Estimating the number of serious road injuries per vehicle type in the Netherlands using Multiple Imputation of Latent Classes

Laura Boeschoten, Ton de Waal, Jeroen Vermunt

November 20, 2018

Abstract

Statistics published by official agencies are often generated using population registries, which are likely to contain classification errors and missing values. A method that simultaneously handles classification errors and missing values is Multiple Imputation of Latent Classes (MILC). In this paper, the MILC method is applied to estimate the number of serious road injuries per vehicle type in the Netherlands and to stratify the number of serious road injuries per vehicle type into relevant subgroups using data from two registries. For this specific application, the MILC method is extended to handle the large number of missing values in the stratification variable 'region of accident' and to include more stratification covariates. After applying the extended MILC method, a multiply imputed dataset is generated that can be used to create statistical figures in a straightforward manner, and that incorporates uncertainty due to classification errors and missing values in the estimate of the total variance.

Keywords

Combined dataset; Missing values; Classification error; Latent class analysis; Multiple imputation

1 Introduction

When statistics are published by government or other official agencies, population registries are often utilized to generate these statistics. Here, caution is advised as population registries are collected for administrative purposes so they may not align conceptually with the target of interest. Furthermore, they are likely to contain process delivered classification errors. Another issue is that population registries are likely to not have registered every single unit in the population of interest, so the population registry is not complete.

An official agency dealing with the issues of classification errors and missing units in registers when generating statistics is the Institute for Road Safety Research (in Dutch Stichting Wetenschappelijk Onderzoek Verkeersveiligheid, abbreviated as SWOV). An important statistic SWOV publishes every year is the number of serious road injuries in the Netherlands. The number of serious road injuries is important because it is used to define the road safety target (Reurings & Stipdonk, 2011). To gain more insight in the total number of serious road injuries, it can be further stratified by vehicle type, injury severity and region (Reurings & Bos, 2012). When estimating the number of serious road injuries in the Netherlands, SWOV uses information from police and hospital registries. These registries contain classification errors and are incomplete. SWOV estimates the number of units missing in both registries by a method based on capture-recapture (Reurings & Stipdonk, 2011). However, a procedure to correct for classification errors and missing values within the observed cases has not been applied.

A method to simultaneously deal with classification errors and missing values within the observed cases is the recently proposed method Multiple Imputation of Latent Classes (MILC) (Boeschoten et al., 2017). The MILC method combines two existing statistical methods: multiple imputation and latent class analysis. To apply the MILC method, it is necessary to have multiple population registries that can be linked on a unit level. Both registries are required to contain identifier variables to link the information for a specific case in one registry to its information in the other registry. In such a combined dataset, variables are selected that measure the same construct but originate from the different registries. They are used as indicators of a latent variable of which it can be said that it contains the 'true scores' which is estimated using a latent class model. Information from the latent class model is then used to create multiple imputations of the 'true variable'. The multiply imputed datasets can be used to generate statistics of interest, graphs or frequency tables. Uncertainty due to classification errors and missing cases is reflected in the differences between the imputations and is incorporated in the estimate of the total variance (Rubin, 1987, p.76).

In this paper, the MILC method is applied on a linked dataset containing a police and a hospital registry, to estimate the number of serious road injuries per vehicle type. Next, two variables measuring vehicle type are used as indicators of a latent variable measuring the 'true' vehicle type. To stratify the serious road injuries into relevant groups, covariates are included in the latent class model.

A statistic that is currently not straightforward to estimate is the number of serious road injuries per vehicle type per region, because the variable 'region of accident' is only observed in the police registry and contains many missing cases. To estimate this statistic, the MILC method is extended in two ways. First, the MILC method is extended to simultaneously estimate two latent variables ('vehicle type' and 'region of accident'). For the latent variable 'vehicle type', two imperfectly measured indicators are specified. For the latent variable 'region of accident', one indicator (containing missing values) is assumed to be a perfect representation of the latent variable, next to a second, imperfectly measured, indicator. Second, the MILC method is extended to incorporate more covariates for investigating relevant stratifications in general. In the remainder of this paper, we refer to this as the 'extended MILC method'.

In the next section, a more detailed description of the data on which the extended MILC method is applied is given. In the third section, a detailed description is given of how the extended MILC method is applied to these datasets. In addition, an illustrative simulation study is performed in this section. Here, the results obtained after applying the extended MILC method are compared to results obtained after applying a more traditional hierarchical assignment procedure. In the fourth section, the output from the latent class model and the number of serious road injuries are discussed.

2 Background

The extended MILC method is applied on a unit linked dataset containing a police and a hospital registry. It is applied separately to datasets from 1994, 2009 and 2014 as the quality of the registries has changed substantially over time. In this section, the process of constructing these datasets is described and variables of interest are discussed in more detail.

For every year, units observed in the two sources are linked by using information on person and accident characteristics (Reurings & Stipdonk, 2009). Changes in registration systems over time influenced the success rate of the linking procedure. In addition, a weighting factor was determined for many of the individual cases (Bos et al., 2017).

2.1 Variables measuring 'Vehicle type'

As can be seen in Table 1, the variable 'Vehicle type' is observed in both the police and the hospital registry and has nine categories. The categories make a distinction between injuries caused by motorized vehicles (with an 'M' in the category label) and non-motorized vehicles (with an 'N' in the category label). For example, there is a category 'M-bicycle' and 'N-bicycle'. The difference between these categories is that for the category 'M-bicycle', the injured person was on a bike and got into an accident with a motorized vehicle, while for the category 'N-bicycle', the injured person was on a bike and there was no motorized vehicle involved in the accident. The distinction between motorized and non-motorized is important because it provides information on the cause of the injury. For example, when the number of injuries increase in the category 'N-bicycle', it can be caused by unsafe bicycle lanes. If the number of injuries increase in the category 'M-bicycle', it can be caused by a high speed limit on roads shared by cars and bicycles.

As shown in Table 1, many injuries were classified differently by the police and the hospital. In addition, it can also be seen that injuries in the 'non-motorized' ('N') categories are particularly often missing in the police registry, as the police is generally not involved in, for example, one-sided bicycle accidents. Also note that the category 'N-other' is not observed in the police registry at all.

2.2 Variables describing relevant subgroups

Besides estimating the number of serious road injuries per vehicle type, stratifications in relevant subgroups need to be made, such as age, gender, injury severity or region of accident. To be able to make such stratifications, the variables need to be included as covariates in the latent class model that is used to estimate 'true vehicle type'.

The reason for estimating the latent class model, is to create imputations for 'true vehicle type' for every observed case. To be able to stratify all cases, the covariates need to be observed completely as well. For the variables 'age', 'gender' and 'injury severity' this is the case. For the variable 'region of accident', this is a problem, as this variable is only observed in the police registry.

To solve the issue of missing values in the variable 'region of accident' the traditional MILC method is extended in such a way that missing values in the variable 'region of accident' are imputed simultaneously

Table 1: Cross-table between the variables measuring vehicle type originating from the police registry (colums) and from the hospital registry (rows) for the years 1994, 2009 and 2013. Note that there are no observations for the category 'Non motorized - other' in the police registry. Also note that 'NA' means 'missing value'

		NA	1.	2.	3.	4.	5.	6.	7.	9.	Total
1994											
	NA	-	561	245	318	122	42	137	90	14	1,529
	1. M car	918	2,596	11	72	12	22	25	2	1	3,659
	2. M moped	702	29	1,131	21	60	2	8	2	1	1,956
	3. M bicycle	397	40	70	1,111	2	1	53	25	4	1,703
	4. M motorcycle	347	16	41	2	633	3	0	0	0	1,042
	5. M other	450	408	106	104	35	50	116	8	2	1,279
	6. M pedestrian	421	128	37	231	4	5	537	5	5	1,373
	7. N bicycle	3,625	28	41	221	3	3	11	296	3	4,231
	8. N other	34	1	0	2	0	4	0	2	0	43
	9. N pedestrian	94	2	2	2	0	0	20	6	22	148
	Total	6,988	3,809	1,684	2,084	871	132	907	436	52	16,963
2009											
	NA	-	209	111	126	38	20	62	26	6	598
	1. M car	779	969	8	29	8	17	3	0	0	1,813
	2. M moped	1,117	4	611	10	23	20	2	0	0	1,787
	3. M bicycle	565	23	17	701	0	9	20	9	0	1,344
	4. M motorcycle	668	9	74	2	367	6	0	0	0	1,126
	5. M other	350	51	40	21	11	23	23	1	1	521
	6. M pedestrian	363	39	15	62	2	2	202	2	2	689
	7. N bicycle	6,369	17	22	161	2	4	5	144	4	6,728
	8. N other	99	0	2	4	0	0	0	4	1	110
	9. N pedestrian	136	0	1	4	0	0	6	8	16	171
	Total	10,446	1,321	901	1,120	451	101	323	194	30	14,887
2013											
	NA	-	59	29	33	15	36	11	5	1	189
	1. M car	877	566	3	1	4	65	3	0	0	1,519
	2. M moped	2,220	8	419	3	167	63	2	1	0	2,883
	3. M bicycle	944	4	11	451	0	155	10	7	0	1,582
	4. M motorcycle	69	0	10	0	21	3	0	0	0	103
	5. M other	556	18	8	1	19	27	4	0	0	633
	6. M pedestrian	392	2	3	30	0	64	123	0	1	615
	7. N bicycle	7,230	12	7	41	1	29	2	44	1	7,367
	8. N other	13	0	0	0	0	0	0	0	0	13
	9. N pedestrian	117	0	0	1	0	4	2	0	5	129
	Total	12,418	669	490	561	227	446	157	57	8	15,033

while the latent variable 'true vehicle type' is estimated. To create these imputations, information is used from the variable 'region of hospital', which is observed for the cases that contain missing values for the variable 'region of accident'. The two variables have a strong, but not perfect, relationship. For example, from the serious road injuries in 2013 of which the injured person was in a hospital in Groningen, 53 were also registered to have taken place in Groningen, while 12 of those accidents were registered to have taken place in Friesland, a neighbouring region of Groningen. There was also one person in a hospital in Groningen of which the accident was registered to be in Zuid-Holland, which is quite far away from Groningen (see Figure 1 for the regions of the Netherlands). A reason for this observation can be classification error in one of the registries or incorrect linkage of a case in the police registry to a case in the hospital registry (wrongfully assuming the cases contained the same person). However, it is also possible that this person indeed had a road accident in Zuid-Holland and was transfered to a hospital in Groningen because it was closer to the person's home or it could provide a form of specialized health care.

3 Applying the extended MILC method

In this section, it is described step-by-step how the extended MILC method is applied to estimate the number of serious road injuries per vehicle type in the Netherlands. The procedure of applying the MILC method starts with the dataset that is linked and processed as described in the previous section.

3.1 Bootstrapping for parameter uncertainty

In order to account for parameter uncertainty when applying the extended MILC method, we use a non-parametric bootstrap procedure. This involves creating M bootstrap samples by drawing observations from the observed data set with replacement. Subsequently, for each bootstrap sample, the latent class model of interest is estimated and the M imputations are created using the M sets of parameter values obtained. This is preferable over creating imputations based the maximum-likelihood estimates obtained with the observed data, which would imply ignoring the uncertainty regarding the estimated parameters of the latent class model. Thus, by applying a non-parametric bootstrap procedure, parameter uncertainty is incorporated in the final pooled standard error estimates of the statistics of interest.

Table 2: Cross-table between the variables region of hospital (columns) and region of accident (rows) for the years 1994, 2009 and 2013. Note that 'NA' means 'missing value'.

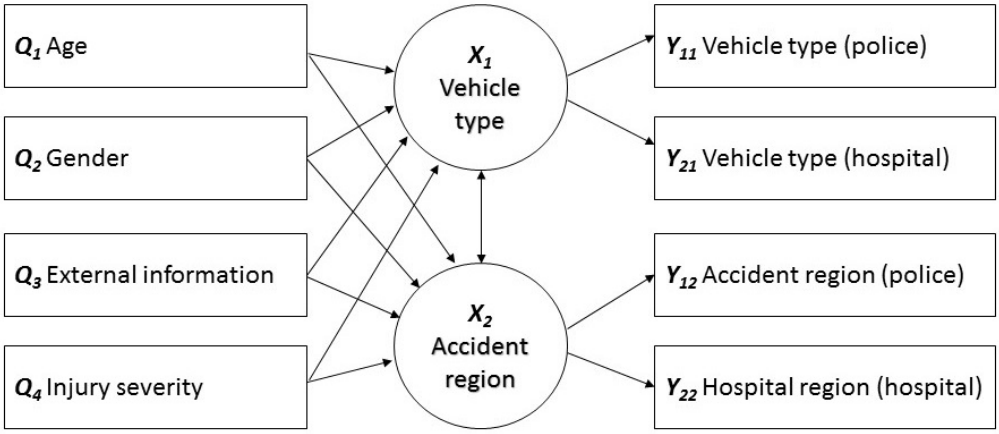
		1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	Total
1994														
	NA	345	419	213	627	772	499	1,152	1,140	123	997	590	111	6,988
	1. Groningen	314	4	2	5	2	1	4	2	0	0	2	1	337
	2. Friesland	17	393	5	7	0	1	1	5	0	3	1	2	435
	3. Drenthe	57	3	230	14	1	1	1	3	0	1	0	0	311
	4. Overijssel	2	3	26	711	7	4	6	4	0	9	2	4	778
	5. Gelderland	2	0	3	112	977	108	10	23	1	46	4	3	1,289
	6. Utrecht	3	3	1	2	52	564	38	7	2	6	3	1	682
	7. Noord-Holland	4	2	2	6	15	11	1,538	29	1	14	9	4	1,635
	8. Zuid-Holland	6	4	7	8	16	22	30	1,564	4	20	8	2	1,691
	9. Zeeland	0	0	0	0	2	0	1	9	212	24	0	0	248
	10. Noord-Brabant	1	2	1	5	60	6	17	35	2	1,550	33	0	1,712
	11. Limburg	0	2	1	1	19	2	5	3	1	12	690	1	737
	12. Flevoland	0	1	0	6	6	5	10	3	0	0	0	89	120
	Total	751	836	491	1,504	1,929	1,224	2,813	2,827	3,46	2,682	1,342	218	16,963
2009														
	NA	435	586	267	667	1,523	865	2,014	1,728	151	1,185	840	185	10,446
	1. Groningen	186	5	2	2	3	0	3	1	0	4	1	0	207
	2. Friesland	23	200	3	3	0	1	1	0	0	0	0	0	231
	3. Drenthe	48	0	91	16	1	0	1	3	1	4	2	0	167
	4. Overijssel	2	2	3	265	2	0	2	5	0	1	1	0	283
	5. Gelderland	1	2	0	51	516	58	5	5	0	20	2	0	660
	6. Utrecht	1	2	0	3	26	323	23	1	1	2	0	0	382
	7. Noord-Holland	0	3	2	1	10	11	673	11	2	6	12	0	731
	8. Zuid-Holland	2	3	1	3	6	19	13	683	0	6	4	0	740
	9. Zeeland	0	0	0	0	0	0	2	3	80	8	1	0	94
	10. Noord-Brabant	1	0	0	0	23	4	9	14	0	491	6	0	548
	11. Limburg	0	0	0	1	7	0	4	1	1	3	300	1	318
	12. Flevoland	1	13	0	22	5	3	6	3	0	0	0	27	80
	Total	700	816	369	1,034	2,122	1,284	2,756	2,458	236	1,730	1,169	213	14,887
2013														
	NA	392	534	372	857	1,696	870	2,643	2,286	324	1,475	815	154	12,418
	1. Groningen	53	0	0	0	0	0	0	1	1	2	1	0	58
	2. Friesland	12	77	0	1	0	1	1	0	0	0	0	0	92
	3. Drenthe	18	0	36	8	1	1	0	0	0	0	1	0	65
	4. Overijssel	0	0	0	180	1	0	0	2	0	0	2	0	185
	5. Gelderland	0	0	0	37	313	30	2	2	0	5	2	0	391
	6. Utrecht	0	0	0	0	13	178	15	1	1	3	2	0	213
	7. Noord-Holland	2	0	0	1	3	7	492	3	0	0	1	1	510
	8. Zuid-Holland	1	0	1	2	4	8	14	439	1	6	1	2	479
	9. Zeeland	0	0	0	0	0	0	0	8	57	5	0	0	70
	10. Noord-Brabant	1	1	2	0	19	1	3	11	0	293	3	0	334
	11. Limburg	0	0	0	2	14	0	0	3	1	3	141	0	164
	12. Flevoland	1	2	0	10	2	2	15	0	0	0	0	22	54
	Total	480	614	411	1,098	2,066	1,098	3,185	2,756	385	1,792	969	179	15,033

Figure 1: A map of the Netherlands

3.2 Specifying the latent class model

The second step of the extended MILC method is specifying the latent class (LC) model. The LC model is estimated separately to each bootstrap sample so that the differences between the parameters in the different LC models reflect parameter uncertainty. A graphical overview of the specified LC model can be found in Figure 2. First, the latent variable measuring vehicle type (X_1) is specified. The variables measuring vehicle type originating from the police registry (Y_{11}) and from the hospital registry (Y_{21}) are specified as indicators of this latent variable. Note that this notation differs from traditional notation where X variables are predictors and Y variables are responses, e.g. in regression analysis. As was discussed in the

Figure 2: Graphical overview of the latent class model specified in Latent GOLD.



background section, the vehicle type indicator variables contain nine categories in total, six representing motorized vehicles and three representing non-motorized vehicles. However, specifying nine latent classes would be problematic, since the number of observed non-motorized accidents in the police registry is very low. Therefore, the non-motorized categories are grouped into one category resulting in the specification of a seven class model. By saving the original scores of this indicator variable in separate variables, these can be re-assigned to the accidents which were assigned to the latent class 'accidents without motorized vehicle' after multiple imputation. For this, the proportions of the categories in the observed data are used.

Second, all covariates of interest need to be included in the LC, because otherwise point estimates describing the relationship between a latent variable and an excluded covariate will be biased (Bolck et al., 2004). As discussed in the background section, the variable 'region of accident' cannot be included directly as a covariate as it contains a large proportion of missing values. Therefore, multiple imputations are created for this variable to be able to stratify for the variable 'vehicle type' over the different regions in the Netherlands. For this purpose, a second latent variable is specified to measure 'region of accident' (X_2). The first indicator is the variable 'region of accident' measured in the police registry Y_{12}. The second indicator variable is 'region of hospital' (Y_{22}). Since the first indicator is actually the variable for which imputations are created, the relationship between the latent variable and the indicator variable is restricted such that if the indicator variable is observed, this score is assigned directly to the latent variable as well. Only if this indicator variable contains a missing value, the outcomes of this latent class model are used.

Other covariates needed to make relevant stratifications can be included in the LC model directly, since they do not contain any missing values. The other covariates included in the LC model are:

- Age: $0-17$; $18-44$; $46-69$; $70+$ (Q_1).
- Gender: Male; Female (Q_2).
- External information: Standard; Falling; Non-public road; No driving vehicle; Other (Q_3).
- Injury severity using Abbreviated Injury Scale (AIS), an anatomical scoring system where injuries are ranked on a scale from one to six. As 'one' represents 'minor injuries' and 'six' represents 'unsurvivable injuries', these do not fit in the scope of this research, as this research pertains to 'serious road injuries'. Therefore, the following scores on AIS are included: 'two' means 'Moderate'; 'three' means 'Serious'; 'four' means 'Severe'; 'five' means 'Critical' (Q_4) (Wong, 2011).

To ensure that all parameters can be estimated for each bootstrap sample, only main effects of the covariates are included in the LC model.

The latent class model for response pattern $\mathrm{P}(\mathbf{Y}=\mathbf{y}|\mathbf{Q}=\mathbf{q})$ is:

$$\begin{aligned}\mathrm{P}(\mathbf{Y}=\mathbf{y}|\mathbf{Q}=\mathbf{q}) = \sum_{x_1=1}^{7}\sum_{x_2=1}^{12}\prod_{l_1=1}^{2}\mathrm{P}(Y_{l_1,1}=y_{l_1,1}|X_1=x_1)\times \\ \prod_{l_2=1}^{2}\mathrm{P}(Y_{l_2,2}=y_{l_2,2}|X_2=x_2)\times \\ \mathrm{P}(X_1=x_1, X_2=x_2|\mathbf{Q}=\mathbf{q}).\end{aligned} \tag{1}$$

In this latent class model, X_1 represents the latent variable 'vehicle type' with seven classes and X_2 represents the latent variable 'region of accident' with 12 classes. Furthermore, $\mathbf{Q}$ represents the covariate variables and $\mathbf{Y}$ represents the indicator variables, where l_1 stands for the two indicator variables corresponding to X_1 and l_2 for the two indicator variables corresponding to X_2 (which corresponds to what can be seen in Figure 2). The latent class model is estimated using Latent GOLD 5.1 (Vermunt & Magidson, 2015), where the recommendations by Vermunt et al. (2008) for large datasets have been followed to ensure convergence. See the Appendix for the Latent GOLD syntax used.

By specifying the previously described latent class model, the first assumption made is that the probability of obtaining a specific response pattern is a weighted average of all conditional response probabilities, also known as the mixture assumption. Second, the assumption is made that the observed indicators are independent of each other given a unit's score on the underlying true measure. In other words, this me-

ans that if a classification error is made in the police registry, we assume that this is independent of the probability of also having a classification error in the hospital registry. For most cases this assumption can be considered realistic, since the police registry and the hospital registry are generally filled out by two different and independent persons. In rare situations, dependencies might arise. For example, in a 'hit-and-run' situation, both registries will probably be filled out based on information provided by the victim and are therefore not independent. Third, the assumption is made that the misclassification in the indicators is independent of the covariates. It is unlikely that scores on covariates such as age or gender will influence this. However, for example for the variable 'external information', it can be the case that if an accident takes place outside the public road, it is more difficult for the police to reach this location and therefore the probability of an error can increase. Fourth, the assumption is made that the covariate variables are free of error. This is, of course, an unrealistic assumption, especially given the substantial amounts of classification error found in the 'vehicle type' indicator variables. At this point we unfortunately do not have any information available about the extent of possible classification errors in the other variables. However, these errors are considered less problematic as long as they are random. Lastly, assumptions are made with respect to the missingness mechanisms present in the data. More specifically, the mechanism that governs the probability each data point has of being missing is considered Missing At Random (MAR) for the variables 'vehicle-type observed in the police registry' (Y_{11}) and 'region of accident' (Y_{12}), as the probability of being missing is larger for 'non-motorized' vehicles, which is measured by the hospital registry (Y_{21}). Formally, it can be stated that Y_{11} consists of a part $Y_{11,\text{obs}}$ and $Y_{11,\text{mis}}$ and that a vector R can be defined

$$R = 0 \text{ if } Y_{11,\text{obs}} \tag{2}$$

$$R = 1 \text{ if } Y_{11,\text{mis}}. \tag{3}$$

As we assume the missingness mechanism to be MAR, the distribution of missing values is related to Y_{21}:

$$\text{P}(R = 0|Y_{11,\text{obs}}, Y_{11,\text{mis}}, Y_{21}) = \text{P}(R = 0|Y_{11,\text{obs}}, Y_{21}). \tag{4}$$

What holds for Y_{11}, holds for Y_{12} as well. Furthermore, the mechanism that governs the probability of being missing is considered Missing Completely At Random for the variable 'vehicle type observed in the

hospital registry' (Y_{21}). Here,

$$\mathrm{P}(R = 0|Y_{21,\mathrm{obs}}, Y_{21,\mathrm{mis}}, Y_{11}) = \mathrm{P}(R = 0). \tag{5}$$

The latent class model gives different forms of relevant output. The first form of relevant output is the entropy R^2. Entropy can be formally defined as:

$$EN(\alpha) = -\sum_{j=1}^{N}\sum_{x=1}^{X}\alpha_{jx}\log\alpha_{jx}, \tag{6}$$

where α_{jx} is the probability that observation j is a member of class x, X the number of classes, and N is the number of units in the combined dataset. Rescaled to values between zero and one, entropy R^2 is measured by:

$$R^2 = 1 - \frac{EN(\alpha)}{N\log X}, \tag{7}$$

where one means perfect prediction (Dias & Vermunt, 2008). Boeschoten et al. (2017) show that the performance of the MILC method is closely related to the entropy R^2 of the corresponding latent class model.

A second form of relevant output are the conditional response probabilities. They provide us the probability of obtaining a specific response on the indicator conditional on belonging to a certain latent class. These values can be used to investigate the relationships between the indicator variables and the latent variables into detail. For example, they show us the probability of having the score 'M-car' on the indicator originating from the police registry given that the model assigned a case to the latent class 'M-car', but also the probability of having the score 'M-bicycle' on the indicator given that the model assigned a case to the latent class 'M-car'. Here, the former should be much higher compared to the latter. By comparing the conditional response probabilities to the cross-table between the variables measuring vehicle type originating from the police registry and the hospital registry (as seen in Table 1), it can be investigated if the latent classes identified as certain categories of vehicle-type are related to other categories in the indicator variables in a comparable way as in the observed data. In this way, it is checked if the latent class model reflects the main relations found in the observed data, which is an important indication of adequate imputations in the next step.

Third, the posterior membership probabilities represent the probability that a unit belongs to a latent

class given its combination of scores on the indicators and covariates used in the latent class model. These values are used to create multiple imputations for the latent variables, and the exact procedure for this is described in the next section.

3.3 Creating multiple imputations

The posterior membership probabilities are used to create multiple imputations of the latent variables containing the 'true scores'. The posterior membership probabilities can be estimated by applying Bayes' rule to the latent class model described in Equation 1:

$$\mathrm{P}(X_1 = x_1, X_2 = x_2 | \mathbf{Y} = \mathbf{y}, \mathbf{Q} = \mathbf{q}) = \frac{\mathrm{P}(X_1 = x_1, X_2 = x_2, \mathbf{Y} = \mathbf{y} | \mathbf{Q} = \mathbf{q})}{\mathrm{P}(\mathbf{Y} = \mathbf{y} | \mathbf{Q} = \mathbf{q})}, \tag{8}$$

where

$$\begin{aligned} \mathrm{P}(X_1 = x_1, X_2 = x_2, \mathbf{Y} = \mathbf{y} | \mathbf{Q} = \mathbf{q}) = & \prod_{l_1=1}^{2} \mathrm{P}(Y_{l_1,1} = y_{l_1,1} | X_1 = x_1) \times \\ & \prod_{l_2=1}^{2} \mathrm{P}(Y_{l_2,2} = y_{l_2,2} | X_2 = x_2) \times \\ & \mathrm{P}(X_1 = x_1, X_2 = x_2 | \mathbf{Q} = \mathbf{q}). \end{aligned} \tag{9}$$

and $\mathrm{P}(\mathbf{Y} = \mathbf{y} | \mathbf{Q} = \mathbf{q})$ is defined in Equation 1.

Since there are two latent variables specified in this model, the joint posterior membership probabilities are obtained which represent the probability that a unit is a member of a specific latent class in the latent variable 'vehicle type', and a member of a specific latent class in the latent variable 'accident region'. Since the variable 'vehicle type' has seven classes and the variable 'accident region' has 12 classes, there are 84 posterior membership probabilities which add up to one, and there is a different set of posterior membership probabilities for each combination of scores on the indicators and covariates. Parameter estimation was constrained in such a way that if a case had an observed score on the variable 'accident region' in the police registry, this score is automatically assigned to the latent variable as well. In those cases, there are only seven posterior membership probabilities with a value larger than zero (those representing the different classes for 'vehicle type' in combination with that specific region); all other posterior membership probabilities are exactly zero.

For each case in the original dataset, the posterior membership probabilities corresponding to its combination of scores on the indicators and covariates are used as a multinomial distribution to draw a joint score on both latent variables. These joint scores are then used to create separate imputations for the variables 'vehicle type' and 'accident region'.

By drawing multiple times from the posterior membership probabilities, multiple imputations for both latent variables are created. The scores assigned to the latent variables can be different for the different imputations. The differences between them reflect the uncertainty due to the missing and conflicting values in the indicator variables. Boeschoten et al. (2017) concluded that a low number of imputations, such as five is already sufficient for a correct estimation of the standard errors. However, in that simulation study the number of classes was much lower compared to the number of classes needed for this dataset. To evaluate what the appropriate number of imputations would be, the number of imputations was gradually increased and the fraction of missing information was compared between the differing numbers of imputations (Graham et al., 2007), resulting in 20 imputations. This is in line with the recommendations by Wang et al. (2005).

3.4 Pooling of the results

At this point, 20 imputations are created for 'vehicle type' and 'region of accident' for every unit in the combined dataset. The goal is to obtain estimates of interest using these imputed variables. This is done by obtaining the estimate of interest for every imputed variable, and pooling these estimates using the pooling rules defined by Rubin (Rubin, 1987, p.76). Although our context differs from the traditional statistical context for which the pooling rules were originally developed, the rules are considered appropriate for the context of multiple imputation for measurement error (Reiter & Raghunathan, 2007). For this specific research, the main estimates of interest are frequency tables.

The first step is to calculate a pooled frequency table. In other words, we take the average over the imputations for every cell in the frequency table. This can be for the imputed variable 'vehicle type', for the imputed variable 'region of accident' or for a cross-table between (one of) these variables and covariate(s). A pooled cell count is obtained by:

$$\hat{\theta}_j = \frac{1}{m} \sum_{i=1}^{m} \hat{\theta}_{ij}, \tag{10}$$

where θ refers to a cell count, j refers to a specific cell in the frequency table, i refers to one imputation and m refers to the total number of imputations.

Next, an estimate of the uncertainty around these frequencies is of interest. Therefore, the pooled frequencies need to be transformed into pooled proportions:

$$\hat{\mathrm{p}}_j = \frac{\frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_{ij}}{\sum_{j=1}^{s}\frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_{ij}}, \tag{11}$$

where s refers to the number of cells in the frequency table.

Since we work with a multiply imputed dataset, an estimate of the variance is obtained that is a combination of sampling uncertainty and uncertainty due to missing and conflicting values in the dataset. This is the total variance that consists of a 'within imputation' and 'between imputation' component:

$$\mathrm{VAR}_{\mathrm{total}_j} = \overline{\mathrm{VAR}}_{\mathrm{within}_j} + \mathrm{VAR}_{\mathrm{between}_j} + \frac{\mathrm{VAR}_{\mathrm{between}_j}}{m}. \tag{12}$$

$\overline{\mathrm{VAR}}_{\mathrm{within}_j}$ is the within imputation variance of $\hat{\mathrm{p}}_j$ calculated by

$$\overline{\mathrm{VAR}}_{\mathrm{within}_j} = \frac{1}{m}\sum_{i=1}^{m}\mathrm{VAR}_{\mathrm{within}_{ij}}, \tag{13}$$

where $\mathrm{VAR}_{\mathrm{within}_{ij}}$ is estimated as the variance of $\hat{\mathrm{p}}_{ij}$:

$$\frac{\hat{\mathrm{p}}_{ij}(1-\hat{\mathrm{p}}_{ij})}{N}, \tag{14}$$

where N is the total size of the observed dataset and $\hat{\mathrm{p}}_{ij}$ is estimated as

$$\hat{\mathrm{p}}_{ij} = \frac{\hat{\theta}_{ij}}{\sum_{j=1}^{s}\hat{\theta}_{ij}}. \tag{15}$$

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

$$\mathrm{VAR}_{\mathrm{between}_j} = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{\mathrm{p}}_{ij} - \hat{\mathrm{p}_j})(\hat{\mathrm{p}}_{ij} - \hat{\mathrm{p}_j})'. \tag{16}$$

When $\mathrm{VAR}_{\mathrm{total}_j}$ is estimated, it can be used to estimate the standard error of $\hat{\mathrm{p}}_j$

$$\mathrm{SE}(\hat{\mathrm{p}}_j) = \sqrt{\mathrm{VAR}_{\mathrm{total}_j}}. \tag{17}$$

From here, the confidence interval around $\hat{\mathrm{p}}_j$ can be estimated by

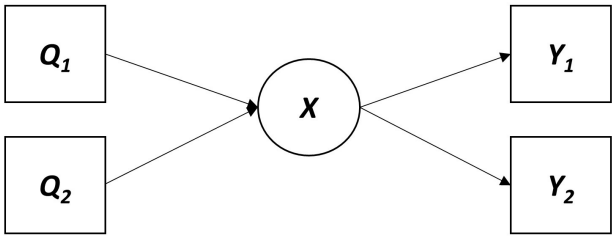
$$\hat{\mathrm{p}}_j \pm 0.975 \times \mathrm{SE}(\mathrm{P}_j), \tag{18}$$

where 0.975 corresponds to the $1 - \frac{\alpha}{2}$ quantile of a standard normal distribution for $\alpha = 0.05$. The values obtained here can simply be multiplied by N to obtain the 95% confidence intervals around the observed frequencies $\hat{\theta}_j$. Note that a standard normal distribution is assumed so problems can be encountered when dealing with very small proportions.

3.5 Performance of the MILC method

Boeschoten et al. (2017) introduced the MILC method, and evaluated the method under a range of conditions in terms of data quality. To investigate how the MILC method performs in comparison to the hierarchical assignment procedure traditionally used by SWOV(Bos et al., 2017), an illustrative simulation study is performed.

Figure 3: Graphical overview of the latent class model used for the simulation study.



In the theoretical population used for this simulation study, latent variable X has two categories with probabilities:

$$\begin{matrix} X = 1 \\ X = 2 \end{matrix} \begin{pmatrix} 0.6 \\ 0.4 \end{pmatrix}. \tag{19}$$

The probability distribution of $\mathrm{P}(X, Q_1)$ is:

$$\begin{array}{c} \\ X=1 \\ X=2 \end{array} \begin{array}{c} \begin{array}{cc} Q_1=1 & Q_1=2 \end{array} \\ \begin{pmatrix} 0.48 & 0.12 \\ 0.32 & 0.08 \end{pmatrix} \end{array} \tag{20}$$

and the probability distribution of $\mathrm{P}(X, Q_2)$ is:

$$\begin{array}{c} \\ X=1 \\ X=2 \end{array} \begin{array}{c} \begin{array}{ccc} Q_2=1 & Q_2=2 & Q_2=3 \end{array} \\ \begin{pmatrix} 0.36 & 0.18 & 0.06 \\ 0.24 & 0.12 & 0.04 \end{pmatrix} \end{array} \tag{21}$$

From this population stucture, $1,000$ samples are drawn. In each sample, indicator (Y_1) of X is created with 5% misclassification and a Missing At Random (MAR) mechanism, where the probability of being missing is related to a person's score on the Q_2 covariate.

$$Q_2 = 1, \mathrm{P}(Y_1 = \mathrm{NA}) = 0.20; \tag{22}$$

$$Q_2 = 2, \mathrm{P}(Y_1 = \mathrm{NA}) = 0.15; \tag{23}$$

$$Q_2 = 3, \mathrm{P}(Y_1 = \mathrm{NA}) = 0.10. \tag{24}$$

A second indicator (Y_2) of X is created with 15% misclassification and 5% missing cases which are Missing Completely At Random (MCAR). The latent class models had an entropy R^2 value of approximately 0.75.

The MILC method as described in sections 3.1, 3.2 and 3.3 is applied on the sample datasets, where five bootstrap samples are drawn and subsequently five imputations of X are created. As an illustration, the MILC method is also applied without the bootstrap procedure; so with one latent class model directly estimated on the observed data and five imputations drawn from one single set of posterior membership probabilities. Furthermore, the hierarchical assignment procedure as used by SWOV is also applied. At SWOV, the score observed in the police registry (Y_1) is assigned if it is observed. Otherwise, the score observed in the hospital registry (Y_2) is assigned.

The imputations are evaluated in terms of bias, coverage of the 95% confidence interval, confidence interval width, average standard error of the estimates divided by the standard deviation over the estimates

and the root mean squared error. Furthermore, the proportion of correctly classified cases is evaluated for imputation and hierarchical assignment.

To evaluate the methods, the marginals of the imputed latent variable (W) are compared to the hierarchically assigned variable (W_{ass}). In addition the estimated relationships of the latent variable with covariates ($W \times Q_1$, $W_{\text{ass}} \times Q_1$, $W \times Q_2$ and $W_{\text{ass}} \times Q_2$) are examined.

In Table 3 the results of the simulation study comparing the MILC method (with and without bootstrap) and the hierarchical assignment procedure are shown. We first discuss the performance of the MILC method in comparison to the hierarchical assignment method. The results obtained with hierarchical assignment especially show substantial amounts of bias for $W_{\text{ass}} \times Q_2$ as compared to both implementations of the MILC method. For the unbiased parameters obtained when applying hierarchical assignment, the RMSE is in general lower and more stable compared to the RMSE of MILC. The fact that with hierarchical assignment, bias is especially found in the results relating to Q_2, can be explained by the fact that the missingness mechanism of Y_1 is defined by Q_2.

Comparison of the MILC method with and without bootstrap shows clearly that standard errors are very much underestimated when no bootstraps are performed; that is, coverage rates are too low and the ratios between the average standard error and the standard deviation across replications are far below one. In contrast, these ratios are larger than one when the bootstrap is included in the MILC method, meaning the standard errors are somewhat overestimated. The large difference between the two approaches is caused by the fact that the statistics we are interested in are tables containing the latent variable X. By not applying the bootstrap, one seriously underestimates the uncertainty about the latent class proportions. The fact that the bootstrap procedure yields slightly too large standard errors can be considered to be less problematic than having (much) too small standard errors.

The percentage of incorrectly classified cases is 4.5% for $X = 1$ and 10.1% for $X = 2$ when hierarchical assignment is applied (these results are not shown in Table 3). When the MILC method (including bootstrap) is applied, the percentage of incorrectly classified cases is 8.6% for $X = 1$ and 20.5% for $X = 2$. Although this percentage is substantially larger for MILC compared to hierarchical assignment, this does not seem to affect the ability of the methods to produce unbiased results with appropriate standard errors.

Table 3: Results of a simulation study where the hierarchical assignment procedure is compared to the MILC method, which is performed with and without a non-parametric bootstrap. Results are shown for the imputed mixture variable, denoted by W, and of the relationship of W with covariates Q_1 and Q_2. Results are given in terms of bias, coverage of the 95% confidence interval, confidence interval width, the average standard error of the estimate divided by the standard deviation over the estimates, and the root mean squared error.

	Bias	Coverage	CI width	se/sd	RMSE
Hierarchical assignment					
$W_{\text{ass}} = 1$	-0.0134	0.2180	0.0193	0.9981	0.1158
$W_{\text{ass}} = 2$	0.0134	0.2180	0.0193	0.9981	0.1158
$W_{\text{ass}} = 1 \times Q_1 = 1$	-0.0106	0.4220	0.0196	0.9894	0.1029
$W_{\text{ass}} = 2 \times Q_1 = 1$	-0.0028	0.8380	0.0126	0.9964	0.0532
$W_{\text{ass}} = 1 \times Q_1 = 2$	0.0107	0.3590	0.0184	0.9963	0.1036
$W_{\text{ass}} = 2 \times Q_1 = 2$	0.0027	0.8380	0.0108	1.0191	0.0518
$W_{\text{ass}} = 1 \times Q_2 = 1$	0.0012	0.3560	0.0134	0.9433	0.0346
$W_{\text{ass}} = 2 \times Q_2 = 1$	-0.1676	0.6390	0.0107	1.0053	0.4094
$W_{\text{ass}} = 1 \times Q_2 = 2$	-0.2910	0.7770	0.0066	0.9898	0.5394
$W_{\text{ass}} = 2 \times Q_2 = 2$	-0.1115	0.3050	0.0121	0.9702	0.3339
$W_{\text{ass}} = 1 \times Q_2 = 3$	-0.2261	0.5920	0.0092	1.0201	0.4755
$W_{\text{ass}} = 2 \times Q_2 = 3$	-0.3022	0.7990	0.0056	1.0552	0.5498
MILC method, bootstrap excluded					
$W = 1$	-0.0317	0.1300	0.0216	0.1425	0.1781
$W = 2$	0.0317	0.1300	0.0216	0.1425	0.1781
$W = 1 \times Q_1 = 1$	-0.0252	0.1660	0.0213	0.1751	0.1586
$W = 2 \times Q_1 = 1$	-0.0066	0.4410	0.0132	0.3912	0.0810
$W = 1 \times Q_1 = 2$	0.0253	0.1660	0.0205	0.1683	0.1591
$W = 2 \times Q_1 = 2$	0.0064	0.3980	0.0118	0.3628	0.0800
$W = 1 \times Q_2 = 1$	-0.0191	0.2270	0.0201	0.2151	0.1380
$W = 2 \times Q_2 = 1$	-0.0095	0.3470	0.0157	0.3278	0.0976
$W = 1 \times Q_2 = 2$	-0.0031	0.5820	0.0096	0.5341	0.0560
$W = 2 \times Q_2 = 2$	0.0191	0.1980	0.0188	0.2029	0.1380
$W = 1 \times Q_2 = 3$	0.0095	0.3150	0.0142	0.2980	0.0977
$W = 2 \times Q_2 = 3$	0.0031	0.5510	0.0085	0.4962	0.0559
MILC method including bootstrap					
$W = 1$	-0.0304	0.8880	0.1790	1.5797	0.1745
$W = 2$	0.0304	0.8880	0.1790	1.5797	0.1745
$W = 1 \times Q_1 = 1$	-0.0241	0.8950	0.1439	1.5811	0.1553
$W = 2 \times Q_1 = 1$	-0.0063	0.9050	0.0383	1.4324	0.0796
$W = 1 \times Q_1 = 2$	0.0243	0.8940	0.1437	1.5744	0.1559
$W = 2 \times Q_1 = 2$	0.0062	0.9160	0.0378	1.4887	0.0785
$W = 1 \times Q_2 = 1$	-0.0183	0.8880	0.1087	1.5375	0.1352
$W = 2 \times Q_2 = 1$	-0.0091	0.9020	0.0560	1.5192	0.0956
$W = 1 \times Q_2 = 2$	-0.0030	0.9290	0.0205	1.4125	0.0551
$W = 2 \times Q_2 = 2$	0.0183	0.8910	0.1085	1.5562	0.1352
$W = 1 \times Q_2 = 3$	0.0092	0.9050	0.0555	1.5085	0.0957
$W = 2 \times Q_2 = 3$	0.0030	0.9280	0.0200	1.4670	0.0548

4 Results

First, results in terms of relevant model output will be discussed. Second, substantial results obtained after creating multiple imputations for the latent variables are given.

4.1 Latent class model output

Table 4: Entropy R^2 values for the latent variables 'Vehicle type' and 'Region of accident' for the years 1994, 2009 and 2013.

	Vehicle type	Region of accident
1994	0.8219	0.9050
2009	0.7444	0.8267
2013	0.8031	0.8077

The first relevant model output from the latent class models comes in terms of the entropy R^2. A separate entropy R^2 value is estimated for the two latent variables and for each year. The results are shown in Table 4. These results are obtained after applying an LC model on the original dataset. Here it can be seen that the entropy R^2 value in 2013 increased compared to 2009 for the variable 'Vehicle type'. Pankowska et al. (2017) showed in their simulation studies that when a latent class model is used to correct for misclassification in combined datasets, the model also treats inconsistencies due to incorrect linkage as misclassification and thereby corrects for it in a similar way. This implies that the increase in terms of entropy R^2 in 2013 in comparison to 2009 for the latent variable 'vehicle type' makes sense as the police improved their registration system in 2013. This improvement caused an increase in the number of correctly linked cases and therefore also improved the entropy R^2.

In Table 5, the probability of correct classification for the indicators of both latent variables are shown, for the three different time-points, obtained after applying an LC model to the original dataset. Class-specific response probabilities indicate the probability of having a score on the indicator variable that is equal to the latent class. A high probability of correct classification indicates that when a specific case belongs to a certain latent class, the probability is large that this same score was obtained on an indicator variable. For example, the probability of correct classification of the 1994 indicator variable 'Hospital' for the latent class 'Vehicle type = M car' is 0.8226. This means that the probability of having scored 'M car'

Table 5: Class-specific response probabilities for latent variables 'vehicle type' and 'region of accident' for the years 1994, 2009 and 2013.

		1994		2009		2013	
Vehicle type		Hospital	Police	Hospital	Police	Hospital	Police
	1. M car	0.8226	0.9782	0.8004	0.9742	0.9590	0.8973
	2. M moped	0.8458	0.9781	0.7194	0.9786	0.9693	0.8848
	3. M bicycle	0.7393	0.9170	0.7635	0.9620	0.9263	0.7376
	4. M motorcycle	0.8353	0.9686	0.8876	0.9129	0.0774	0.7577
	5. M other	0.6890	0.0578	0.5276	0.2629	0.0000	0.4243
	6. M pedestrian	0.7132	0.8213	0.8758	0.8104	0.5358	0.6412
	7. N all	0.9920	0.6162	0.9916	0.5273	0.9931	0.3897
Region of accident		Region of hospital	Region of accident	Region of hospital	Region of accident	Region of hospital	Region of accident
	1. Groningen	0.9351	1	0.8798	1	0.9167	1
	2. Friesland	0.9063	1	0.8740	1	0.8433	1
	3. Drenthe	0.7338	1	0.5897	1	0.6556	1
	4. Overijssel	0.9103	1	0.9290	1	0.9675	1
	5. Gelderland	0.7551	1	0.7961	1	0.8119	1
	6. Utrecht	0.8292	1	0.8259	1	0.8149	1
	7. Noord-Holland	0.9378	1	0.9267	1	0.9673	1
	8. Zuid-Holland	0.9240	1	0.9248	1	0.9094	1
	9. Zeeland	0.8506	1	0.8248	1	0.7941	1
	10. Noord-Brabant	0.9084	1	0.9055	1	0.8884	1
	11. Limburg	0.9397	1	0.9466	1	0.8725	1
	12. Flevoland	0.7771	1	0.5374	1	0.4694	1

on the indicator variable 'Vehicle type measured by hospital' is 0.8226 given that this case truly belongs to the latent class 'M car'.

When looking at the probabilities of correct classification for a specific latent class, the two probabilities corresponding to the two indicators are often not equal. This may be due to differences in the quality of the data. A low probability of correct classification can be caused by the fact that for this specific latent class, this category is observed many times in one indicator (here this is often the indicator 'Hospital'), while in the other indicator ('Police'), these cases are often missing. This can clearly be seen for the latent class 'N-all'. Conditional on truly belonging in this latent class, the probability of obtaining this score on the hospital indicator was 0.9920 in 1994. In other words, almost everyone who is assigned to this class by the model, obtained this score in the hospital registry as well. However, the probability of obtaining this score by the police is only 0.6162. A substantial part of the cases belonging to this latent class obtained another score or no score at all by the police.

In general, it can be seen that the probabilities of correct classification for the police indicator in 1994

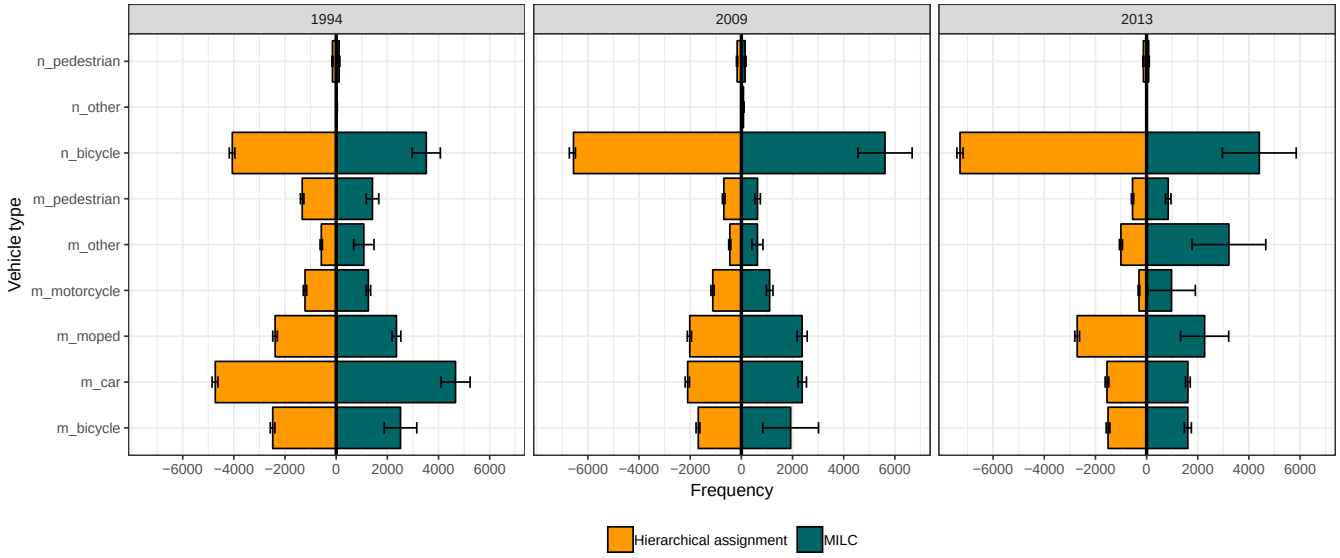
and 2009 are larger compared to the hospital indicator for all motorized classes except the class 'motorized, other' and the 'all non-motorized' category. However, in 2013 all probabilities of correct classification are higher for the hospital indicator compared to the police indicator. This result might be related to the improvement in the linking in 2013. An exception is the category 'M motorcycle', which is the only category with a probability of correct classification below 0.90 in the hospital registry. This is caused by the fact that some of the hospitals used a different registration system, that categorizes both motorcycles and mopeds into the motorcycle category.

When investigating the probabilities of correct classification for the latent variable 'region of accident', it can be seen that they are all exactly 1 for the indicator variable 'region of accident'. Conditional on being in a specific class in the latent variable 'region of accident', the probability of obtaining the same score on the indicator variable 'region of accident' is 1. This restriction was imposed on the latent class model. The probabilities of correct classification of the indicator variable 'region of hospital' now show us the probability that conditional on an accident truly happening in a specific region, what is the probability of also going to a hospital in that same region. These probabilities are generally quite high and stable over the different time-points. The regions Drenthe and Flevoland stand out because the probability of going to a hospital in these regions when having a serious road accident in this region is somewhat lower compared to other regions.

4.2 Pooled results output

In Figure 4, the number of serious road injuries per vehicle type are shown for the three different years investigated. For every year, the results obtained after applying the hierarchical assignment procedure are compared to results obtained when the extended MILC method is applied. Here, it can be seen that in general the frequencies obtained after applying the extended MILC method are quite similar compared to the results obtained after applying the hierarchical assignment procedure. When the extended MILC method is applied, the number of cases assigned to the category 'M-other' is larger while the number of cases assigned to the category 'N-bicycle' is smaller compared to the hierarchical assignment procedure, particularly in 2013. This corresponds to a large amount of missing cases for 'N-bicycle' and a substantial amount of cases differently categorized by the police and hospital. Furthermore, in 2013 the number of cases categorized as 'M-other' by the hospital increased, while this category was often classified differently

Figure 4: The three graphs represent results obtained for three different years. On the left side of each graph, the number of serious road injuries per vehicle type and corresponding 95% confidence intervals are shown when the hierarchical assignment procedure is applied. On the right side of each graph, pooled frequencies and 95% confidence intervals are shown when the extended MILC method is applied.



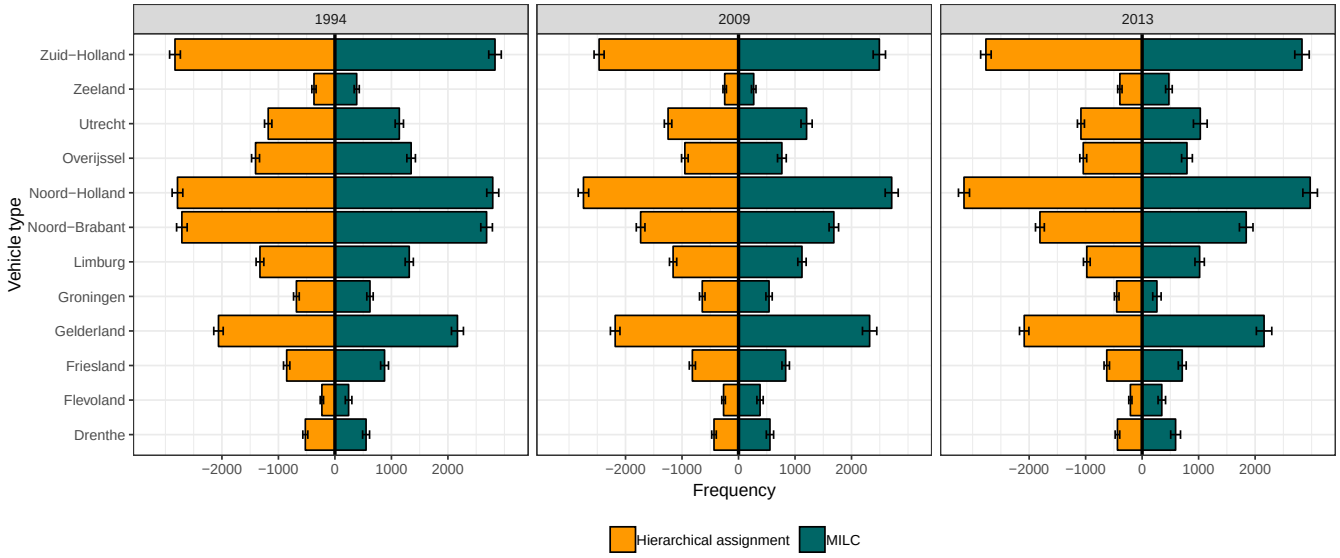
by the police (see Table 1). At last, it can be seen that the width of the 95% confidence intervals are substantially larger for all categories when the extended MILC method is applied.

In Figure 5, the number of serious road injuries per region are shown for the three different years investigated. For every year, the results obtained after applying the hierarchical assignment procedure are compared to results obtained when the extended MILC method is applied, which are very similar. The 95% confidence intervals are larger when the extended MILC method was applied compared to the hierarchical assignment procedure, but the difference is not as substantial as was the case for the variable 'vehicle type' in Figure 4.

5 Discussion

In this article, an extension of the MILC method was developed and applied to estimate the number of serious road injuries per vehicle type and to stratify this number in relevant subgroups. Information on serious road injuries was found in registries from both police and hospitals, which are both incomplete and contain misclassification. These variables were used as indicators of a latent variable of which it can be said that it contains the 'true scores'. Posterior membership probabilities obtained from this latent class model were then used to create multiple imputations of these 'true scores'. Simultaneously, multiple

Figure 5: The three graphs represent results obtained for three different years. On the left side of each graph, frequencies of serious road injuries per region and corresponding 95% confidence intervals are shown when the hierarchical assignment procedure is applied. On the right side of each graph, pooled frequencies and 95% confidence intervals are shown when the extended MILC method is applied.



imputations were created for the missing values in the variable 'region of accident' by using this variable as a perfectly measured indicator of the latent variable 'region of accident' and supplementing it by specifying the variable 'region of hospital' as an imperfectly measured indicator.

Multiple imputations were created for the variable 'vehicle type' and for the variable 'region of accident'. All variables are now fully imputed for every case in the dataset. Descriptive statistics of these variables, or estimates of relationships with other variables can now be investigated in a straightforward manner.

The extended MILC method was applied on datasets for the years 1994, 2009 and 2013. The quality of the data for these years was very different, which can be seen in the number of observations per registry per year and which is reflected in the entropy R^2 of the corresponding latent class model. In general the quality of the data was sufficient for applying the MILC method. In contrast, the results of the extended MILC method were compared to the results obtained when the hierarchical assignment procedure was applied (traditionally used to generate these statistics). A clear difference was that the extended MILC method generated wider 95% confidence interval widths. Based on the results obtained from the simulation study performed in section 3.5, it can be concluded that these wider confidence interval widths were indeed necessary to obtain nominal coverage rates.

A number of issues are worth reflecting on a bit further. First of all, it is important to note that our results heavily depend on the model assumptions made. In particular, the assumption is made that the classification errors are independent of covariates (also known as 'ICE' and 'homogeneous CE'). Furthermore, the assumption is made that the covariate variables are free of error. Violating this assumption does not necessarily have to be an issue if these errors are random. However, there is currently no literature on this topic, so more research in this specific area is needed in order to be able to adapt the model. A more crucial assumption is that the missingness is at random (MAR). Although from a theoretical perspective this assumption is likely to hold, it could however lead to substantial bias in cases where this assumption is violated.

A second issue is how the extended MILC method dealt with non-motorized vehicles. This was an ad hoc procedure to handle an issue that could not be handled by the latent class model. This ad-hoc procedure turned out to be useful. It can be investigated whether a comparable procedure could be applied to handle a moped/motorcycle issue in the 2013 dataset and whether there are other issues that can be solved like this.

This particular dataset contained a number of issues, of which a substantial part has been investigated by means of a simulation study. The results of this simulation study made clear that the extended MILC method was able to handle the missing values in the indicator variables and that the non-parametric bootstrap was required to obtain nominal coverage rates. It is however not investigated if and how large numbers of categories influence the results. Therefore, the number of imputations was increased and evaluated using methods to evaluate the number of imputations for missing values. A more thorough investigation could provide insight into whether these methods are suitable to evaluate the number imputations needed when the MILC method is applied, and how many imputations are needed to evaluate datasets with larger numbers of categories.

Furthermore, in the initial model proposed by Boeschoten et al. (2017), bootstrap samples were taken of the original data to incorporate parameter uncertainty in the estimate of the total variance. This appeared to be problematic for larger models with many interactions as those used in our application, because not all parameters can be estimated for every bootstrap sample. Alternatives to incorporate parameter uncer-

tainty can be Bayesian MCMC or a parametric bootstrap. However, it should also be investigated whether such a step is still necessary for larger sample sizes as parameter uncertainty can become minimal in such cases. As the simulation study showed that it was necessary to incorporate parameter uncertainty when creating imputations for this specific case, a model with only main effects was used to enable estimation of all parameters.

At last, it is important to note that missing values in the combined dataset and classification errors in the observed data are not the only issues when estimating the total number of serious road injuries per vehicle type. There is also a number of serious road injuries that are neither observed by the hospital nor by the police. Weighting and capture/recapture methods are typically used to obtain an estimate of the total number of serious road injuries; approaches which can easily be combined with MILC by applying the methods on the imputations separately. A variance estimate would then include uncertainty about the total number of injuries which is typically estimated by making use of bootstrapping. This can also be applied separately to every imputation before pooling of the results is applied (Gerritse et al., 2016).

By creating multiple imputations using a latent class model, multiply imputed versions of variables that contained missing values and/or classification errors are created. These can be used to easily provide frequencies, to further divide these frequencies into relevant subgroups or to create statistical figures. This application showed that the initial MILC method can be extended to handle problems that are dataset-specific. Furthermore, this application highlighted various new problems that one may need to deal with when applying the MILC approach. In future research, these will be investigated more thoroughly to fully exploit the potential of the MILC method for dealing with classification error problems.

Acknowledgement

The authors would like to thank Niels Bos and Jacques Commandeur from the Institute of Road Safety Research (SWOV) for providing us with the data and for their useful comments on earlier versions of this manuscript.

Appendix A Latent GOLD syntax

```
options
  maxthreads=all;
algorithm
  tolerance=1e-008 emtolerance=0.01 emiterations=20000  nriterations=0;
startvalues
  seed=0 sets=200 tolerance=1e-005 iterations=500;
bayes
  categorical=1 variances=1 latent=1 poisson=1;
missing
  includeall;
output
  profile;
outfile
  'posteriors1.dat' classification
keep
  LRM2, BRON2, wfactor;
variables
  caseweight b1;
  dependent LRM nominal 7, BRON nominal 7, prov_hosp nominal 12, prov_acc nominal 12;
  independent ernst nominal, external nominal, gender nominal, age nominal;
  latent X nominal 7, Xacc nominal 12;
equations
  LRM       <- 1 | X;
  BRON      <- 1 | X;
  prov_acc  <- (a~wei)Xacc;
  prov_hosp <- 1 | Xacc;
  X         <- 1 | ernst + external + gender + age;
  Xacc      <- 1 | ernst + external + gender + age;
  X <-> Xacc;
a={1 0 0 0 0 0 0 0 0 0 0 0
   0 1 0 0 0 0 0 0 0 0 0 0
   0 0 1 0 0 0 0 0 0 0 0 0
   0 0 0 1 0 0 0 0 0 0 0 0
   0 0 0 0 1 0 0 0 0 0 0 0
   0 0 0 0 0 1 0 0 0 0 0 0
   0 0 0 0 0 0 1 0 0 0 0 0
   0 0 0 0 0 0 0 1 0 0 0 0
   0 0 0 0 0 0 0 0 1 0 0 0
   0 0 0 0 0 0 0 0 0 1 0 0
   0 0 0 0 0 0 0 0 0 0 1 0
   0 0 0 0 0 0 0 0 0 0 0 1};
```

To ensure convergence and to minimize the probability of obtaining local maxima, the number of random start sets is set to 200 with 500 iterations each. The use of Newton Rapson iterations is suppressed and the number of EM iterations is increased to $20,000$, following the suggestions by Vermunt et al. (2008).

To reduce computation time, the storing of parameters and the computation of standard errors is suppressed, since conditional and posterior response probabilities are of main interest.

To ensure that in the latent variable 'Accident of region' ($Xacc$ in the Latent GOLD syntax) the value observed in the indicator variable 'Accident of region' ($prov\_acc$ in the Latent GOLD syntax) is assigned in cases where this variable is observed, the relationship between $Xacc$ and $prov\_acc$ is restricted using the matrix denoted by 'a' in the Latent GOLD syntax.

References

Boeschoten, L., Oberski, D., & 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.

Bolck, A., Croon, M., & Hagenaars, J. (2004). Estimating latent structure models with categorical variables: One-step versus three-step estimators. Political Analysis, 12(1), 3–27.

Bos, N., Stipdonk, H., & Commandeur, J. (2017). Ernstig verkeersgewonden 2016. SWOV Instituut voor Wetenschappelijk Onderzoek Verkeersveiligheid. Retrieved from https://www.swov.nl/publicatie/ernstig-verkeersgewonden-2016

Dias, J. G., & Vermunt, J. K. (2008). A bootstrap-based aggregate classifier for model-based clustering. Computational Statistics, 23(4), 643–659. Retrieved from http://dx.doi.org/10.1007/s00180-007-0103-7 doi: 10.1007/s00180-007-0103-7

Gerritse, S. C., Bakker, B. F., de Wolf, P.-P., & van der Heijden, P. G. (2016). Undercoverage of the population register in the netherlands, 2010. CBS Discussion paper, 2016.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? some practical clarifications of multiple imputation theory. Prevention science, 8(3), 206–213.

Pankowska, P., Bakker, B., Oberski, D., & Pavlopoulos, D. (2017, 3). Estimating employment mobility using linked data from different sources. does linkage error matter?

Reiter, J. P., & Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation. Journal of the American Statistical Association, 102(480), 1462–1471.

Reurings, M. C. B., & Bos, N. M. (2012). Ernstig verkeersgewonden in de jaren 2009 en 2010: update van de cijfers.

Reurings, M. C. B., & Stipdonk, H. L. (2009, December). Ernstig gewonde verkeersslachtoffers in nederland in 1993-2008. Stichting Wetenschappelijk Onderzoek Verkeersveiligheid SWOV. Retrieved from https://www.swov.nl/publicatie/ernstig-gewonde-verkeersslachtoffers-nederland-1993-2008

Reurings, M. C. B., & Stipdonk, H. L. (2011). Estimating the number of serious road injuries in the netherlands. Annals of epidemiology, 21(9), 648–653.

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys (wiley series in probability and statistics).

Vermunt, J. K., & Magidson, J. (2015). Upgrade manual for Latent GOLD 5.1. Statistical Innovations Inc.

Vermunt, J. K., Van Ginkel, J. R., Der Ark, V., Andries, L., & Sijtsma, K. (2008). Multiple imputation of incomplete categorical data using latent class analysis. Sociological Methodology, 38(1), 369–397.

Wang, C.-P., Brown, C. H., & Bandeen-Roche, K. (2005). Residual diagnostics for growth mixture models. Journal of the American Statistical Association, 100(471), 1054-1076. Retrieved from https://doi.org/10.1198/016214505000000501 doi: 10.1198/016214505000000501

Wong, E. (2011). Abbreviated injury scale. In J. S. Kreutzer, J. DeLuca, & B. Caplan (Eds.), Encyclopedia of clinical neuropsychology (pp. 5–6). New York, NY: Springer New York. Retrieved from https://doi.org/10.1007/978-0-387-79948-3_2 doi: 10.1007/978-0-387-79948-3_2