Estimating classification error under edit restrictions in combined survey-register data using Multiple Imputation Latent Class modelling (MILC)

Laura Boeschoten, Daniel Oberski & Ton de Waal

June 27, 2016

Abstract

Both registers and surveys can contain classification errors. These errors can be estimated by making use of information that is obtained when making use of a combined dataset. We propose a new method based on latent class modelling that estimates the number of classification errors in the multiple sources, and simultaneously takes impossible combinations with other variables into account. Furthermore, we use the latent class model to multiply impute a new variable, which enhances the quality of statistics based on the combined dataset. The performance of this method is investigated by a simulation study, which shows that whether the method can be applied depends on the entropy R^2 of the LC model and the type of analysis a researcher is planning to do. Furthermore, the method is applied to a combined dataset from Statistics Netherlands.

1 Introduction

National Statistical Institutes (NSIs) often use large datasets to estimate population tables on many different aspects of society. A way to create these rich datasets as efficiently and cost effectively as possible is by utilizing already available register data. This has several advantages. First, data does not have to be collected twice, saving collection and processing costs as well as reducing the burden on the respondents. Second, registers often contain very specific information that could not have been collected by surveys (Zhang, 2012). Third, statistical figures can be published more quickly, as conducting surveys can be time consuming. However, when more information is required than already available, registers can be supplemented with survey data (De Waal, 2015). Caution is then advised as surveys likely contain classification errors. Datasets consisting of a combination of registers and surveys are used by, among others, the Innovation

Panel (Understanding Society, 2016), the Millennium Cohort Study (Institute of Education, 2012), the Avon Longitudinal Study of Parents and Children (Ness, 2004), the System of Social Statistical Databases of the Netherlands and the 2011 Dutch Census (Schulte Nordholt et al., 2014).

When using registers for research, we should be aware that they are collected for administration purposes and can contain classification errors. This may be due to mistakes made when entering the data, delays in adding data to the register (Bakker et al., 2009) or differences between the variables being measured in the register and the variable of interest (Groen, 2012). This means that both registers and surveys can contain classification errors. In this paper, we classify classification errors as visibly or invisibly present. These errors may be resolved by making use of new information that is provided by the combined dataset. For instance, invisibly present errors in surveys or registers can be resolved when responses on both are compared in the combined dataset using a latent variable model. We do this by using multiple indicators from different datasets that measure the same latent variable. The invisibly present errors are measured in this way using structural equation models (Bakker, 2012; Scholtus & Bakker, 2013), latent class models (Guarnera & Varriale, 2015; Oberski, 2015) and latent markov models (Pavlopoulos & Vermunt, 2013). When taking covariate information into account, some errors can be observed directly. These visibly present errors are commonly resolved using edit rules. Edit rules describe which combinations of values are (not) allowed. An example of a combination which is not allowed, is the score "own" on the variable *home ownership* and the score "yes" on the variable *rent benefit*. An incorrect combination of values can be replaced by a combination that adheres to the edit rules. Methods for finding and correcting visibly present errors are optimization solutions (Fellegi-Holt method for categorical data; branch-and-bound algorithm; adjusted branch-and-bound algorithm; nearest-neighbour imputation, De Waal et al., 2011, pp. 115-156) and multiple imputation solutions (nonparametric Bayesian multiple imputation, Si & Reiter, 2013).

The previously discussed solutions are tailored to only handle either visibly or invisibly present errors, they are not able the handle these errors simultaneously. Furthermore, a disadvantage of current methods for invisibly present errors is that they do not offer possibilities to take the errors into account in further statistical analyses; they only give an indication of the extent of the classification errors. In addition, uncertainty caused by both invisibly and visibly present errors is not taken into account when further statistical analyses are performed. An exception is the method developed by Manrique-Vallier & Reiter (2015), which simultaneously handles invisibly and visibly present errors using a mixture model in combination with edit rules. This method fixes the extent of invisibly present errors to a presumed known number. However, in practice the number of invisibly present errors is unknown and estimation of this number is desired.

We propose a new method that simultaneously handles invisibly and visibly present classification errors and takes them both into account when performing further statistical analyses. Invisibly present errors

are taken into account by comparing responses on indicators measuring the same latent variable within a combined dataset allowing the estimation of the number of invisibly present errors. Visibly present errors are handled by making use of relevant covariate information in a restricted Latent Class (LC) model. In the hypothetical cross table between the latent "true" variable and the restriction covariate, the cells containing impossible combinations are restricted to contain zero observations. To correct for both invisibly and visibly present errors and to take uncertainty into account when performing further statistical analyses, we make use of Multiple Imputation (MI). Because Multiple Imputation and Latent Class analysis are combined in this new method, the method will be further denoted as MILC.

In the following section, we describe the MILC method in more detail. In the third section, a simulation study is performed to assess the novel method. In the fourth section, we apply the MILC method on a combined dataset from Statistics Netherlands.

2 The MILC method

The MILC method takes visibly and invisibly present errors into account by combining Multiple Imputation (MI) and Latent Class (LC) analysis. Figure 1 gives a graphical overview of this procedure. The method starts with the original combined dataset. In the first step, m bootstrap samples are taken from the original dataset. In the second step, an LC model is created for every bootstrap sample. In the third step, m new empty variables are created in the original dataset. The m empty variables are imputed using the corresponding m LC models. In the fourth step, estimates of interest are obtained from the m variables and in the last step, the estimates are pooled using Rubin's rules for pooling. These five steps are now discussed in more detail.

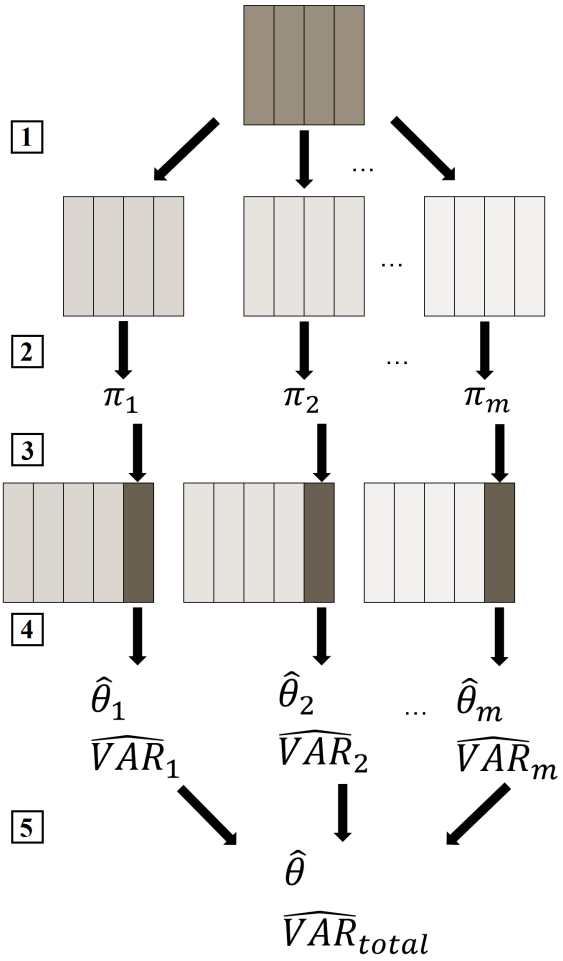

Figure 1: Procedure of latent class multiple imputation for a multiply observed variable in a combined dataset

The MILC method starts by taking m bootstrap samples from the original combined dataset. These bootstrap samples are drawn because we want the imputations we create in a later step to take parameter uncertainty into account. Therefore, we do not use one latent class model based on one dataset, but we use m latent class models based on m bootstrap samples of the original dataset (Van der Palm et al., 2016).

Although LC analysis is typically used as a tool for analyzing multivariate categorical response data (Vermunt & Magidson, 2004), we use LC analysis in a different context; namely to estimate both visibly and invisibly present classification errors in categorical variables. We have multiple datasets linked on a unit level, containing the same variable, which can be used as indicators measuring one latent variable. In this way, we estimate the invisibly present classification errors within the combined dataset. This latent variable can be seen as the "true variable", and is denoted by X. For example, we have l dichotomous indicators $(Y_1, ..., Y_l)$ measuring the variable *home ownership* (1="own", 2="rent") in multiple datasets linked on person level. Differences between the responses of a person are caused by what we described as invisibly present classification error in one (or more) of the indicators. Because the indicators all have an equal number of categories (denoted by $D_1, ..., D_l$), the number of categories (C) in the "true variable" X, is equal to $D_1 = ... = D_l$. A specific category is denoted by x, where $x = 1, ..., C$.

The LC model is based on three assumptions. The first assumption is that the probability of obtaining marginal response pattern $\mathbf{y}$, $P(\mathbf{Y} = \mathbf{y})$, is a weighted average of the C class-specific probabilities $P(\mathbf{Y} = \mathbf{y}|X = x)$:

$$P(\mathbf{Y} = \mathbf{y}) = \sum_{x=1}^{C} P(X = x)P(\mathbf{Y} = \mathbf{y}|X = x). \tag{1}$$

If we have three indicators measuring home ownership, the response pattern $\mathbf{y}$ can for example be "own", "own", "rent". $P(X = x)$ denotes the proportion of units belonging to category x in the "true variable", for example to the category "own". The second assumption is that the observed indicators are independent of each other given an individual's score on the latent "true variable". This means that when a mistake is made when filling in a specific question in a survey, this is unrelated to what is filled in for the same question in another survey or in a register. This is called the assumption of local independence,

$$P(\mathbf{Y} = \mathbf{y}|X = x) = \prod_{l=1}^{L} P(Y_l = y_l|X = x). \tag{2}$$

Combining equation 1 and equation 2 yields the following model for response pattern $P(\mathbf{Y} = \mathbf{y})$:

$$P(\mathbf{Y} = \mathbf{y}) = \sum_{x=1}^{C} P(X = x) \prod_{l=1}^{L} P(Y_l = y_l|X = x). \tag{3}$$

In equation 3, only the indicators are used to estimate the likelihood of being in a specific true category. However, it is also possible to make use of covariate information to estimate the LC model. Often, this is even necessary to obtain model identification. The third assumption we make here is that the measurement errors are independent of the covariates. For example, a covariate which can help in identifying whether someone owns or rents a house is *marital status*, this covariate is denoted by Q and can be added to equation 3.

$$P(\mathbf{Y} = \mathbf{y}|Q = q) = \sum_{x=1}^{C} P(X = x|Q = q) \prod_{l=1}^{L} P(Y_l = y_l|X = x), \tag{4}$$

Covariate information can also be used to impose a restriction to the model, to make sure that the model does not create a combination of a category of the "true" variable and a score on a covariate that is in practice impossible. For example, when an LC model is estimated to measure the variable *home ownership* using three indicator variables, and a covariate (denoted by Z) measures *rent benefit*, the impossible combination of owning a house and receiving rent benefit should not be created.

In this paper, we distinguish between three different models to take impossible combinations into account when applying the MILC method. In the first model, only indicators and covariates without restrictions are in the model, we call this the *unconditional model*. Note that while other covariates are found in this model, only the restriction covariate is left out. This model is equal to equation 4. In the second model, we also take the restriction covariate Z into account without explicitly mentioning the restriction itself, we call this the *conditional model*:

$$P(\mathbf{Y} = \mathbf{y}|Q = q, Z = z) = \sum_{x=1}^{C} P(X = x|Q = q, Z = z) \prod_{l=1}^{L} P(Y_l = y_l|X = x), \tag{5}$$

In the third model, we use a restriction to fix the cell proportion of the impossible combination to 0, we call this the *restricted conditional model*. In the example where Z measures *rent benefit*, and the latent "true" variable measures *home ownership*, the imposed restriction can be:

$$P(X = \text{own}|Z = \text{rent benefit}) = 0. \tag{6}$$

By using such a restriction, we can take impossible combinations with other variables into account, while we estimate an LC model for the "true" variable.

The next step is to impute latent "true" variable X. This imputed variable is denoted by W. When estimating W, uncertainty caused by the classification errors should be correctly taken into account. Therefore, multiple imputation is used to estimate W. m empty variables $(W_1, ..., W_m)$ are created and imputed by drawing one of the categories using their posterior membership probabilities from the m LC models. Poste-

rior membership probabilities are obtained by applying Bayes' rule to equation 4, equation 5 or equation 6. For example, the posterior membership probabilities for the *conditional model* are obtained by:

$$P(X = x|Y = y, Q = q, Z = z) = \frac{P(X = x|Q = q, Z = z) \prod_{l=1}^{L} P(Y_l = y_l|X = x, Q = q, Z = z)}{\sum_{x=1}^{C} P(X = x|Q = q, Z = z) \prod_{l=1}^{L} P(Y_l = y_l|X = x, Q = q, Z = z)}. \quad (7)$$

With the *restricted conditional model*, we wanted to make sure that cases were not assigned to categories on the latent "true" variable which would result in impossible combinations with scores on other variables, such as the combination "rent benefit" × "own". Therefore, the restriction set in equation 6 is also used here.

After we created m variables by imputing them using the posterior membership probabilities obtained from the m LC models, the estimates of interest can be obtained. This can be, for example, a cross table between imputed "true" variable W and covariate Z. The m estimates can now be pooled by making use of the rules defined by Rubin for pooling (Rubin, 1987, p.76). The pooled estimate is obtained by

$$\hat{\theta} = \frac{1}{m} \sum_{i=1}^{m} \hat{\theta}_i. \quad (8)$$

The total variance is estimated as

$$\mathrm{VAR}_{\mathrm{total}} = \overline{\mathrm{VAR}}_{\mathrm{within}} + \mathrm{VAR}_{\mathrm{between}} + \frac{\mathrm{VAR}_{\mathrm{between}}}{m}, \quad (9)$$

where $\overline{\mathrm{VAR}}_{\mathrm{within}}$ is the within imputation variance and $\mathrm{VAR}_{\mathrm{between}}$ is the between imputation variance.

Furthermore, $\overline{\mathrm{VAR}}_{\mathrm{within}}$ is calculated by

$$\overline{\mathrm{VAR}}_{\mathrm{within}} = \frac{1}{m} \sum_{i=1}^{m} \mathrm{VAR}_{\mathrm{within}_i} \quad (10)$$

and $\mathrm{VAR}_{\mathrm{between}}$ is calculated by

$$\mathrm{VAR}_{\mathrm{between}} = \frac{1}{m-1} \sum_{i=1}^{m} (\hat{\theta}_i - \hat{\theta})(\hat{\theta}_i - \hat{\theta})'. \quad (11)$$

Besides the uncertainty caused by missing or conflicting data, $\mathrm{VAR}_{\mathrm{between}}$ also contains parameter uncertainty, which was taken into account by the bootstrap performed in the first step of the MILC method.

3 Simulation

3.1 Simulation approach

To empirically evaluate the performance of MILC, we conducted a simulation study using R (R Core Team, 2014). We start by creating an infinite population using Latent Gold, containing five variables: three dichotomous indicators (Y_1, Y_2, Y_3) measuring the latent dichotomous variable (X); one dichotomous covariate (Z) which has an impossible combination with a score of the latent variable; one other dichotomous covariate (Q). Datasets are generated by making use of the restricted conditional model, and variations are made corresponding to the scenarios described in the following sections.

When evaluating an imputation method, the relation between the imputed latent variable and other variables should be preserved, since these relations might be the subject of research later on. When investigating the performance of MILC, there are two relations we are particularly interested in. We are interested in the relation between the imputed latent variable W and the covariate Z which has an impossible combination with a score on the latent variable. The four cell proportions of the 2×2 are denoted by: $W1 \times Z1$, $W2 \times Z1$, $W1 \times Z2$ and $W2 \times Z2$. The cell $W1 \times Z2$ is the impossible combination and should contain 0 observations. We compare the cell proportions of a 2×2 table of the population latent variable X and Z with the cell proportions of a table of the imputed latent variable W and Z from the samples. Furthermore, we are interested in the relation between W and covariate Q. To investigate this relation, we compare the intercept and the coefficient of a logistic regression of the latent population variable X on Q with the intercept and coefficient of the logistic regression of the imputed W and Q.

To investigate these relations, we look at three performance measures. First, we look at the bias of the estimates of interest. The bias is equal to the difference between the average estimate over all replications and the population value. Next, we look at the coverage of the 95% confidence interval. This is equal to the proportion of times that the population value falls within the 95% confidence interval constructed around the estimate over all replications. To confirm that the standard errors of the estimates were being properly estimated, the ratio of the average standard error of the estimate over the standard deviation of the estimates was also examined.

We expect the performance of MILC to be influenced by the measurement quality of the indicators, the marginal distribution of covariates Z and Q, the sample size and the number of multiple imputations. The quality of the indicators is represented by classification probabilities. They represent the probability of a specific score on the indicator given the latent class. If the quality of the indicators is low, it will also be more difficult for MILC to assign cases to the correct latent classes. Classification probabilities of 0.90 and higher are considered realistic for population registers. The quality of survey data is considered to be lower,

therefore we also investigate classification probabilities of 0.70 and 0.80. The marginal distribution of Z, $P(Z)$, is also expected to influence the performance of MILC. A larger $P(Z = 2)$ for example can give more information to the latent class model to assign scores to the correct latent class. Sample size may influence the standard errors and thereby the confidence intervals. The performance of MILC can also depend on the number of multiple imputations. Investigation of several multiple imputation methods have shown that 5 imputations are often sufficient (Rubin, 1987). However, with complex data, it can be the case that more imputations are needed. As a result, the simulation conditions can be summarized as follows:

- Classification probabilities: 0.70; 0.80; 0.90; 0.95; 0.99.
- $P(Z = 2)$: 0.01; 0.05; 0.10; 0.20.
- Sample size: $1,000$; $10,000$.
- Logit coefficients of X regressed on Q of $\log(0.45/(1-0.45) = -0.2007$, $\log(0.55/(1-0.55) = 0.2007$ and $\log(0.65/(1-0.65) = 0.6190$ corresponding to an estimated odds ratio of 0.81, 1.22 and 1.86. The intercept was fixed to 0
- Number of imputations: 5; 10; 20.

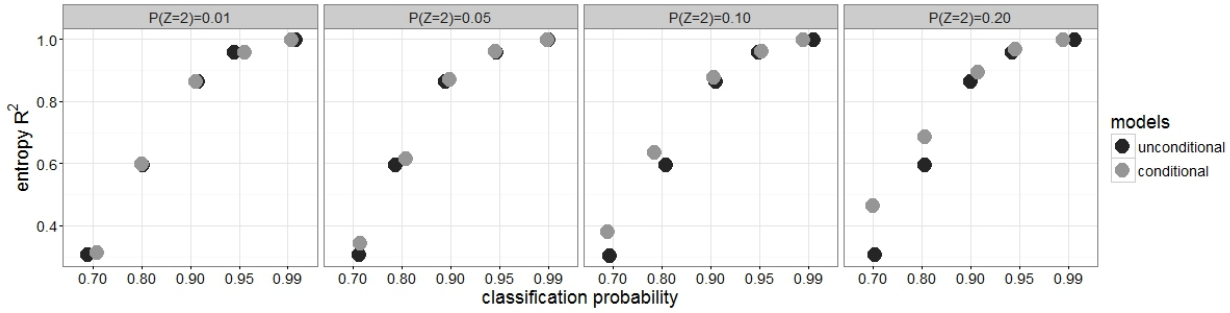

Figure 2: Entropy R^2 of the unconditional and conditional model with different values for the classification probability and $P(Z = 2)$. The restricted conditional model has the same entropy R^2 as the conditional model because the models contain the same variables

To illustrate the measurement quality corresponding to different conditions, Figure 2 shows the entropy R^2 of the models under different values for $P(Z = 2)$ and classification probabilities. The entropy indicates how well one can predict class membership based on the observed variables, and is measured by:

$$EN(\alpha) = -\sum_{i=1}^{N}\sum_{c=1}^{C}\alpha_{ic}log\alpha_{ic}, \tag{12}$$

where α_{ic} is the estimated probability that observation i is a member of class c. Rescaled with values between 0 and 1, entropy R^2 is measured by

$$R^2 = 1 - \frac{EN(\alpha)}{N \log C}, \tag{13}$$

where 1 means perfect prediction (Dias & Vermunt, 2008). The *conditional* and the *restricted conditional model* have the same entropy R^2 because these models consist of the same variables. All models with classification probabilities of 0.90 and above, have a high entropy R^2 and are able to predict class membership well. When the classification probabilities are 0.70, the entropy R^2 is especially low. However, for the conditional and the restricted conditional model, the entropy R^2 under classification probability 0.70 increases as $P(Z = 2)$ increases. A larger $P(Z = 2)$ means that covariate Z contains more information for predicting class membership. Because covariate Z is not in the *unconditional model*, it makes sense that entropy R^2 remains stable for different values for $P(Z = 2)$ under this model. Furthermore, Figure 2 demonstrates that the performance of MILC is evaluated over an extreme range of entropy R^2 values and gives an indication of what we can expect from the MILC method under different simulation conditions.

3.2 Simulation results

In this section we discuss our simulation results in terms of bias, coverage of the 95% confidence interval, and the ratio of the average standard error of the estimate over the standard deviation of the estimates. We do this in three sections. In the first section we discuss the 2×2 of the imputed latent variable W and restriction covariate Z. In the second section, we investigate the relation between imputed latent variable W and covariate Q. In the third section we investigate the influence of m, the number of bootstrap samples and multiple imputations. In the simulation results discussed in the first two sections, we used $m = 5$. When investigating the different simulation conditions, we focus on the performance of the three models discussed, the *unconditional model*, the *conditional model* and the *restricted conditional model*.

3.2.1 The relation of imputed latent variable W with restriction covariate Z

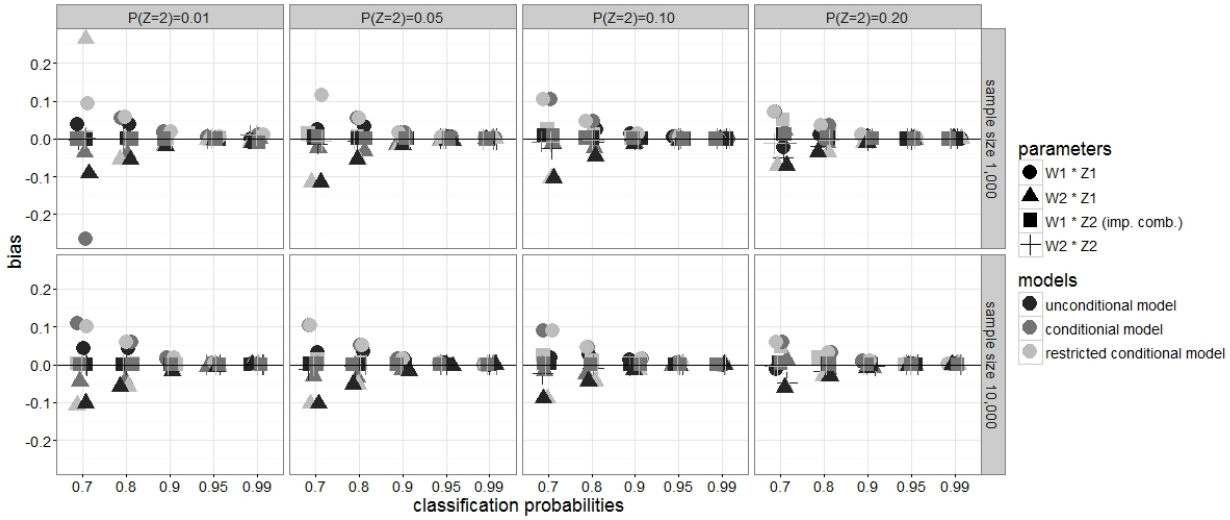

Figure 3: Bias of the 4 cell proportions of the 2×2 table of imputed latent "true" variable W and covariate Z for the unconditional, conditional and restricted conditional model with different values for the classification probabilities, $P(Z = 2)$ and sample size, $m = 5$.

When we investigate the results in terms of bias in Figure 3, it is immediately clear that the conditional model and the restricted conditional model produce bias when the classification probabilities of the indicators are 0.70. When the classification probabilities are 0.80, we still detect some bias, but already a lot less compared to the 0.70 condition. At classification probabilities of 0.90, 0.95 or 0.99, there is no detectable bias. This trend is not influenced by sample size and we see a decrease in the amount of bias for the conditional and the restricted conditional model when $P(Z = 2)$ increases. The trends we detect for bias over the different conditions coincide with the trends we saw in Figure 2 for the entropy R^2. In the conditions where the entropy R^2 is high, the bias is very small. For the conditions when we detect bias, where the classification probabilities are 0.70 or 0.80, the entropy R^2 was also low. Especially the bias for the conditional and the restricted conditional model is large. Furthermore, the bias in these models decreases as $P(Z = 2)$ increases, which corresponds to an increasing entropy R^2.

In general, the bias of the conditional and the restricted conditional model is larger than the bias of the unconditional model when the classification probability is 0.70. This is caused by the characteristics of covariate Z, which is part of the conditional and the restricted conditional model, but not of the unconditional model. When classification probabilities are low, more individuals are assigned to the wrong cluster. With covariate Z, the model is able to detect the individuals that are in class $W1$ and that have the score $Z2$ and places them in the correct cluster, $W2$. However, individuals that are in $W2$ while they should be in

$W1$ cannot be detected because Z gives no information to correct them. The correction only goes one way, so $W1$ becomes larger while $W2$ becomes smaller, introducing bias. When the classification probabilities are higher, less individuals are placed in the wrong cluster and need to be corrected by Z. It should also be mentioned that the bias is very symmetric. This makes sense, because the cell proportions are highly dependent on each other. For example, if the cell proportion of $W1 \times Z1$ is too large, $W2 \times Z1$ is very likely to be too small. Note that cell $W1 \times Z2$ (representing the impossible combination) has a bias of exactly 0 in model 3 under all conditions. Because we fixed this cell to 0, it is not possible for this cell proportion to vary, so no bias and no variance are calculated here.

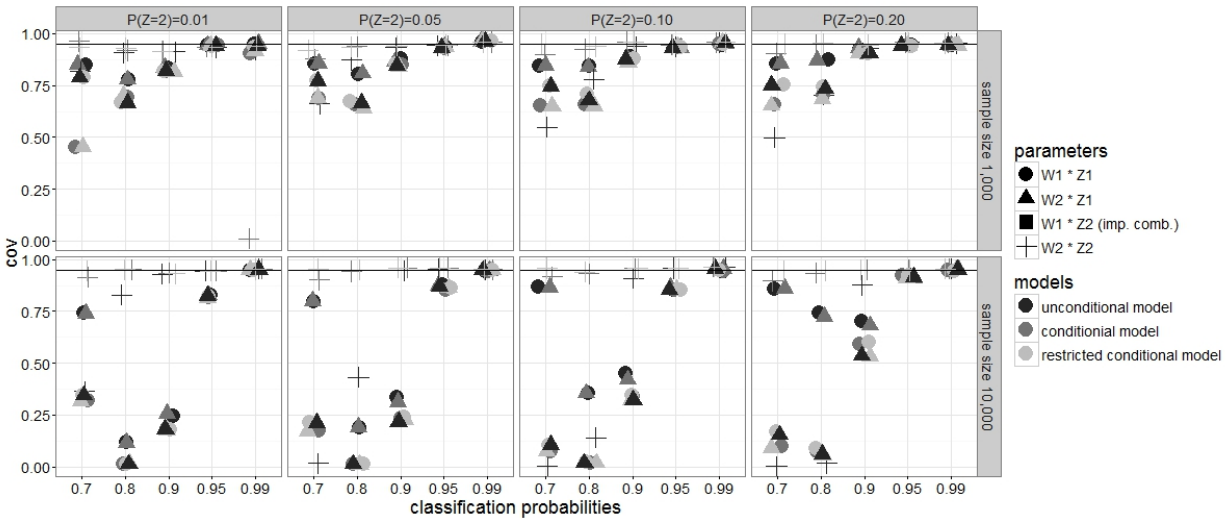

Figure 4: Coverage of the 95% confidence interval of the 4 cell proportions of the 2×2 table of imputed latent "true" variable W and covarate Z for the unconditional, conditional and restricted conditional model with different values for the classification probabilities, $P(Z = 2)$ and sample size, $m = 5$. Coverage for $W1 \times Z2$ are not shown, the confidence intervals cannot be properly estimated due to the fact that the estimates are (almost) 0.

When investigating the results for coverage of the 95% confidence intervals around the cell proportions, we see that the results differ for the different sample sizes. This is caused by the fact that even though the bias is not influenced by the sample size, the standard errors and therefore the confidence intervals are. Confidence intervals of biased estimates are therefore less likely to contain the population value. Furthermore, if the classification probabilities are larger, individuals are more likely to end up in the correct cluster, which also results in less variance, resulting in smaller confidence intervals. Confidence intervals for the impossible combination $W1 \times Z2$ cannot properly be estimated as the cell proportions are (very close to) 0. Therefore,

they are not shown in Figure 4.

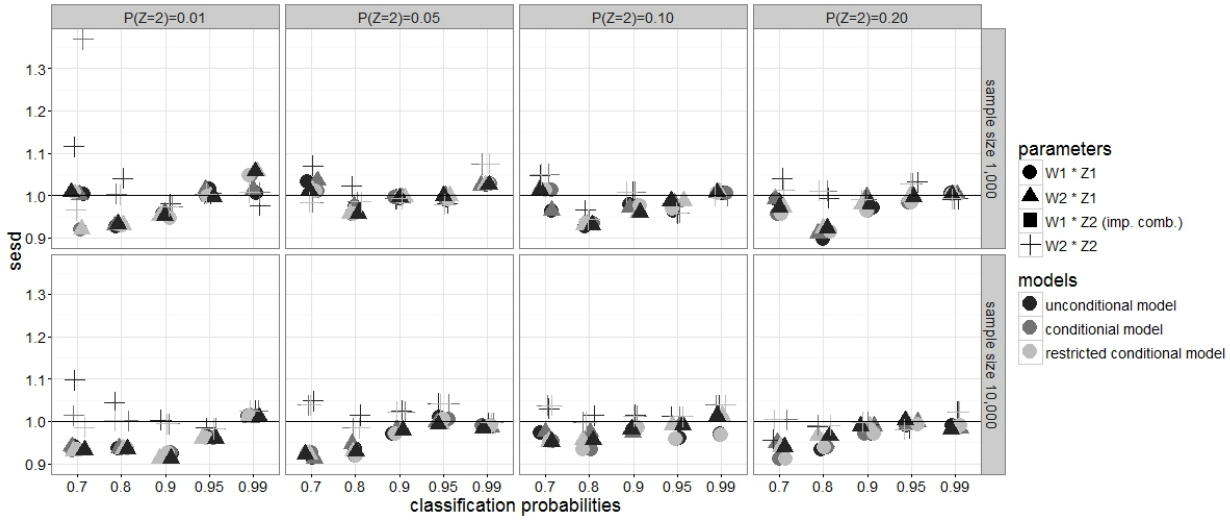

Figure 5: $se/sd(\hat{\theta})$ of the 4 cell proportions of the 2×2 table of imputed latent "true" variable W and covariate Z for the unconditional, conditional and restricted conditional model with different values for the classification probabilities, $P(Z = 2)$ and sample size, $m = 5$. Results for $W1 \times Z2$ are not shown, the standard errors cannot be properly estimated due to the fact that the estimates are (almost) 0.

The ratio of the average standard error of the estimate over the standard deviation of the simulated estimates tells us whether the standard errors of the estimates are properly estimated. The conclusions we can draw here correspond to what we saw in the coverage results. Here we also see that for model 1 and 2, we get unreliable results for cell proportion $W1 \times Z2$. Again, $se/sd(\hat{\bar{\theta}})$ cannot be calculated for cell proportion $W1 \times Z2$, because of the boundary issues previously discussed. All other values in all models are very close to 1, with a more variation when the classification probabilities are 0.70. We may thus conclude that the standard errors of the estimates are properly estimated.

Overall, estimating cross tables containing impossible combinations of an imputed latent variable and a restriction covariate can be done when the LC model of the combined dataset has an entropy R^2 of 0.90, or when the sample size is large, an entropy R^2 of 0.95.

3.2.2 Relationship between imputed latent variable W and covariate Q

In the simulation results discussed in section 4.1, the relation between imputed latent variable W and covariate Z containing an impossible combination was investigated. In the different models discussed here, there was also another covariate, Q. Note that covariate Z is still in the models as well, but we now focus on the relation between W and Q. More specifically, we investigate three different strengths of relations

by using intercepts of 0 and logit coefficients of W regressed on Q of -0.2007; 0.2007; 0.6190. Because the intercept is 0 in all conditions, we focus on the coefficients of Q when investigating the simulation results.

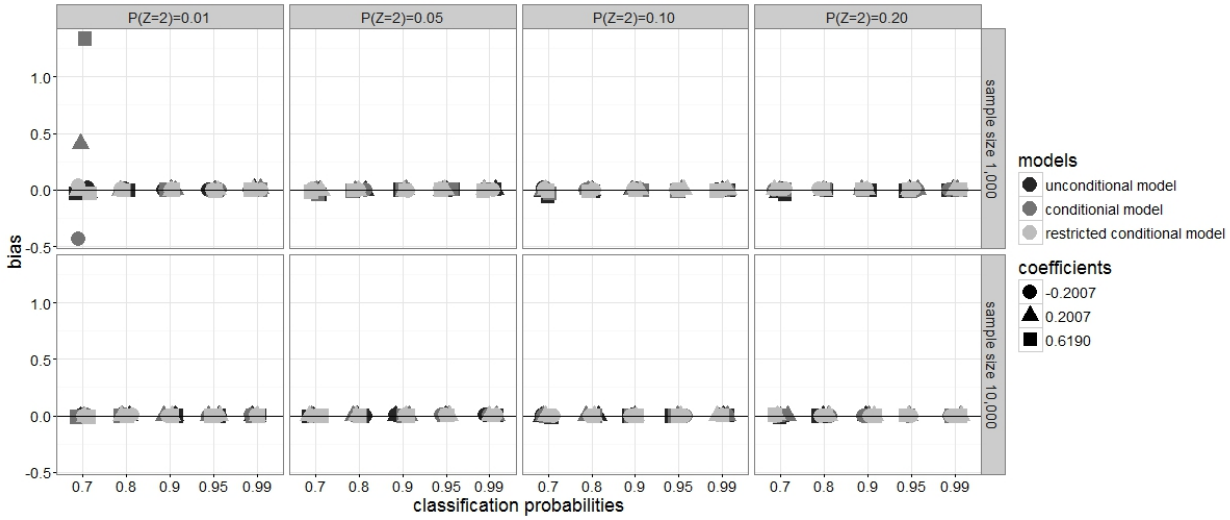

Figure 6: Bias of the logistic regression coefficient of imputed latent "true" variable W on covariate Q for the unconditional, conditional and restricted conditional model with different values for the coefficient, the classification probabilities, $P(Z = 2)$ and sample size.

In Figure 6 we see that the bias is almost 0 in all conditions, except when the sample size is $1,000$, the classification probabilities are 0.70, $P(Z = 2) = 0.01$ and the conditional model is used. Here we find large bias for all the different logistic regression coefficients, especially when the logit coefficient is 0.6190.

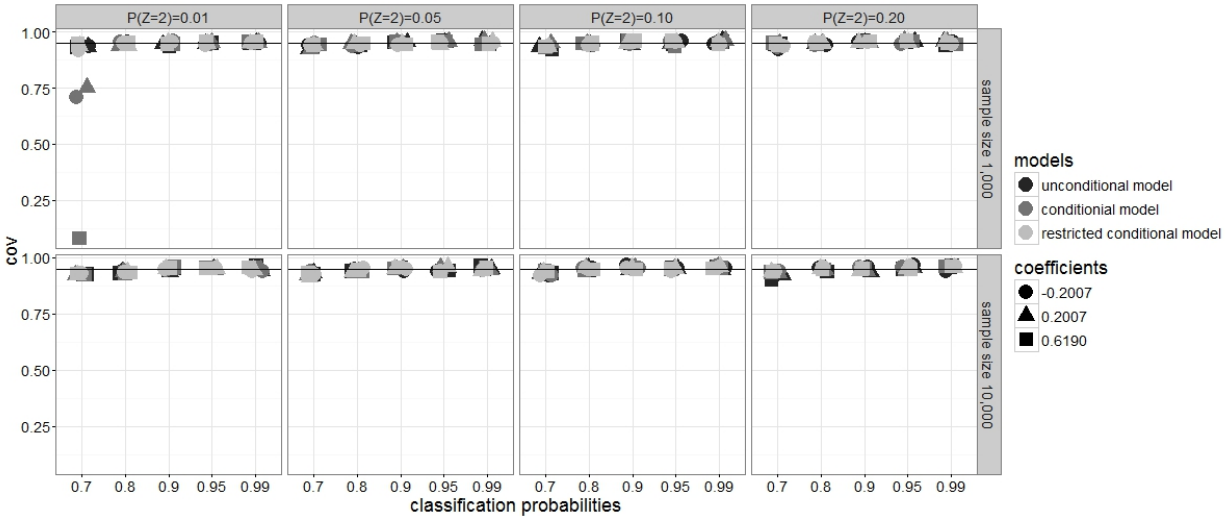

Figure 7: Coverage of the 95% confidence interval of the logistic regression coefficient of imputed latent "true" variable W on covariate Q for the unconditional, conditional and restricted conditional model with different values for the coefficients, the classification probabilities, $P(Z = 2)$ and sample size.

When investigating the results in terms of coverage of the 95 % confidence interval, as can be seen in Figure 7, we can draw comparable conclusions as we did in terms of bias. In the condition with sample size $1,000$, classification probabilities of 0.70 and $P(Z = 2) = 0.01$, we see undercoverage for all three models with different logistic regression coefficients investigated, and most strongly for the logit coefficient of 0.6190 under the conditional model.

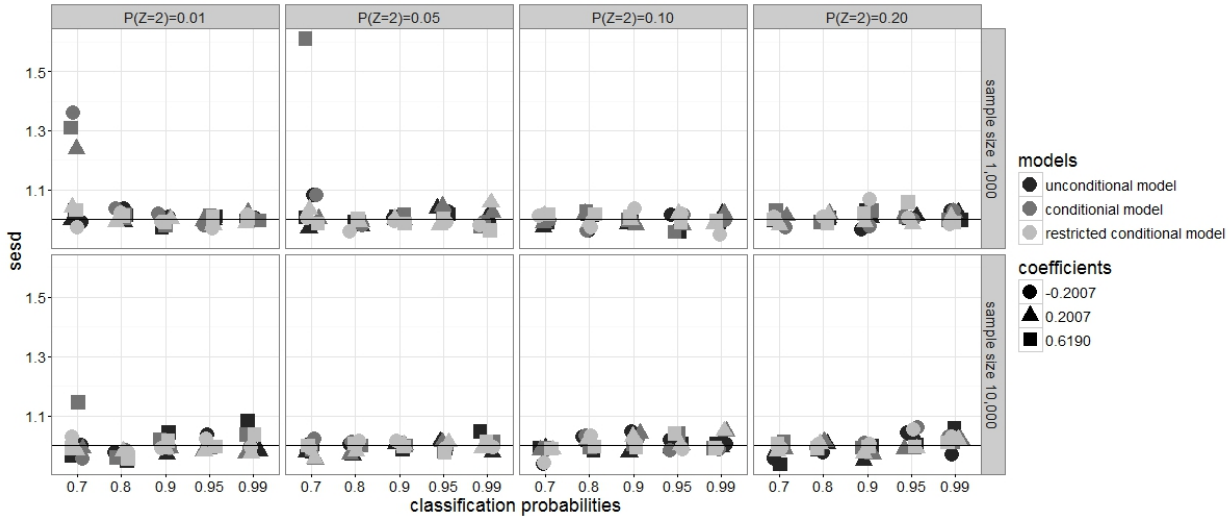

Figure 8: $se/sd(\hat{\theta})$ of the logistic regression coefficient of imputed latent "true" variable W on covariate Q for the unconditional, conditional and restricted conditional model with different values for the coefficients, the classification probabilities, $P(Z = 2)$ and sample size.

In Figure 8 we see results comparable to what we saw for the bias and coverage of the 95% confidence interval. The standard errors are overestimated under the conditions with sample size $1,000$, classification probabilities of 0.70 and $P(Z = 2) = 0.01$ with the different logit coefficients under the conditional model. The standard errors are also overestimated when $P(Z = 2) = 0.05$ and the logit coefficient is 0.6190 under the conditional model. We also see an overestimation when the sample size is $10,000$, classification probabilities are 0.70, $P(Z = 2) = 0.01$, logit coefficient is 0.6190 and the conditional model is used. For the other simulation conditions we see that the standard errors are properly estimated.

Overall, estimates can be obtained in terms of the relation of an imputed latent variable and a covariate when the LC model of the combined dataset has an entropy R^2 of 0.60 or larger.

3.2.3 Number of imputations

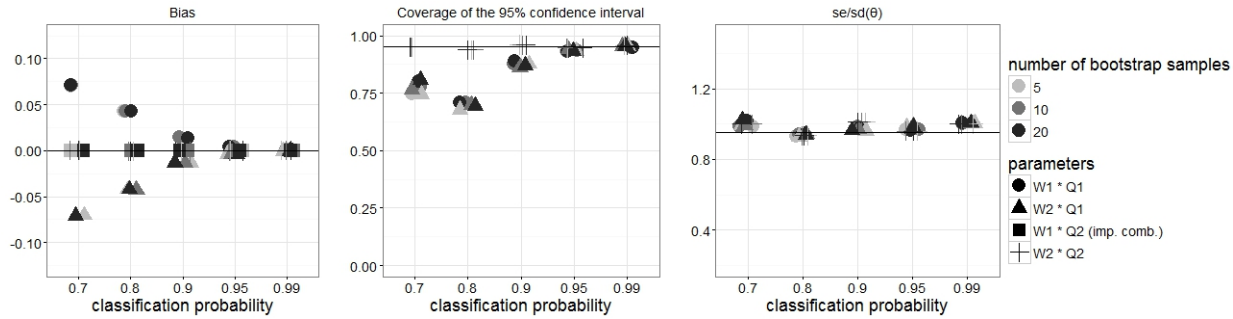

Figure 9: Bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of 4 cells in the 2×2 table of covariate Z imputed latent "true" variable W, data created and evaluated with the restricted conditional model and 5,10 or 20 of bootstrap samples and imputations. The sample size is $1,000$ and $P(Z = 2) = 0.10$.

When we investigate the effect of the number of bootstrap samples, we see that overall these effects are very small. The coverage of the 95% confidence interval and the $se/sd(\bar{\hat{\theta}})$ remain very much alike when the number of bootstrap samples changes. Bias is very comparable under the higher classification probabilities as well. However, for the lower classification probability, 0.70, the bias increases when the number of bootstrap samples becomes larger. This is the only condition in which we already detected some severe bias.

4 Application

4.1 Data

Home ownership is an interesting variable for social research. It has been related to a number of properties, such as inequality (Dewilde & Decker, 2016), employment insecurity (Lersch & Dewilde, 2015) and government redistribution (André & Dewilde, 2016). Therefore, we apply the MILC method on a combined dataset to measure home ownership. This combined dataset consists of data from the LISS (Longitudinal Internet Studies for the Social sciences) panel from 2013, which is administered by CentERdata (Tilburg University, The Netherlands) and a register from Statistics Netherlands from 2013. From this combined dataset, we use two variables indicating whether a person is a home-owner or rents a home (or "other") as indicators for the imputed "true" latent variable *home-owner/renter or other*. The combined dataset also contains a variable measuring whether someone receives rent benefit ("huurtoeslag") from the government. A person can only receive rent benefit if this person rents a house. In a cross-table between the imputed latent variable *home-owner/renter* and *rent benefit*, there should be 0 persons in the cell "home-owner $\times$

receiving rent benefit". If people indeed receive rent benefit and own a house, this could be interesting for researchers as well. However, we assume this to be measurement error, and therefore want this specific cell to contain zero persons. Research has previously been done the relation between home ownership and marital status (Mulder, 2006). A research question here could be whether married individuals more often live in a house they own compared to non-married individuals. Therefore, a variable indicating whether a person is married or not is included in the latent class model as a covariate. The three datasets used to combine the data are discussed in more detail below:

- **Registration of adresses and buildings (BAG):** A register containing data on addresses and buildings originating from municipalities from 2013. We obtained register information from persons who filled in the LISS studies and who declared that we are allowed to combine their survey information with registers. In total, this left us with 3011 individuals. From the BAG we used a variable indicating whether a person "owns"/"rents"/""other" the house he or she lives in. Because our research questions mainly relate to home-owners, we recoded this variable into "owns"/"rents or other".

- **LISS background study:** A survey on general background variables from January 2013. From this survey we also have 3011 individuals. We used the variable *marital status*, indicating whether someone is "married"/"separated"/"divorced"/"widowed"/"never been married". As we are only interested in whether a person is married or not, we recoded this variable in such a way that "married" and "separated" individuals are in the recoded "married" category (because separated individuals are technically still married) and the "divorced", "widowed" and "never been married" individuals are in the "not married" category. We also used a variable indicating whether someone is a "tenant"/"sub-tenant"/"(co-)owner"/"other". We recoded this variable in such a way that we distinguish between "(co-)owner" and "(sub-)tenant or other".

- **LISS housing study:** A survey on housing from June 2013. From this survey we used the variable *rent benefit*, indicating whether someone "receives rent benefit"/"the rent benefit is paid out to the lessor"/"does not receive rent benefit"/"prefers not to say". Because we are not interested in whether someone receives the rent benefit directly or indirectly, we recoded the first two categories into "receiving rent benefit". No one selected the option "prefers not to say". For this variable, we only have 779 observations. This is caused by the fact that another variable, indicating whether someone rents their house, was used as a selection variable. Dependent interviewing has been used here. Only the individuals indicating that they rent their house in this variable were asked if they receive rent benefit. This selection variable could also have been used as an indicator in our LC model. However, because of the strong relation between this variable and the rent benefit variable we decided to leave it out of

the model.

These datasets are linked on a unit level, and matching is done on person identification numbers. In addition, matching could also have been done on date, since the surveys were conducted at different time points within 2013. However, mismatches on dates are a source of measurement error, and are therefore left in for illustration purposes. Not every individual is observed in every dataset. This causes that some missing values are introduced when the different datasets are linked on a unit level. Full Information Maximum Likelihood was used to handle the missing values.

Table 1: Entropy R^2, classification probabilities for the indicators and marginal probabilities for the covariates for the unconditional, the conditional and the restricted conditional model. Note that the *rent benefit* variable takes information of 779 individuals into account and *marital status* variable of 3011.

			unconditional model	conditional model	restricted conditional model
entropy R^2			0.9334	0.9377	0.9380
classification probability	LISS background	P(rent|cluster 1)	0.8937	0.8938	0.9344
		P(own|cluster 2)	0.9997	0.9997	0.9992
	BAG register	P(rent|cluster 1)	0.9501	0.9500	0.9496
		P(own|cluster 2)	0.9749	0.9749	0.9525
P(rent benefit)				0.3004	0.3004
P(married)			0.5284	0.5284	0.5284

The MILC method is applied to impute the latent variable *home owner/renter* by using two indicator variables and two covariates. The *unconditional model*, the *conditional model* and the *restricted conditional model* are applied. In Table 1 classification statistics about the models are given. They give an indication of how we can compare the results of these models to the information we obtained in the simulation study. Both the entropy R^2 and the classification probabilities are comparable to conditions we tested in the simulation study and in which the MILC method appeared to work very well. The classification probabilities for the LISS background survey and the BAG register indicate that they both have a high quality, but are both imperfect. Furthermore, P(married) and P(rent benefit) cannot be compared directly to the set up of the simulation study, but information provided by the covariates is taken into account in the entropy R^2.

For the two variables measuring home ownership, we can see from the cell totals in Table 2 whether individuals who say to own their home, also receive rent benefit, which is not allowed. However, in practice these discrepancies can be caused by the fact that people make mistakes when filling in a survey, or for example because people were moving during the period the surveys took place. Furthermore, the total number of individuals who can be found in the table of the LISS background study are only 779, and for the BAG register 772. This is because only the people indicating that they rented a house in the LISS Housing study

Table 2: The first block represents the (pooled) marginal proportions of the variable *own/rent*. The second block represents the (pooled) proportions of the variable *own/rent* for persons receiving rent benefit. The third block represents the (pooled) proportions of the variable *own/rent* for persons not receiving rent benefit. Within each block, the first two rows represent the BAG register and the LISS background survey, used as the indicators for the MILC method. The last three rows represent the three different models used to apply the MILC method. For each proportion a (pooled) estimate and a (pooled) 95% confidence interval is given.

	P(own)		P(rent)	
	estimate	95% CI	estimate	95% CI
BAG register	0.6450	[0.6448; 0.6451]	0.3550	[0.3549; 0.3511]
LISS background	0.6830	[0.6829; 0.6832]	0.3170	[0.3168; 0.3171]
unconditional	0.6422	[0.6420; 0.6423]	0.3578	[0.3577; 0.3580]
conditional	0.6505	[0.6503; 0.6506]	0.3495	[0.3494; 0.3497]
restricted conditional	0.6591	[0.6590; 0.6593]	0.3409	[0.3407; 0.3410]
	P(own $\times$ rent benefit)		P(rent $\times$ rent benefit)	
	estimate	95% CI	estimate	95% CI
BAG register	0.0051	[0.0001; 0.0102]	0.2953	[0.2632; 0.3273]
LISS background	0.0104	[0.0032; 0.0175]	0.2889	[0.2568; 0.3209]
unconditional	0.0013	[0.0007; 0.0018]	0.2940	[0.2934; 0.2945]
conditional	0.0064	[-0.0263; 0.0391]	0.2888	[0.2561; 0.3215]
restricted conditional	0.0000	-	0.2953	[0.2624; 0.3281]
	P(own $\times$ no rent benefit)		P(rent $\times$ no rent benefit)	
	estimate	95% CI	estimate	95% CI
BAG register	0.0552	[0.0391; 0.0713]	0.6444	[0.6107; 0.6781]
LISS background	0.0285	[0.0167; 0.0403]	0.6723	[0.6391; 0.7054]
unconditional	0.0154	[0.0149; 0.0159]	0.6842	[0.6837; 0.6848]
conditional	0.0154	[-0.0173; 0.0481]	0.6842	[0.6515; 0.7169]
restricted conditional	0.0205	[-0.0123; 0.0534]	0.6791	[0.6462; 0.7119]

were asked the question whether they received rent benefit. For the LISS background study we see that 8 individuals are in the cell representing the impossible combination of owning a house and receiving rent benefit, and for the register 4 persons are. If we investigate the cell proportions estimated by the MILC method, we see that both the conditional and the unconditional model replicate the structure of the indicators very well, but that individuals are still assigned to the cell of the impossible combination. To get this correctly estimated, we need the restricted conditional model. The marginals of the variable *own/rent* (in the upper block of Table 2) for the different models are all very close to each other, and closer to the estimates in the BAG register than to the estimates of the LISS background study. Also note that individuals with missing values on the variable *rent benefit* are not taken into account in the 2×2 table of *rent benefit* $\times$ *own/rent*.

Table 3: The first two rows represent the BAG register and the LISS background survey, used as the indicators for the MILC method. The last three rows represent the three different models used to apply the MILC method. The columns represent the (pooled) estimate and 95% confidence interval (total) standard error of the intercept and the logit coefficient of the variable *owning/renting* a house.

	intercept		marriage	
	estimate	95% CI	estimate	95% CI
BAG register	2.4661	[2.2090; 2.7233]	-1.2331	[-1.3901; -1.0760]
LISS background survey	2.7620	[2.4896; 3.0343]	-1.3041	[-1.4678; -1.1405]
unconditional model	2.7229	[2.4601; 2.9858]	-1.4060	[-1.6688; -1.1431]
conditional model	2.7148	[2.4506; 2.9791]	-1.3751	[-1.6393; -1.1108]
restricted conditional model	2.8220	[2.5533; 3.0907]	-1.4159	[-1.6846; -1.1472]

Here we investigated whether marriage can predict home ownership. When we consider the BAG register, we see that the estimated odds of owning a home when not married are $e^{-1.2331} = 0.29$ times the odds when married. The exponentiated intercept (11.776) can be interpreted as the odds of owning a home when married. This relation is the same when different types of models are used to apply the MILC method.

5 Discussion

In this paper we introduced the MILC method, which combines latent class analysis and multiple imputation to obtain estimates for variables of which we had multiple indicators in a combined dataset. We distinguished between invisibly present and visibly present errors (commonly solved by edit restrictions), and argued the need for a method that takes them into account simultaneously. We evaluated the MILC method in terms of its ability to correctly take impossible combinations and relations with other variables into account. We assessed these relations by investigating the bias of $\bar{\hat{\theta}}$, coverage of the 95 % confidence interval, and $se/sd(\bar{\hat{\theta}})$ in different conditions in a simulation study. The performance of MILC appeared to be mainly dependent on the entropy R^2. We can conclude here that a different quality of the combined dataset is required for different types of estimates. To obtain correct estimates in terms of cross tables replicating a relation which contains an impossible combination, a high quality of the combined dataset is required in terms of an entropy R^2 of 0.90. Obtaining correct estimates in terms of parameters of a logistic regression model can already be done with data of a much lower quality, namely with an entropy R^2 of 0.60.

An example of a combined dataset containing data from the LISS panel and the BAG register were shown to have adequate entropy R^2 and decent results. Furthermore, we investigated the MILC method using three different types of models, the unconditional model, the conditional model and the restricted conditional model. All models can potentially be used when using the MILC method in practice. However,

if there are impossible combinations within the data that need to be taken into account, only the restricted conditional model is appropriate. In light of our main findings, the MILC method can be seen as an appropriate alternative to methods previously used for handling visibly and invisibly present errors. This was done either separately using latent variable models and edit rules, or simultaneously by Manrique-Vallier & Reiter (2015), by fixing the classification error rates a priori.

In an extension, more attention can be paid to the covariates. In the current approach, we assume that the covariates do not contain classification error. We could adapt the method in such a way that we assume a specific amount of classification error in the covariate, or we could use multiple indicators for the "true" latent covariate if available. Furthermore, variables corresponding to all relationships that a researcher wants to investigate should be included as covariates in the LC model. By adapting the three step method (Bakk et al., 2014) to make it applicable to the MILC method, we could also investigate relations of the imputed latent variable and other variables, not taken into account as covariates in the LC model when the MILC method was applied. Furthermore, investigation can be done on how the MILC method could handle linkage and selection errors.

In addition, MILC is developed to be used for indicators coming from both population registers and sample surveys. When the indicators only come from sample surveys, we can use the standard rules for pooling as defined by Rubin (1987). However, as soon as one of the indicators contains a complete population register, we can choose to either only impute the survey variables, and weight them to appropriately represent the population variables, or we can choose to impute both the survey and population variables, and use adjusted rules for pooling (Vink & van Buuren, 2014). We use these rules because all sampling variability is captured by the between imputation variance in this situation, so the within variance should be left out of the equation. We only used samples in the simulation, but it is important to be aware of necessary adjustments when population registers are used. The fact that this simulation study only considers dichotomous variables could potentially be seen as a drawback. However, we are convinced that the results we obtained by using variables with two categories, apply to variables with more categories as well, as this only influences the complexity of the model.

References

André, S., & Dewilde, C. (2016). Home ownership and support for government redistribution. *Comparative European Politics*, *14*(3), 319–348. Retrieved from `http://dx.doi.org/10.1057/cep.2014.31` doi: 10.1057/cep.2014.31

Bakk, Z., Oberski, D. L., & Vermunt, J. K. (2014). Relating latent class assignments to external variables: standard errors for correct inference. *Political analysis*, mpu003.

Bakker, B. (2012). Estimating the validity of administrative variables. *Statistica Neerlandica*, *66*(1), 8-17.

Bakker, B., et al. (2009). *Trek alle registers open!* Vrije Universiteit.

De Waal, T. (2015). Obtaining numerically consistent estimates from a mix of administrative data and surveys. *Statistical Journal of the IAOS*(Preprint), 1–13.

De Waal, T., Pannekoek, J., & Scholtus, S. (2011). *Handbook of statistical data editing and imputation* (Vol. 563). John Wiley & Sons.

Dewilde, C., & Decker, P. D. (2016). Changing inequalities in housing outcomes across western europe. *Housing, Theory and Society*, *33*(2), 121-161. Retrieved from `http://dx.doi.org/10.1080/14036096.2015.1109545` doi: 10.1080/14036096.2015.1109545

Dias, J. G., & Vermunt, J. K. (2008). A bootstrap-based aggregate classifier for model-based clustering. *Computational Statistics*, *23*(4), 643–659.

Groen, J. A. (2012). Sources of error in survey and administrative data: The importance of reporting procedures. *Journal of Official Statistics*, *28*(2), 173.

Guarnera, U., & Varriale, R. (2015). Estimation and editing for data from different sources. an approach based on latent class model. In *Emerging methods and data revolution.*

Institute of Education. (2012). Millennium cohort study: First survey, 2001-2003 [computer file]. colchester, essex: Uk data archive [distributor], sn: 4683 (11th Edition ed.) [Computer software manual]. doi: http://dx.doi.org/10.5255/UKDA-SN-4683-3

Lersch, P. M., & Dewilde, C. (2015). Employment insecurity and first-time homeownership: Evidence from twenty-two european countries. *Environment and Planning A*, *47*(3), 607-624. Retrieved from `http://epn.sagepub.com/content/47/3/607.abstract` doi: 10.1068/a130358p

Manrique-Vallier, D., & Reiter, J. P. (2015). Bayesian simultaneous edit and imputation for multivariate categorical data.

Mulder, C. H. (2006). Home-ownership and family formation. *Journal of Housing and the Built Environment*, *21*(3), 281–298. Retrieved from `http://dx.doi.org/10.1007/s10901-006-9050-9` doi: 10.1007/s10901-006-9050-9

Ness, A. R. (2004). The avon longitudinal study of parents and children (alspac)–a resource for the study of the environmental determinants of childhood obesity. *European journal of endocrinology*, *151*(Suppl 3), U141–U149.

Oberski, D. L. (2015). Estimating error rates in an administrative register and survey questions using a latent class model. In P. Biemer, B. West, S. Eckman, B. Edwards, & C. Tucker (Eds.), *Total survey error.* New York: Wiley.

Pavlopoulos, D., & Vermunt, J. (2013). Measuring temporary employment. do survey or register tell the truth? *Survey Methodology*, *41*(1), 197-214.

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys* (Vol. 81). John Wiley & Sons.

Scholtus, S., & Bakker, B. (2013). *Estimating the validity of administrative and survey variables through structural equation modeling: A simulation study on robustness* (Discussion paper). Statistics Netherlands.

Schulte Nordholt, E., Van Zeijl, J., & Hoeksma, L. (2014). *Dutch census 2011, analysis and methodology.* The Hague/Heerlen.

Si, Y., & Reiter, J. P. (2013). Nonparametric bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of Educational and Behavioral Statistics*, *38*(5), 499–521.

Understanding Society. (2016). Understanding society: Innovation panel, waves 1-7, 2008-2014. [data collection]. 6th edition [Computer software manual]. UK Data Service. Retrieved from `http://dx.doi.org/10.5255/UKDA-SN-6849-7`

Van der Palm, D. W., Van der Ark, L. A., & Vermunt, J. K. (2016). Divisive latent class modeling as a density estimation method for categorical data. *Journal of Classification*, 1–21. Retrieved from `http://dx.doi.org/10.1007/s00357-016-9195-5` doi: 10.1007/s00357-016-9195-5

Vermunt, J. K., & Magidson, J. (2004). Latent class analysis. *The sage encyclopedia of social sciences research methods*, 549–553.

Vink, G., & van Buuren, S. (2014). Pooling multiple imputations when the sample happens to be the population. *arXiv preprint arXiv:1409.8542*.

Zhang, L.-C. (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, *66*(1), 41–63. Retrieved from http://dx.doi.org/10.1111/j.1467-9574.2011.00508.x doi: 10.1111/j.1467-9574.2011.00508.x