F.	Partners	not EU	26,877	12,022	404	402	401	254	255	258
	F.	Partners	N.S.	1,292	1,837	-19	-18	-18	-23	-24	-24
	F.	Sons/daughters	EU	4,368	7,541	-778	-779	-780	-851	-853	-854
	F.	Sons/daughters	NL	321,364	-8,738	2,483	2,471	2,479	2,620	2,601	2,588
	F.	Sons/daughters	not EU	7,680	8,303	-764	-768	-766	-876	-874	-869
	F.	Sons/daughters	N.S.	1,482	971	-209	-208	-208	-223	-223	-222
	M.	Lone parents	EU	389	591	-10	-11	-11	9	9	9
	M.	Lone parents	NL	14,536	4,791	-553	-552	-554	-134	-131	-130
	M.	Lone parents	not EU	372	707	35	35	35	53	53	53
	M.	Lone parents	N.S.	75	100	27	27	27	28	29	29
	M.	N.A.	EU	16,308	4,444	-306	-304	-305	-349	-349	-350
	M.	N.A.	NL	253,493	-3,733	-714	-708	-717	-2,730	-2,722	-2,713
	M.	N.A.	not EU	13,636	5,548	-904	-903	-902	-1,023	-1,023	-1,020
	M.	N.A.	N.S.	3,469	455	-85	-86	-87	-102	-103	-104
	M.	Partners	EU	18,444	11,881	793	796	794	905	906	906
	M.	Partners	NL	599,278	-38,164	-3,170	-3,128	-3,127	-1,528	-1,490	-1,474
	M.	Partners	not EU	19,776	13,709	1,794	1,793	1,793	1,785	1,790	1,791
	M.	Partners	N.S.	1,682	1,846	69	69	69	78	78	79
	M.	Sons/daughters	EU	4,733	8,319	-382	-382	-384	-370	-371	-374
	M.	Sons/daughters	NL	367,905	-18,435	1,049	1,076	1,072	1,308	1,333	1,346
	M.	Sons/daughters	not EU	8,622	8,966	-1,118	-1,120	-1,117	-1,240	-1,239	-1,233
	M.	Sons/daughters	N.S.	1,592	1,103	90	90	91	77	77	78

Note: "N.S." means "Not stated" and "N.A." means "Not applicable". Note that the categories "Stateless" for "Citizen" and "Not Stated" for "Type of family nucleus" are mitigated as they contained zero observations.

Table 4.4 (continued)
Results in terms of bias and root mean squared error for the 32 observed categories of the joint distribution of the three imputed latent variables "Gender", "Type of family nucleus" and "Citizen"

	Gender × Type of family nucleus × Citizen					MCAR			MAR		
	Gender	Family nucleus	Citizen	Frequency	$Y_{v,1}$	$M=5$	$M=10$	$M=20$	$M=5$	$M=10$	$M=20$
RMSE	F.	Lone parents	EU	2,091	1,434	45	42	41	45	42	40
	F.	Lone parents	NL	76,131	6,621	742	734	724	418	408	394
	F.	Lone parents	not EU	3,120	1,514	67	64	64	71	68	66
	F.	Lone parents	N.S.	646	155	22	21	20	26	25	24
	F.	N.A.	EU	12,436	5,972	449	446	445	447	442	440
	F.	N.A.	NL	293,960	12,001	1,260	1,245	1,222	1,433	1,374	1,348
	F.	N.A.	not EU	9,509	7,317	1,038	1,037	1,037	1,075	1,075	1,076
	F.	N.A.	N.S.	1,221	983	185	185	185	202	201	201
	F.	Partners	EU	20,443	11,186	291	285	282	173	163	157
	F.	Partners	NL	584,547	34,003	2,332	2,285	2,204	2,364	2,248	2,197
	F.	Partners	not EU	26,877	12,023	456	450	447	330	327	327
	F.	Partners	N.S.	1,292	1,838	46	44	43	48	48	47
	F.	Sons/daughters	EU	4,368	7,541	787	787	787	860	862	863
	F.	Sons/daughters	NL	321,364	8,742	2,820	2,796	2,781	2,959	2,903	2,879
	F.	Sons/daughters	not EU	7,680	8,304	779	782	780	892	889	883
	F.	Sons/daughters	N.S.	1,482	972	216	214	214	230	230	229
	M.	Lone parents	EU	389	592	18	17	17	17	17	16
	M.	Lone parents	NL	14,536	4,792	605	600	600	271	260	257
	M.	Lone parents	not EU	372	707	38	38	37	55	55	55
	M.	Lone parents	N.S.	75	101	27	27	27	29	29	29
	M.	N.A.	EU	16,308	4,445	331	328	327	373	371	370
	M.	N.A.	NL	253,493	3,742	1,390	1,349	1,314	2,959	2,931	2,911
	M.	N.A.	not EU	13,636	5,549	913	912	911	1,033	1,031	1,028
	M.	N.A.	N.S.	3,469	456	107	105	104	121	121	120
	M.	Partners	EU	18,444	11,881	808	810	807	919	919	917
	M.	Partners	NL	599,278	38,165	3,898	3,837	3,794	2,755	2,617	2,568
	M.	Partners	not EU	19,776	13,709	1,804	1,803	1,803	1,797	1,800	1,800
	M.	Partners	N.S.	1,682	1,846	88	87	85	98	95	95
	M.	Sons/daughters	EU	4,733	8,319	403	403	403	401	401	402
	M.	Sons/daughters	NL	367,905	18,437	1,728	1,723	1,687	1,905	1,872	1,854
	M.	Sons/daughters	not EU	8,622	8,967	1,129	1,130	1,127	1,252	1,250	1,244
	M.	Sons/daughters	N.S.	1,592	1,104	109	108	107	103	102	101

Note: "N.S." means "Not stated" and "N.A." means "Not applicable". Note that the categories "Stateless" for "Citizen" and "Not Stated" for "Type of family nucleus" are mitigated as they contained zero observations.

4.1.3 Restricted cells

In Table 4.5, the simulation results can be found for the six cells that are restricted in the marginal cross-table between "Age" and "Type of family nucleus". Under "Frequency", it can be seen that these six cells should all contain zero observations. A combination of these scores is logically impossible. Furthermore, it can be seen that due to misclassification in $Y_{2,1}$, observations containing these combinations of scores are present when $Y_{2,1}$ is used to estimate this cross-table directly. In addition, it can be seen that if the MILC method is applied, such impossible combinations of scores will never be present, regardless of the missingness mechanism or the number of imputations. Furthermore, as the cells in this marginal table contain zero observations, all cells of more detailed tables covering these logically impossible combinations of scores automatically also contain zero observations.

Table 4.5
Results in terms of bias and root mean squared error for the six restricted categories from cross-table between "Type of family nucleus" and the covariate "Age"

	Type of family nucleus		Frequency	$Y_{2,1}$	MCAR			MAR		
					$M = 5$	$M = 10$	$M = 20$	$M = 5$	$M = 10$	$M = 20$
Bias	Lone parents	under 5 years	0	377	0	0	0	0	0	0
	Lone parents	5 to 9 years	0	386	0	0	0	0	0	0
	Lone parents	10 to 14 years	0	376	0	0	0	0	0	0
	Partners	under 5 years	0	4,934	0	0	0	0	0	0
	Partners	5 to 9 years	0	5,041	0	0	0	0	0	0
	Partners	10 to 14 years	0	4,937	0	0	0	0	0	0
RMSE	Lone parents	under 5 years	0	377	0	0	0	0	0	0
	Lone parents	5 to 9 years	0	386	0	0	0	0	0	0
	Lone parents	10 to 14 years	0	377	0	0	0	0	0	0
	Partners	under 5 years	0	4,934	0	0	0	0	0	0
	Partners	5 to 9 years	0	5,041	0	0	0	0	0	0
	Partners	10 to 14 years	0	4,937	0	0	0	0	0	0

4.1.4 The complete population frequency table

Figures 4.1 and 4.2 show results in terms of bias and root mean squared error (RMSE) when the complete census table, so the cross-table between the six variables, is estimated. As these are 42,000 cells in total, it is not feasible to evaluate them individually. Figure 4.1 and Figure 4.2 give an overview of how size of the cell frequency is related to the quality of the results. Here it can be seen that if $Y_{v,1}$ are used, results in terms of bias and RMSE are related directly to cell frequency. More specifically, the relationship between cell frequency and absolute bias is approximately linear where the amount of bias is approximately 10% of the cell frequency.

Figure 4.1 Results in terms of bias when the complete cross-table between the latent variables "Gender", "Type of family nucleus" and "Citizen" and the three covariates "Age", "Marital status" and "Place of birth" is estimated. The X-axis represents cell frequency and the Y-axis represents the bias. Results are shown for $Y_{v,1}$, MILC-MCAR-20 and MILC-MAR-20.

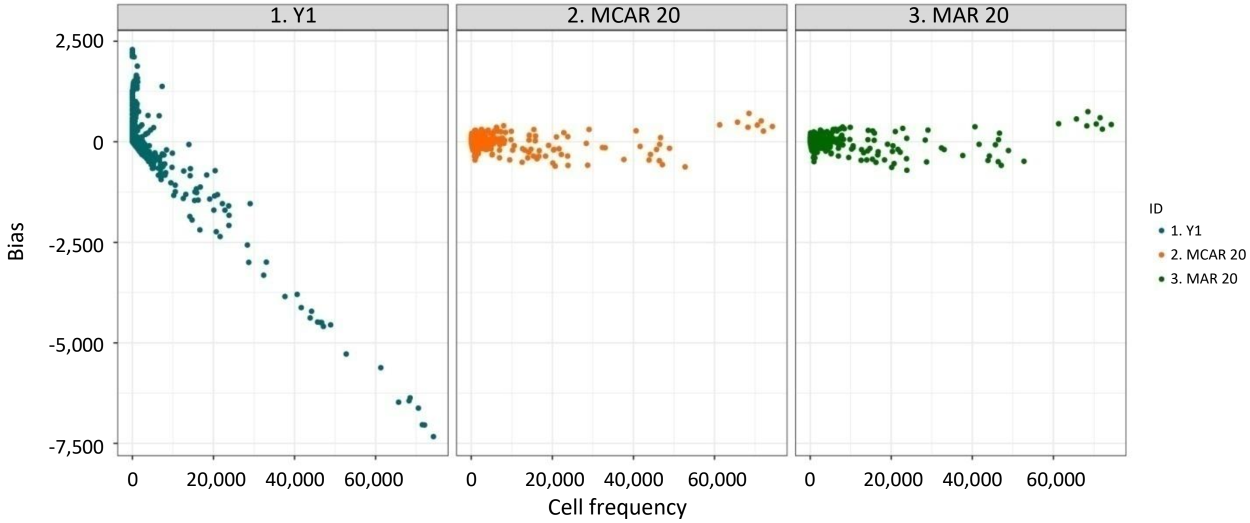
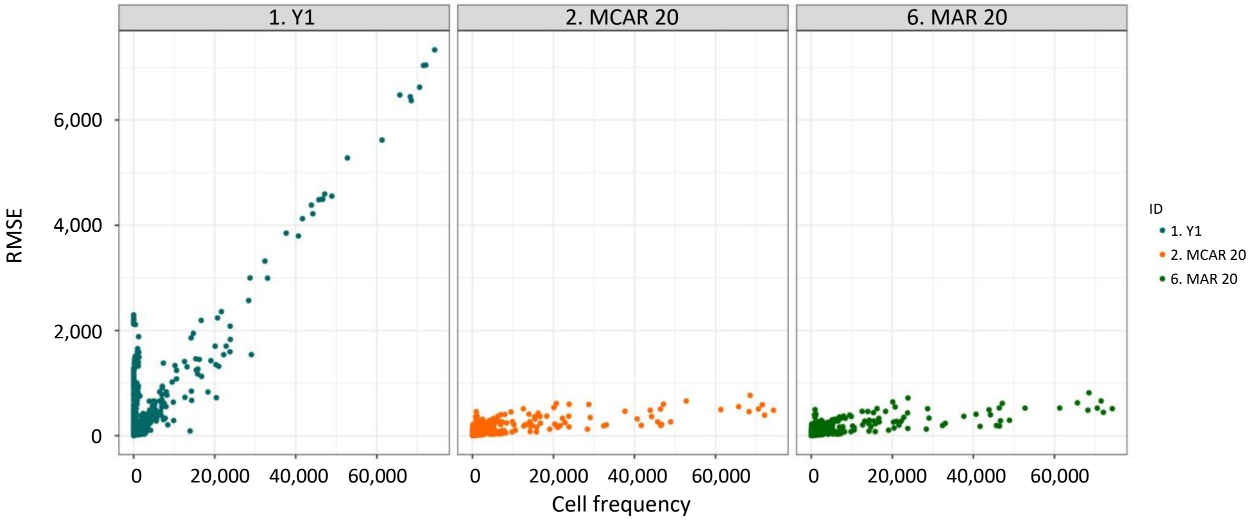


Figure 4.2 Results in terms of root mean squared error (RMSE) when the complete cross-table between the latent variables "Gender", "Type of family nucleus" and "Citizen" and the three covariates "Age", "Marital status" and "Place of birth" is estimated. The X-axis represents cell frequency and the Y-axis represents the RMSE. Results are shown for $Y_{v,1}$, MILC-MCAR-20 and MILC-MAR-20.



4.2 Results in terms of variance

4.2.1 Univariate marginal frequencies of imputed variables

In Table 4.6, the simulation results can be found that cover the univariate marginal frequencies "Gender" in terms of se/sd. As this ratio measures whether the average standard error estimated at each replication in the simulation correctly describes the uncertainty (standard deviation) that is found over the estimates, it should be close to one. In addition, as a completely observed and finite population is assumed, variance is not estimated when $Y_{v,1}$ is used. The results obtained using MILC are generally close to one and comparable to the results in terms of bias as only minor differences can be found between different values for M or between the different missingness mechanisms.

Table 4.6
Results in terms of average standard error of the estimates divided by standard deviation over the estimates (se/sd) for the two categories of the imputed latent variable "Gender"

	Gender	Frequency	$Y_{v,1}$	MCAR			MAR		
				$M=5$	$M=10$	$M=20$	$M=5$	$M=10$	$M=20$
se/sd	F.	1,367,167	-	1.0540	1.0317	1.0363	1.0030	1.0235	1.0237
	M.	1,324,310	-	1.0546	1.0317	1.0363	1.0034	1.0236	1.0236

Note: ("F." is "Female" and "M." is "Male").

In Table 4.7 and 4.8, the simulation results can be found that cover the univariate marginal frequencies for "Type of family nucleus" and "Citizen" respectively in terms of se/sd. The results found here have a very comparable structure compared to the results we found for "Gender".

Table 4.7
Results in terms of average standard error of the estimates divided by standard deviation over the estimates (se/sd) for the four observed categories of the imputed latent variable “Type of family nucleus”

	Type of family nucleus	Frequency	$Y_{v,1}$	MCAR			MAR		
				$M = 5$	$M = 10$	$M = 20$	$M = 5$	$M = 10$	$M = 20$
se/sd	Lone parents	97,360	-	1.0457	1.0510	1.0529	1.0561	1.0337	1.0336
	N.A.	604,032	-	0.9706	0.9874	0.9922	0.9751	0.9829	0.9863
	Partners	1,272,339	-	1.0332	1.0418	1.0456	1.0052	1.0269	1.0298
	Sons/daughters	717,746	-	0.9594	0.9615	0.9606	0.9696	0.9880	0.9938

Note: “N.A.” means “Not applicable”. Note that the category “Not stated” is mitigated as it contained zero observations.

Table 4.8
Results in terms of average standard error of the estimates divided by standard deviation over the estimates for the four observed categories of the imputed latent variable “Citizen”

	Type of family nucleus	Frequency	$Y_{v,1}$	MCAR			MAR		
				$M = 5$	$M = 10$	$M = 20$	$M = 5$	$M = 10$	$M = 20$
se/sd	Citizen EU	79,212	-	1.0417	1.0172	1.0362	1.0768	1.0539	1.0571
	Citizen NL	2,511,214	-	1.0136	1.0113	1.0235	1.0925	1.0645	1.0927
	Citizen not EU	89,592	-	0.9478	0.9632	0.9709	1.0282	0.9916	1.0125
	Not stated	11,459	-	1.0063	1.0208	1.0238	1.1057	1.0861	1.1143

Note: “N.S.” means “Not stated”. Note that the category “Stateless” is mitigated as it contained zero observations.

4.2.2 Joint frequencies of imputed variables

In Table 4.9, the simulation results can be found that cover the joint marginal frequencies of the imputed latent variables “Gender”, “Type of family nucleus” and “Citizen” in terms of absolute se/sd. The results found for these joint frequencies are very comparable to the results we found for the marginal frequencies. For cells with a relatively low frequency, it can be seen that the ratio is in general larger than one, indicating that the variance estimated for these frequencies (and thereby the differences between the imputations) incorporate more uncertainty than is actually found over different replications. Summarizing, the uncertainty for cells containing low frequencies is overestimated.

Results in terms for variance are not shown for the restricted cells, as a variance term cannot be estimated here.

Table 4.9
Results in terms of average standard error of the estimates divided by standard deviation over the estimates for the 32 observed categories of the joint distribution of the three imputed latent variables “Gender”, “Type of family nucleus” and “Citizen”

Gender × Type of family nucleus × Citizen			Frequency	$Y_{v,1}$	MCAR			MAR		
Gender	Family nucleus	Citizen			$M = 5$	$M = 10$	$M = 20$	$M = 5$	$M = 10$	$M = 20$
F.	Lone parents	EU	2,091	-	1.1813	1.2097	1.2032	1.1495	1.1654	1.1997
F.	Lone parents	NL	76,131	-	1.0371	1.0471	1.0504	1.0270	1.0252	1.0349
F.	Lone parents	not EU	3,120	-	1.1659	1.1590	1.1519	1.1607	1.1634	1.1870

Note: “N.S.” means “Not stated” and “N.A.” means “Not applicable”. Note that the categories “Stateless” for “Citizen” and “Not Stated” for “Type of family nucleus” are mitigated as they contained zero observations.

Table 4.9 (continued)
Results in terms of average standard error of the estimates divided by standard deviation over the estimates for the 32 observed categories of the joint distribution of the three imputed latent variables "Gender", "Type of family nucleus" and "Citizen"

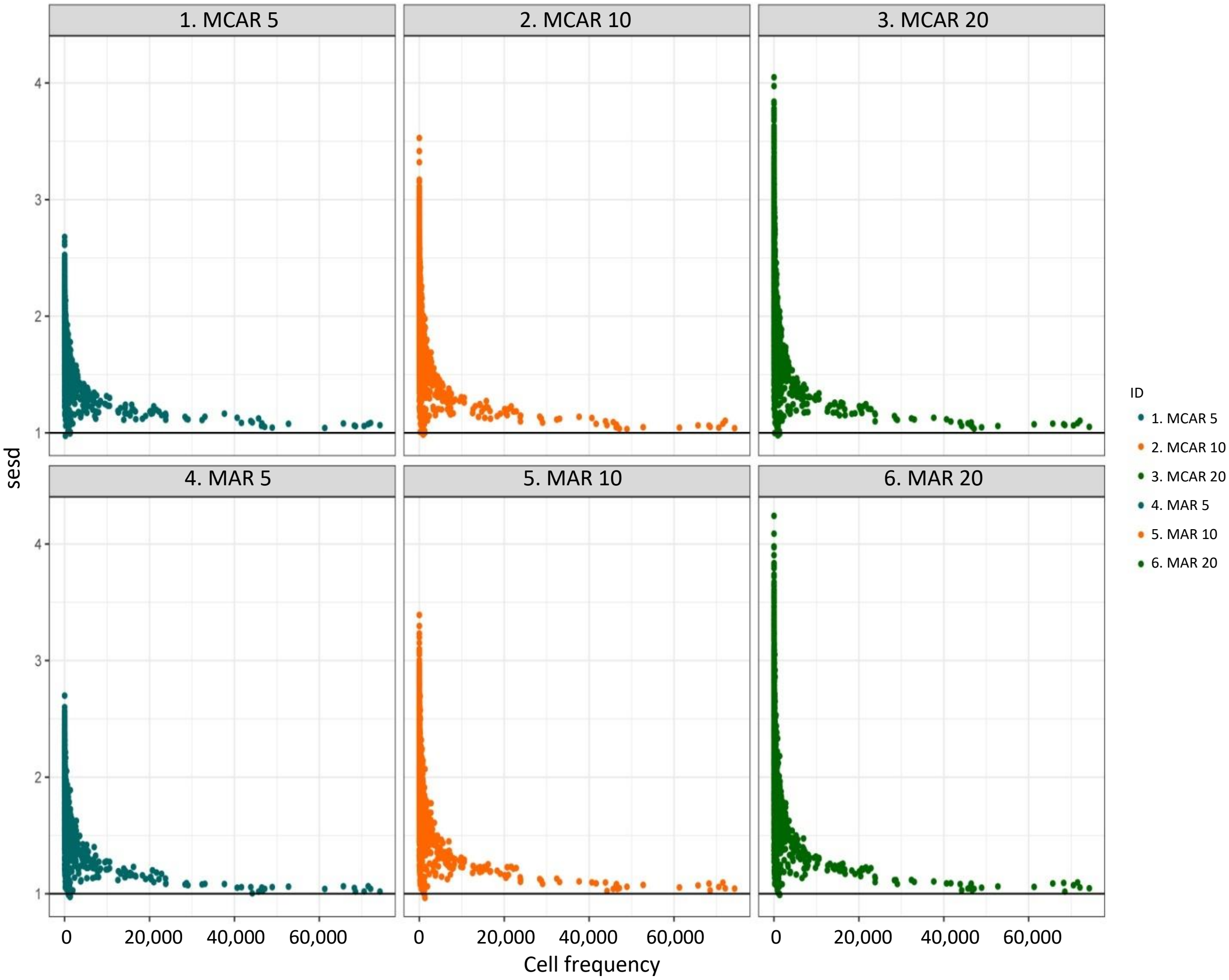
Gender × Type of family nucleus × Citizen					MCAR			MAR		
Gender	**Family nucleus**	**Citizen**	**Frequency**	$Y_{v,1}$	$M = 5$	$M = 10$	$M = 20$	$M = 5$	$M = 10$	$M = 20$
F.	Lone parents	N.S.	646	-	1.0963	1.1004	1.1272	1.1110	1.1000	1.1054
F.	N.A.	EU	12,436	-	1.0850	1.0838	1.1172	1.0888	1.1065	1.1456
F.	N.A.	NL	293,960	-	1.0840	1.0652	1.0575	1.0158	1.0406	1.0461
F.	N.A.	not EU	9,509	-	1.1636	1.1822	1.1892	1.1574	1.1383	1.1562
F.	N.A.	N.S.	1,221	-	1.1789	1.1964	1.2097	1.1959	1.1826	1.2133
F.	Partners	EU	20,443	-	1.0508	1.0537	1.0653	1.0689	1.0684	1.0925
F.	Partners	NL	584,547	-	1.0313	1.0099	1.0189	1.0035	1.0253	1.0197
F.	Partners	not EU	26,877	-	1.0532	1.0766	1.0720	1.0765	1.0725	1.0733
F.	Partners	N.S.	1,292	-	1.1471	1.1566	1.1504	1.2157	1.1855	1.1940
F.	Sons/daughters	EU	4,368	-	1.0135	1.0147	1.0338	1.0430	1.0518	1.0479
F.	Sons/daughters	NL	321,364	-	1.0548	1.0379	1.0527	1.0017	1.0222	1.0221
F.	Sons/daughters	not EU	7,680	-	0.9977	0.9966	0.9909	1.0249	1.0132	1.0416
F.	Sons/daughters	N.S.	1,482	-	1.0344	1.0325	1.0357	1.0836	1.0688	1.0890
M.	Lone parents	EU	389	-	1.3198	1.4136	1.4316	1.2941	1.3575	1.4470
M.	Lone parents	NL	14,536	-	1.0784	1.0762	1.0736	1.0755	1.0690	1.0650
M.	Lone parents	not EU	372	-	1.4159	1.3857	1.4511	1.4814	1.4481	1.4619
M.	Lone parents	N.S.	75	-	1.4330	1.5192	1.5659	1.4598	1.5035	1.5373
M.	N.A.	EU	16,308	-	1.0990	1.0908	1.1165	1.0894	1.1022	1.1366
M.	N.A.	NL	253,493	-	1.0035	1.0100	1.0193	0.9920	1.0175	1.0238
M.	N.A.	not EU	13,636	-	1.1168	1.1100	1.1141	1.0826	1.1054	1.0952
M.	N.A.	N.S.	3,469	-	1.0241	1.0818	1.1052	1.1592	1.1478	1.1780
M.	Partners	EU	18,444	-	1.1618	1.1593	1.1579	1.1473	1.1335	1.1476
M.	Partners	NL	599,278	-	1.0668	1.0444	1.0487	1.0081	1.0329	1.0231
F.	Partners	not EU	19,776	-	1.0932	1.0788	1.0816	1.0674	1.0612	1.0911
F.	Partners	N.S.	1,682	-	1.1068	1.1411	1.1418	1.1335	1.1719	1.1770
F.	Sons/daughters	EU	4,733	-	1.0598	1.0396	1.0548	1.0528	1.0497	1.0414
F.	Sons/daughters	NL	367,905	-	1.0549	1.0347	1.0365	1.0098	1.0298	1.0340
F.	Sons/daughters	not EU	8,622	-	1.0077	1.0093	1.0100	1.0413	1.0449	1.0471
F.	Sons/daughters	N.S.	1,592	-	1.0472	1.0617	1.0699	1.0458	1.0362	1.0627

Note: ("N.S." means "Not stated" and "N.A." means "Not applicable"). Note that the categories "Stateless" for "Citizen" and "Not Stated" for "Type of family nucleus" are mitigated as they contained zero observations.

4.2.3 The complete population frequency table

In Figure 4.3, results can be found in terms of average standard error of the cell frequencies divided by the standard deviation over the frequencies estimated in the 500 replications in the simulation study (se/sd). Here it can be seen that the standard error estimated per cell frequency is especially too large when cell frequencies are close to zero, and become closer to the nominal rate of one as the cell frequencies become larger. Apparently, variability due to missing and conflicting values is overestimated by MILC for cells with a frequency close to zero. In addition, this becomes more apparent when the number of imputations increases and it is not influenced by missingness mechanism.

Figure 4.3 Results in terms of average standard error of the cell frequencies divided by the standard deviation over the frequencies (se/sd) when the complete cross-table between the latent variables "Gender", "Type of family nucleus" and "Citizen" and the three covariates "Age", "Marital status" and "Place of birth" is estimated. The X-axis represents cell frequency and the Y-axis represents the se/sd ratio. Results are shown for MILC-MCAR-5, MILC-MCAR-10, MILC-MCAR-20, MILC-MAR-5, MILC-MAR-10 and MILC-MAR-20.



4.3 Sensitivity to violations of assumptions

The simulation study presented in this paper is aimed at investigating the performance of the MILC method in a situation of misclassification in a finite population setting. When applying the MILC method in practice, a number of assumptions are made and during this simulation study these assumptions were met. To further investigate the sensitivity to violations of these assumptions, additional simulation studies were performed.

An important assumption made when applying the MILC method is that the missingness mechanism is either MCAR or MAR. Therefore, a first sensitivity analysis involves a Missing Not At Random (MNAR) mechanism. More specifically, we generated this mechanism in such a way that the probability of being missing in the survey indicator for "Type of family nucleus" depends on the latent variable "type of family nucleus" and is smallest for the first category and largest for the last category. In Table 4.10, it can be seen that the bias and RMSE increase when the mechanism is MNAR compared to MAR, while the se/sd is not affected. More specifically, it can be seen that the extent of the bias relates to how much the respective class is affected by the mechanism.

A second assumption states that the measurement error present in the indicators is random. To investigate sensitivity to the violation of this assumption, we generated a selective measurement error mechanism where the probability of measurement error in the register indicator for the variable "type of family nucleus" differs per category. Here, again the first category is least affected and the last category most. In Table 4.10 it can be seen that the effect of this selective mechanism are limited. The bias increases in a similar way as the percentage of measurement error in the respective category increases, but these are still relatively low amounts of bias. The se/sd is not affected by the mechanism.

Table 4.10
Results in terms of bias, root mean squared error and se/sd for the four observed categories of the imputed latent variable "Type of family nucleus"

	Type of family nucleus	Frequency	$Y_{2,1}$	MAR	MNAR	Selective	ME covar
Bias	Lone parents	97,360	2,670	224	6,256	105	1,172,993
	N.A.	604,032	8,985	-1,601	27,002	-1,824	534
	Partners	1,272,339	-19,686	932	-11,341	1,116	-1,174,697
	Sons/daughters	717,746	8,030	446	-21,917	603	1,170
RMSE	Lone parents	97,360	2,672	426	6,268	332	1,172,994
	N.A.	604,032	8,989	1,837	27,017	2,060	1,094
	Partners	1,272,339	19,688	1,256	11,377	1,466	1,174,697
	Sons/daughters	717,746	8,034	715	21,924	819	1,291
se/sd	Lone parents	97,360	-	1.0561	1.01936	1.0634	1.0518
	N.A.	604,032	-	0.9751	1.02491	0.9722	1.0471
	Partners	1,272,339	-	1.0052	0.97456	0.9291	0.9649
	Sons/daughters	717,746	-	0.9696	1.02547	1.0962	1.0181

Note: "N.A." means "Not applicable" under different violations of assumptions. Note that the category "Not stated" is mitigated as it contained zero observations.

A third assumption is that covariates do not contain measurement error. This assumption is the most remarkable, as it is typically often not the case that a coviarate does not contain measurement error. It is more likely that these variables will be treated as such because no additional information about their measurement error is known. If information was known, for example because additional survey information was present, it would have been incorporated by means of a latent variable. As in practice however there is always a probability that for some variables such information is not known, we investigate the sensitivity of the method to violation of this assumption. More specifically, we generated 5% misclassification in the covariate "marital status", which has a relatively strong association with the latent variable "type of family nucleus". Indeed, the bias in some categories is highly affected by this misclassification.

5. Discussion

In this paper, the performance of the MILC method was investigated in a situation where misclassification was induced in a finite population setting. Here, an existing population census table was used as a starting point, and for three categorical variables present in this census table, two indicator variables were generated with 5% misclassification each, where one indicator also contains approximately 90% missing values. As a finite population was assumed, the estimated variance only contained a between variance component reflecting the differences between the imputations and thereby the uncertainty caused by the misclassification and missing values in the indicator variables.

The simulation results show that the method, regardless of the number of imputations, produces results with a low bias for marginal frequency distributions, cross-tables between imputed latent variables and covariates and even for the complete six-way cross-table. Striking is the amount of bias that is induced when the indicator observed via the register is used to calculate the cross-tables evaluated in comparison to when MILC is used. It is also shown that if these indicators are used, it is likely that impossible combinations of scores are produced as well, something that can be easily circumvented by specifying edit restrictions in the LC model. This simulation study once again shows that misclassification, even if it is non-systematic, can seriously bias results. In terms of variance, it was seen that if the MILC method is applied, variance estimates are appropriate in general. However, if cell frequencies are relatively small, the variance is overestimated. This problem is more severe if the complete frequency table is evaluated, because this large table contains many cells with low frequencies.

The current set-up of this simulation study knows two major limitations. The first is caused by the large amount of cells in the cross-table. Because of this, a latent class model containing only main effects was used. It was not feasible to use a saturated model as the number of parameters would be very large, and it would be likely that not every parameter is estimable in every bootstrap sample. This would limit the use of starting values, thereby increasing the computation time for the simulation study to an unfeasible amount.

A second limitation is that in our simulation set-up we only considered relatively simple sampling designs for the survey data: simple random sampling (MCAR conditions) and, essentially, stratified simple random sampling (MAR conditions). A future study could examine to what extent the MILC method can also correct for misclassification error with appropriate variance estimates when survey data are obtained by a complex sampling design that involves, for instance, cluster sampling, multistage sampling or sampling with unequal probabilities proportional to size. In the context of missing data it has been found that, although a generally accepted theory is still lacking, in practice multiple imputation often works reasonably well for complex samples, provided that design variables and/or survey weights are included in the imputation model; see, e.g., Rässler (2004, page 14) and the references listed there. It would be interesting to investigate whether this result also applies to multiple imputation in the context of correcting for measurement errors. As an alternative, Zhou, Elliott and Raghunathan (2016) proposed a Bayesian approach to incorporate survey design features into a multiple-imputation analysis.

The starting point of this simulation study was an existing population census table. A nice property here was that we could approach this as a finite and known population. Therefore, we did not have to include (within) sampling variance in our estimate of the total variance. It was insightful to evaluate cell frequencies of both univariate and multivariate cross-tables as results generally appeared to be related to cell-frequency.

References

Bakker, B. (2010). *Micro-integration, state of the art*. Paper presented at the joint UNECE-Eurostat expert group meeting on registered based censuses in The Hague, May 11, 2010. Retrieved from https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2010/wp.10.e.pdf (date visited 2017.04.24).

Bakker, B. (2011). Micro-integration. statistical methods 201108. *Statistics Netherlands*.

Bakker, B., Van Rooijen, J. and Van Toor, L. (2014). The system of social statistical datasets of statistics netherlands: An integral approach to the production of register-based social statistics. *Statistical Journal of the IAOS*, 30(4), 411-424.

Bikker, R., Daalmans, J. and Mushkudiani, N. (2013). Benchmarking large accounting frameworks: A generalized multivariate model. *Economic Systems Research*, 25(4), 390-408.

Boeschoten, L., de Waal, T., and Vermunt, J.K. (2019). Estimating the number of serious road injuries per vehicle type in the netherlands by using multiple imputation of latent classes. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182(4), 1463-1486. Retrieved from https://doi.org/10.1111/rssa.12471.

Boeschoten, L., Filipponi, D. and Varriale, R. (2021). Combining multiple imputation and hidden markov modeling to obtain consistent estimates of employment status. *Journal of Survey Statistics and Methodology*, 9(3), 549-573. Retrieved from https://doi.org/10.1093/jssam/smz052.

Boeschoten, L., Oberski, D. and de Waal, T. (2017). Estimating classification errors under edit restrictions in composite survey-register data using Multiple Imputation Latent Class modelling (MILC). *Journal of Official Statistics*, 33(4), 921-962. Retrieved from https://doi.org/10.1515/jos-2017-0044.

Census Hub (2017, July). *European Statistical System*. Online, July 2017, (last visited 11/07/2017).

Daalmans, J. (2018). Divide-and-Conquer solutions for estimating large consistent table sets. *Statistical Journal of the IAOS*, 34(2), 223-233.

Daalmans, J. (2019). Pushing the Boundaries for Automated Data Reconciliation in Official Statistics. Tilburg University.

de Waal, T., Pannekoek, J. and Scholtus, S. (2011). *Handbook of Statistical Data Editing and Imputation*, New York: John Wiley & Sons, Inc., 563. (ISBN 0470904836, 9780470904831).

de Waal, T., van Delden, A. and Scholtus, S. (2020). Multi-source statistics: Basic situations and methods. *International Statistical Review*, 88(1), 203-228.

Di Fonzo, T., and Martini, M. (2003). Benchmarking systems of seasonally adjusted time series according.

European Commission (2008). Regulation (ec) no 763/2008 of the european parliament and of the council of 9 july 2008 on population and housing censuses. *Official Journal of the European Union*, (L218), 14-20.

European Commission (2009). Commission regulation (ec) no 1201/2009 of 30 november 2009 implementing regulation (ec) no 763/2008 of the european parliament and of the council on population and housing censuses as regards the technical specifications of the topics and of their breakdowns. *Official Journal of the European Union*, (L329), 29-68.

European Commission (2010). Commission regulation (eu) no 1151/2010 of 8 december 2010 implementing regulation (ec) no 763/2008 of the european parliament and of the council on population and housing censuses, as regards the modalities and structure of the quality reports and the technical format for data transmission. *Official Journal of the European Union*, (L324), 1-12.

Geerdinck, M., Goedhuys-van der Linden, M., Hoogbruin, E., De Rijk, A., Sluiter, N. and Verkleij, C. (2014). *Monitor Kwaliteit Stelsel Van Basisregistraties: Nulmeting Van de Kwaliteit Van Basisregistraties in Samenhang, 2014* (13114th Ed.). Henri Faasdreef 312, 2492 JP Den Haag, Centraal Bureau voor de Statistiek. Retrieved from https://www.cbs.nl/-/media/pdf/2016/50/monitor-kwaliteit-stelsel-van-basisregistraties.pdf (date visited 2017.04.25).

Magnus, J.R., van Tongeren, J.W. and de Vos, A.F. (2000). National accounts estimation using indicator ratios. *Review of Income and Wealth*, 46(3), 329-350.

Mashreghi, Z., Haziza, D. and Léger, C. (2016). A survey of bootstrap methods in finite population sampling. *Statistics Surveys*, 10, 1-52.

Pankowska, P., Pavlopoulos, D., Bakker, B. and Oberski, D.L. (2020). Reconciliation of inconsistent data sources using hidden markov models. *Statistical Journal of the IAOS*, 36(4), 1261-1279.

Rässler, S. (2004). The impact of multiple imputation for DACSEIS. (DACSEIS Research Paper Series No. 5).

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63(3), 581-592.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc., 81. Retrieved from dx.doi.org//10.1002/9780470316696 (ISBN 9780471087052) doi: 10.1002/9780470316696.

Särndal, C.-E., Swensson, B. and Wretman, J. (2003). *Model Assisted Survey Sampling*. Springer Science & Business Media.

Schulte Nordholt, E., Van Zeijl, J. and Hoeksma, L. (2014). Dutch census 2011, analysis and methodology. *Statistics Netherlands*. Retrieved from https://www.cbs.nl/-/media/imported/documents/2014/44/2014-b57-pub.pdf (date visited 2017.04.25).

Sefton, J., and Weale, M. (1995). Reconciliation of National Income and Expenditure: Balanced Estimates of National Income for the United Kingdom, 1920-1990. Cambridge University Press, 7.

Stone, R., Champernowne, D.G. and Meade, J.E. (1942). The precision of national income estimates. *The Review of Economic Studies*, 9(2), 111-125.

The Economic and Social Council (2005). Ecosoc resolution 2005/13. 2010 World Population and Housing Census Programme. doi: http://www.un.org/en/ecosoc/docs/2005/resolution%202005-13.pdf.

van Rooijen, J., Bloemendal, C. and Krol, N. (2016). The added value of micro-integration: Data on laid-off employees. *Statistical Journal of the IAOS*, 32(4), 685-692.

Vermunt, J.K., and Magidson, J. (2013a). Latent GOLD 5.0 Upgrade Manual [Computer software manual]. Belmont, MA, Retrieved from https://www.statisticalinnovations.com/wp-content/uploads/LG5manual.pdf (date visited 2017.04.25).

Vermunt, J.K., and Magidson, J. (2013b). Technical guide for *Latent GOLD 5.0: Basic, Advanced, and Syntax*. Statistical Innovations Inc., Belmont, MA. Retrieved from https://www.statisticalinnovations.com/wp-content/uploads/LGtecnical.pdf (date visited 2017.04.25).

Vink, G., and van Buuren, S. (2014). Pooling multiple imputations when the sample happens to be the population. *arXiv* preprint arXiv:1409.8542, Retrieved from https://arxiv.org/abs/1409.8542.

Zhou, H., Elliott, M.R. and Raghunathan, T.E. (2016). A two-step semiparametric method to accommodate sampling weights in multiple imputation. *Biometrics*, 72, 242-252. doi: 10.1111/biom.12413.