Estimating the number of serious road injuries per vehicle type in the Netherlands by using multiple imputation of latent classes

Boeschoten, Laura; de Waal, Ton; Vermunt, Jeroen

Published in:
Journal of the Royal Statistical Society A

DOI:
10.1111/rssa.12471

Publication date:
2019

Document Version
Publisher's PDF, also known as Version of record

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2019)
182, *Part* 4, *pp.* 1463–1486

Estimating the number of serious road injuries per vehicle type in the Netherlands by using multiple imputation of latent classes

Laura Boeschoten and Ton de Waal

Tilburg University, and Centraal Bureau voor de Statistiek, The Hague, The Netherlands

and Jeroen K. Vermunt

Tilburg University, The Netherlands

[Received March 2018. Final revision April 2019]

Summary. Statistics that are published by official agencies are often generated by 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). We apply the MILC method 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 by 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 data set 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: Classification error; Combined data set; Latent class analysis; Missing values; 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 not to 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 (SWOV)). An important statistic that the 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 and Stipdonk, 2011). To gain more insight into the total number of serious road injuries, it can be further stratified by vehicle type, severity of injury and region (Reurings and Bos, 2012).

Address for correspondence: Laura Boeschoten, Department of Methodology and Statistics, Tilburg School of Social and Behavioral Sciences, Tilburg University, Warandelaan 2, Tilburg 5000 E, The Netherlands.
E-mail: L.Boeschoten@uvt.nl

When estimating the number of serious road injuries in the Netherlands, the SWOV uses information from police and hospital registries. These registries contain classification errors and are incomplete. The SWOV estimates the number of units that are missing in both registries by a method based on capture–recapture (Reurings and Stipdonk, 2011). However, a procedure to correct for classification errors and missing values within the observed cases has not been applied.

A method to deal simultaneously with classification errors and missing values within the observed cases is the recently proposed multiple imputation of latent classes (MILC) method (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 at a unit level. All registries are required to contain identifier variables for their cases which makes it possible to link the information for a specific case in one registry to its information in the other registries. In such a combined data set, 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 are estimated by 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 data sets 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), page 76).

In this paper, the MILC method is applied to a linked data set 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. Because of the way in which this data set is constructed, a special feature of this data set is that, whenever one of these two indicators is missing, the other is observed. 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 observed in the police registry only and contains many missing cases. To estimate this statistic, the MILC method is extended in two ways. First, the MILC method is extended to estimate two latent variables simultaneously (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 to 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 the unit-linked police–hospital data sets. In addition, an illustrative simulation study is performed. Here, the results that are obtained after applying the extended MILC method are compared with results that are 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.

The programs that were used to analyse the data can be obtained from

`https://rss.onlinelibrary.wiley.com/hub/journal/1467985x/series-a-datasets`

2. Background

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

For every year, units that are observed in the two sources are linked by using information on personal and accident characteristics (Reurings and 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

Table 1. Cross-table between the variables measuring vehicle type originating from the police registry (columns) and from the hospital registry (rows) for the years 1994, 2009 and 2013†

Year	*Category*	*Missing value*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*9*	*Total*
1994	Missing value	—	561	245	318	122	42	137	90	14	1529
	1 M–car	918	2596	11	72	12	22	25	2	1	3659
	2 M–moped	702	29	1131	21	60	2	8	2	1	1956
	3 M–bicycle	397	40	70	1111	2	1	53	25	4	1703
	4 M–motorcycle	347	16	41	2	633	3	0	0	0	1042
	5 M–other	450	408	106	104	35	50	116	8	2	1279
	6 M–pedestrian	421	128	37	231	4	5	537	5	5	1373
	7 N–bicycle	3625	28	41	221	3	3	11	296	3	4231
	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	6988	3809	1684	2084	871	132	907	436	52	16963
2009	Missing value	—	209	111	126	38	20	62	26	6	598
	1 M–car	779	969	8	29	8	17	3	0	0	1813
	2 M–moped	1117	4	611	10	23	20	2	0	0	1787
	3 M–bicycle	565	23	17	701	0	9	20	9	0	1344
	4 M–motorcycle	668	9	74	2	367	6	0	0	0	1126
	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	6369	17	22	161	2	4	5	144	4	6728
	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	10446	1321	901	1120	451	101	323	194	30	14887
2013	Missing value	—	59	29	33	15	36	11	5	1	189
	1 M–car	877	566	3	1	4	65	3	0	0	1519
	2 M–moped	2220	8	419	3	167	63	2	1	0	2883
	3 M–bicycle	944	4	11	451	0	155	10	7	0	1582
	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	7230	12	7	41	1	29	2	44	1	7367
	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	12418	669	490	561	227	446	157	57	8	15033

†Note that there are no observations for the category 'Non-motorized—other' in the police registry.

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 experienced an accident with a motorized vehicle, whereas, for the category N–bicycle, the injured person was on a bike and no motorized vehicle was 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 increases in the category N–bicycle, it can be caused by unsafe bicycle lanes. If the number of injuries increases in the category M–bicycle, it can be caused by a high speed limit on roads that are 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 are 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, severity of injury or region of accident. To be

Fig. 1. Map of the Netherlands

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 so. For the variable region of accident, this is a problem, as this variable is observed in the police registry only.

To solve the issue of missing values in region of accident the traditional MILC method is extended in such a way that missing values in region of accident are imputed simultaneously whereas the latent variable true vehicle type is estimated. To create these imputations, information is used from region of hospital, which is observed for the cases that contain missing values for 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, whereas 12 of those accidents were registered to have taken place in Friesland (Table 2), which is a neighbouring region of Groningen. There was also one person in a hospital in Groningen for whom the accident was registered to be in Zuid-Holland, which is quite far away from Groningen (see Fig. 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 that 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 healthcare.

3. Applying the extended multiple imputation of latent classes 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 data set that is linked and processed as described in the previous section.

3.1. Bootstrapping for parameter uncertainty

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 by using the M sets of parameter values that are obtained. This is preferable over creating imputations based on the maximum likelihood estimates that are 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.

3.2. Specifying the latent class model

The second step of the extended MILC method is specifying the latent class model. The latent class model is estimated separately for each bootstrap sample so that the differences between the parameters in the different latent class models reflect parameter uncertainty. A graphical overview of the specified latent class model can be found in Fig. 2. First, the latent variable measuring vehicle type, X_1, is specified. The variables measuring vehicle type originating from

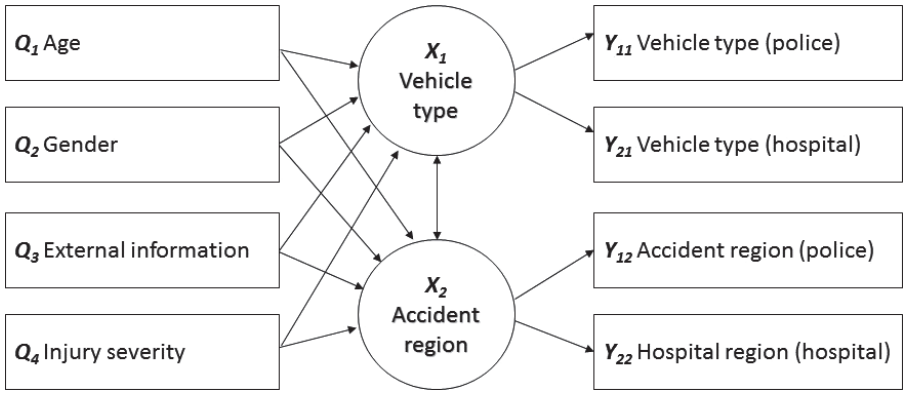

Fig. 2. Graphical overview of the latent class model specified in Latent GOLD

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 Section 2, 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 reassigned 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 latent class, 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 Section 2, 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 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 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 are the outcomes of this latent class model used.

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

(a) age, 0–17, 18–44, 46–69 and 70 years or older (Q_1),
(b) gender, male or female (Q_2),
(c) external information, standard, falling, non-public road, no driving vehicle and other (Q_3), and
(d) injury severity by using the abbreviated injury scale, which is an anatomical scoring system where injuries are ranked on a scale from 1 to 6. As '1' represents 'minor injuries' and '6' 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 the abbreviated injury scale are included: '2' means 'moderate'; '3' means 'serious'; '4' means 'severe'; '5' 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 latent class model.

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

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

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 Fig. 2). The latent class model is estimated by using Latent GOLD 5.1 (Vermunt and Magidson, 2015), where the recommendations by Vermunt *et al.* (2008) for large data sets have been followed to ensure convergence. See Appendix A for the Latent GOLD syntax that was 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, which is 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 means 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 people. In rare situations, dependences might arise. For example, in a 'hit-and-run' situation, both registries will probably be filled out on the basis of information that is 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 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 that is found in the vehicle type indicator variables. At this point we unfortunately do not have any information 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 in the data. More specifically, the mechanism that governs the probability that each data point has of being missing is considered missing at random 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_{11} can be defined:

$$R_{11} = \begin{cases} 0 & \text{if } Y_{11,\text{obs}}, \quad (2) \\ 1 & \text{if } Y_{11,\text{mis}}. \quad (3) \end{cases}$$

Table 2. Cross-table between the variables region of hospital (columns) and region of accident (rows) for the years 1994, 2009 and 2013

Year		*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*Total*
1994	Missing value	345	419	213	627	772	499	1152	1140	123	997	590	111	6988
	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	1289
	6 Utrecht	3	3	1	2	52	564	38	7	2	6	3	1	682
	7 Noord-Holland	4	2	2	6	15	11	1538	29	1	14	9	4	1635
	8 Zuid-Holland	6	4	7	8	16	22	30	1564	4	20	8	2	1691
	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	1550	33	0	1712
	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	1504	1929	1224	2813	2827	346	2682	1342	218	16963
2009	Missing value	435	586	267	667	1523	865	2014	1728	151	1185	840	185	10446
	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	1034	2122	1284	2756	2458	236	1730	1169	213	14887

(*continued*)

Table 2 (*continued*)

Year		*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*Total*
2013	Missing value	392	534	372	857	1696	870	2643	2286	324	1475	815	154	12418
	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	1098	2066	1098	3185	2756	385	1792	969	179	15033

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

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

If a value is missing in Y_{11}, it is by definition missing in Y_{12} as well, as unit missingness is considered here and both variables originate from the same data set. 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,

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

The distribution of missing values in Y_{11} and Y_{12} is related to Y_{21}, which in itself also contains missing values. Generally this would mean that we are not dealing with a missingness at random mechanism for Y_{11} and Y_{12}. However, because of the special structure of our data set in which Y_{11} and Y_{12} never contain missing values if Y_{21} contains missing values and vice versa, we are still dealing with a missingness at random mechanism. Cases containing missing values on all above-mentioned variables are by definition not included in the data set and are treated separately.

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

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

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 data set. Rescaled to values between 0 and 1, entropy R^2 is measured by

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

where 1 means perfect prediction (Dias and Vermunt, 2008). Boeschoten *et al.* (2017) showed 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 is the conditional response probabilities. They provide us with the probability of obtaining a specific response on the indicator conditionally on belonging to a certain latent class. These values can be used to investigate the relationships between the indicator variables and the latent variables in 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 with the latter. By comparing the conditional response probabilities with 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 whether the latent classes that are identified as certain categories of vehicle type are related to other categories in the indicator variables in a comparable way with that in the observed data. In this way, it is checked whether the latent class model reflects the main relationships that are 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 that are 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 the Bayes rule to the latent class model that is described in equation (1):

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

where

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

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

Since two latent variables are 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 region of accident. Since vehicle type has seven classes and region of accident has 12 classes, there are 84 posterior membership probabilities which add up to 1, 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 region of accident 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 0 (those representing the different classes for vehicle type in combination with that specific region); all other posterior membership probabilities are exactly 0.

For each case in the original data set, 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 vehicle type and region of accident.

By drawing multiple times from the posterior membership probabilities, multiple imputations for both latent variables are created. The scores that are 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 5, is already sufficient for a correct estimation of the standard errors. However, in that simulation study the number of classes was much lower compared with the number of classes that is needed for this data set. 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 data set. The goal is to obtain estimates of interest by using these imputed variables. This is done by obtaining the estimate of interest for every imputed variable, and pooling these estimates by using the pooling rules that were defined by Rubin (Rubin (1987), page 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 and 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:

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

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

Since we work with a multiply imputed data set, an estimate of the variance is obtained that is a combination of sampling uncertainty and uncertainty due to missing and conflicting values in the data set. 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 $\hat{p}_j$ 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 $\hat{p}_{ij}$:

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

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

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

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

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

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

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

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

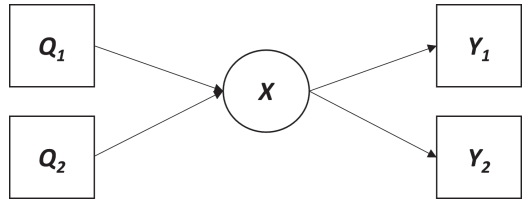

Fig. 3. Graphical overview of the latent class model used for the simulation study

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

where 0.975 corresponds to the $(1-\alpha/2)$-quantile of a standard normal distribution for $\alpha = 0.05$. The values that are 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 multiple imputation of latent classes method

Boeschoten *et al.* (2017) introduced the MILC method and evaluated the method under a range of conditions in terms of data quality. In addition, Boeschoten *et al.* (2018b) extended the method for situations with longitudinal data and Boeschoten *et al.* (2018a) extended the method in such a way that covariates can be included at later time points. All research performed on the MILC method so far has shown a strong relationship between the performance of the method and the entropy R^2-value of the latent class model. To investigate how the MILC method performs in comparison with the hierarchical assignment procedure that is traditionally used by the SWOV (Bos *et al.*, 2017), an illustrative simulation study is performed using a latent class model as shown in Fig. 3.

In the theoretical population that is used for this simulation study, latent variable X has two categories with probabilities 0.6 for $X = 1$ and 0.4 for $X = 2$. The probability distribution of $P(X, Q_1)$ is

$$\begin{array}{c} \\ X=1 \\ X=2 \end{array} \begin{array}{c} \begin{array}{cc} Q_1=1 & Q_1=2 \end{array} \\ \begin{pmatrix} 0.48 & 0.12 \\ 0.32 & 0.08 \end{pmatrix} \end{array} \tag{19}$$

and the probability distribution of $P(X, Q_2)$ is

$$\begin{array}{c} \\ X=1 \\ X=2 \end{array} \begin{array}{c} \begin{array}{ccc} Q_2=1 & Q_2=2 & Q_2=3 \end{array} \\ \begin{pmatrix} 0.36 & 0.18 & 0.06 \\ 0.24 & 0.12 & 0.04 \end{pmatrix} \end{array} \tag{20}$$

From this population structure, 1000 samples are drawn. In each sample, indicator Y_1 of X is created with 5% misclassification and a missingness at random mechanism, where the probability of being missing is related to a person's score on the Q_2-covariate:

$$Q_2 = 1,\ P(Y_1 = \mathrm{NA}) = 0.20; \tag{21}$$

$$Q_2 = 2,\ P(Y_1 = \mathrm{NA}) = 0.15; \tag{22}$$

$$Q_2 = 3,\ P(Y_1 = \mathrm{NA}) = 0.10. \tag{23}$$

Table 3. Results of a simulation study where the hierarchical assignment procedure is compared with the MILC method, which is performed with and without a non-parametric bootstrap†

	Bias	*Coverage*	*Confidence interval width*	*se/sd*	*RMSE*
Hierarchical assignment					
$W_{ass}=1$	−0.0134	0.2180	0.0193	0.9981	0.0143
$W_{ass}=2$	0.0134	0.2180	0.0193	0.9981	0.0143
$W_{ass}=1\times Q_1=1$	−0.0106	0.4220	0.0196	0.9894	0.0117
$W_{ass}=2\times Q_1=1$	−0.0028	0.8380	0.0126	0.9964	0.0043
$W_{ass}=1\times Q_1=2$	0.0107	0.3590	0.0184	0.9963	0.0117
$W_{ass}=2\times Q_1=2$	0.0027	0.8380	0.0108	1.0191	0.0038
$W_{ass}=1\times Q_2=1$	0.0012	0.3560	0.0134	0.9433	0.0052
$W_{ass}=2\times Q_2=1$	−0.1676	0.6390	0.0107	1.0053	0.1677
$W_{ass}=1\times Q_2=2$	−0.2910	0.7770	0.0066	0.9898	0.2910
$W_{ass}=2\times Q_2=2$	−0.1115	0.3050	0.0121	0.9702	0.1116
$W_{ass}=1\times Q_2=3$	−0.2261	0.5920	0.0092	1.0201	0.2261
$W_{ass}=2\times Q_2=3$	−0.3022	0.7990	0.0056	1.0552	0.3022
MILC method, bootstrap excluded					
$W=1$	−0.0317	0.1300	0.0216	0.1425	0.042
$W=2$	0.0317	0.1300	0.0216	0.1425	0.042
$W=1\times Q_1=1$	−0.0252	0.1660	0.0213	0.1751	0.0335
$W=2\times Q_1=1$	−0.0066	0.4410	0.0132	0.3912	0.0093
$W=1\times Q_1=2$	0.0253	0.1660	0.0205	0.1683	0.0336
$W=2\times Q_1=2$	0.0064	0.3980	0.0118	0.3628	0.0089
$W=1\times Q_2=1$	−0.0191	0.2270	0.0201	0.2151	0.0257
$W=2\times Q_2=1$	−0.0095	0.3470	0.0157	0.3278	0.0131
$W=1\times Q_2=2$	−0.0031	0.5820	0.0096	0.5341	0.0048
$W=2\times Q_2=2$	0.0191	0.1980	0.0188	0.2029	0.0255
$W=1\times Q_2=3$	0.0095	0.3150	0.0142	0.2980	0.0131
$W=2\times Q_2=3$	0.0031	0.5510	0.0085	0.4962	0.0046
MILC method including bootstrap					
$W=1$	−0.0304	0.8880	0.1790	1.5797	0.0420
$W=2$	0.0304	0.8880	0.1790	1.5797	0.0420
$W=1\times Q_1=1$	−0.0241	0.8950	0.1439	1.5811	0.0335
$W=2\times Q_1=1$	−0.0063	0.9050	0.0383	1.4324	0.0093
$W=1\times Q_1=2$	0.0243	0.8940	0.1437	1.5744	0.0336
$W=2\times Q_1=2$	0.0062	0.9160	0.0378	1.4887	0.0089
$W=1\times Q_2=1$	−0.0183	0.8880	0.1087	1.5375	0.0257
$W=2\times Q_2=1$	−0.0091	0.9020	0.0560	1.5192	0.0131
$W=1\times Q_2=2$	−0.0030	0.9290	0.0205	1.4125	0.0048
$W=2\times Q_2=2$	0.0183	0.8910	0.1085	1.5562	0.0255
$W=1\times Q_2=3$	0.0092	0.9050	0.0555	1.5085	0.0131
$W=2\times Q_2=3$	0.0030	0.9280	0.0200	1.4670	0.0046

†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 se of the estimate divided by the standard deviation sd over the estimates and the root-mean-squared error RMSE.

A second indicator Y_2 of X is created with 15% misclassification and 5% missing cases which are missing completely at random. 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 to the sample data sets, 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, with one

Table 4. Entropy R^2-values for the latent variables vehicle type and region of accident for the years 1994, 2009 and 2013

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

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 the SWOV is also applied. At the SWOV, the score that is observed in the police registry, Y_1, is assigned if it is observed. Otherwise, the score that is 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 RMSE. 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 with 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 the bootstrap) and the hierarchical assignment procedure are shown. We first discuss the performance of the MILC method in comparison with the hierarchical assignment method. The results that were obtained with hierarchical assignment especially show substantial amounts of bias for $W_{\text{ass}} \times Q_2$ compared with both implementations of the MILC method. For the unbiased parameters that were obtained when applying hierarchical assignment, RMSE is in general lower and more stable compared with the RMSE of the MILC method. 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.

A comparison of the MILC method with and without the bootstrap shows clearly that standard errors are very much underestimated when no bootstraps are performed, i.e. coverage rates are too low and the ratios between the average standard error and the standard deviation across replications are far below 1. In contrast, these ratios are larger than 1 when the bootstrap is included in the MILC method, meaning that the standard errors are somewhat overestimated. The large difference between the two approaches is caused by the fact that the statistics that 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 the bootstrap) is applied, the percentage of incorrectly classified cases is 8.6% for $X = 1$ and 20.5% for $X = 2$. With hierarchical assignment, the score on one indicator variable is used per case, and the misclassification corresponds to the misclassification that is specified in these variables. When the MILC method is applied, two indicator variables are used

Table 5. Class-specific response probabilities for latent variable vehicle type for the years 1994, 2009 and 2013

Vehicle type	*Results for 1994*		*Results for 2009*		*Results for 2013*	
	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

Table 6. Class-specific response probabilities for latent variable region of accident for the years 1994, 2009 and 2013

Region of accident	*Results for 1994*		*Results for 2009*		*Results for 2013*	
	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

to generate the variables that are under evaluation here. When assigning scores, maintaining the relationships with other variables is apparently considered more important than correctly classifying individual cases. Including interaction terms in the latent class model may possibly lead to more accurate results for the MILC method. Whether this really is so remains to be examined, though.

4. Results

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

4.1. Latent class model output

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 a latent class model on the original data set. Here it can be seen that the entropy R^2-value in 2013 increased compared with 2009 for 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 data sets, 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 with 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. The higher entropy R^2-values that were found for 1994 are likely to be caused by the fact that registration was performed more carefully and thoroughly by the police at that period, which also resulted in the lower amount of missing values, as can be seen in Table 1.

In Tables 5 and 6 the probability of correct classification for the indicators of both latent variables are shown, for the three different time points, obtained after applying a latent class model to the original data set. 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 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), whereas, in the other indicator ('police'), these cases are often missing. This can clearly be seen for the latent class N–all. Conditionally 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 with the hospital indicator for all motorized classes except the class M–other and the 'all non-motorized' category. However, in 2013 all probabilities of correct classification are higher for the hospital indicator compared with 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 in 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. Conditionally 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, conditionally on an accident truly happening in a specific region, what is the probability of also going to a hospital in that same

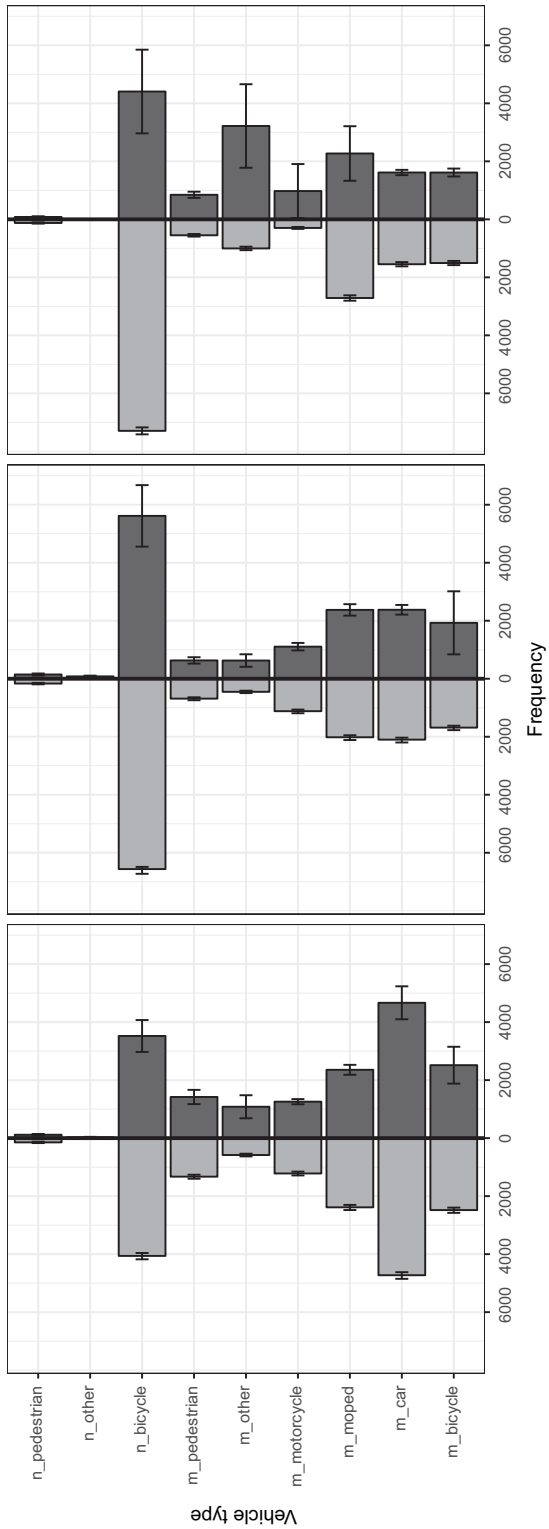

Fig. 4. Results obtained for (a) 1994, (b) 2009 and (c) 2013: on the left-hand side of each graph, the number of serious road injuries per vehicle type and corresponding 95% confidence intervals are shown when the hierarchical assignment procedure is applied (▒); on the right-hand side of each graph, pooled frequencies and 95% confidence intervals are shown when the extended MILC method is applied (■); note that the results that are presented in this paper by using the hierarchical assignment procedure are not necessarily exactly equal to official statistics produced by the SWOV

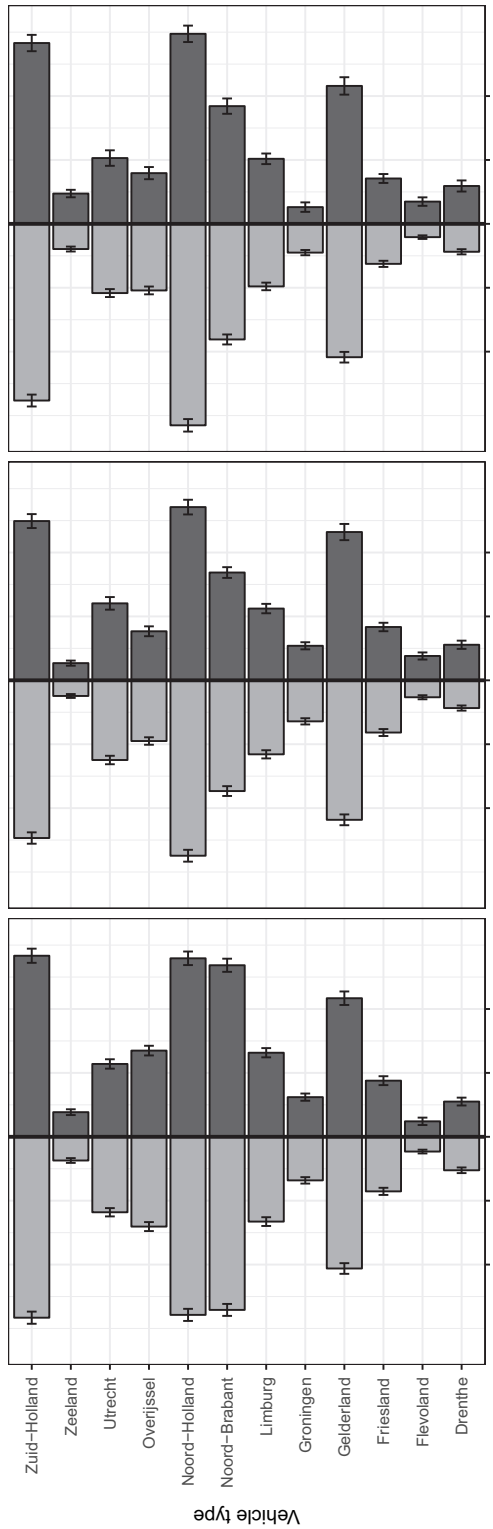

Fig. 5. Results obtained for (a) 1994, (b) 2009 and (c) 2013: on the left-hand side of each graph, frequencies of serious road injuries per region and corresponding 95% confidence intervals are shown when the hierarchical assignment procedure is applied (▒); on the right-hand side of each graph, pooled frequencies and 95% confidence intervals are shown when the extended MILC method is applied (■)

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 with other regions.

4.2. Pooled results output

In Fig. 4, the number of serious road injuries per vehicle type is shown for the three years that were investigated. For every year, the results that were obtained after applying the hierarchical assignment procedure are compared with results obtained when the extended MILC method is applied. Here, it can be seen that in general the frequencies that are obtained after applying the extended MILC method are quite similar compared with the results that are obtained after applying the hierarchical assignment procedure. When the extended MILC method is applied, the number of cases that are assigned to the category M–other is larger whereas the number of cases that were assigned to the category N–bicycle is smaller compared with 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, whereas this category was often classified differently by the police (see Table 1). At last, it can be seen that the widths of the 95% confidence intervals are substantially larger for all categories when the extended MILC method is applied. This directly results from the misclassification between the hospital and the police registry. Because of this misclassification, the latent class model is less certain about which value to assign to a specific case, resulting in differences between imputations and a larger estimate of the total variance. Note also that hierarchical assignment assumes that values that are observed in the police register are error free. Since this assumption is unlikely to be correct, uncertainty in the hierarchical assignment procedure is underestimated.

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

5. Discussion

In this paper, 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 that were obtained from this latent class model were then used to create multiple imputations of these true scores. Simultaneously, multiple imputations were created for the missing values in region of accident by using this variable as a perfectly measured indicator of the latent variable region of accident and supplementing it by specifying region of hospital as an imperfectly measured indicator.

Multiple imputations were created for vehicle type and for region of accident. All variables are now fully imputed for every case in the data set. 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 data sets 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. The results of the extended MILC method were compared with the results that were 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 that were 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.

Some issues are worth reflecting on a little further. First, it is important to note that our results heavily depend on the model assumptions that are made. In particular, the assumption is made that the classification errors are independent of covariates. 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 to be able to adapt the model. A more crucial assumption is that the missingness is at random. 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 or motorcycle issue in the 2013 data set and whether there are other issues that can be solved like this.

This particular data set contained several 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 could 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 whether and how large numbers of categories influence the results. Therefore, the number of imputations was increased and evaluated by 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 of imputations that are needed when the MILC method is applied, and how many imputations are needed to evaluate data sets with larger numbers of categories.

Furthermore, in the initial model that was 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 than those used in our application, because not all parameters can be estimated for every bootstrap sample. Alternatives to incorporate parameter uncertainty can be Bayesian Markov chain Monte Carlo sampling 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.

Lastly, it is important to note that missing values in the combined data set 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 are also 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 the MILC method 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 provide frequencies easily, to divide these frequencies further into relevant subgroups or to create statistical figures. This application showed that the initial MILC method can be extended to handle problems that are data set 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 exploit the potential of the MILC method fully for dealing with classification error problems.

Acknowledgements

The authors thank Niels Bos and Jacques Commandeur from the Institute of Road Safety Research for providing us with the data and for their useful comments on earlier versions of this paper.

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–Raphson iterations is suppressed and the number of expectation–maximization iterations is increased to 20000, following the suggestions by Vermunt *et al.* (2008).

To reduce computational 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 region of accident (`Xacc` in the Latent GOLD syntax) the value that is observed in the indicator variable region of accident (`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 by using the matrix denoted by '`a`' in the Latent GOLD syntax.

References

Boeschoten, L., Oberski, D. and de Waal, T. (2017) Estimating classification errors under edit restrictions in composite survey-register data using multiple imputation latent class modelling (MILC). *J. Off. Statist.*, **33**, 921–962.

Boeschoten, L., Oberski, D. L., Waal, T. D. and Vermunt, J. K. (2018a) Updating latent class imputations with external auxiliary variables. *Structrl Equn Modlng*, **25**, 750–761.

Boeschoten, L., Varriale, R. and Filipponi, D. (2018b) Combining multiple imputation and hidden Markov modeling to obtain consistent estimates of 'true employment status'. *Unpublished*. Tilburg University, Tilburg.

Bolck, A., Croon, M. and Hagenaars, J. (2004) Estimating latent structure models with categorical variables: one-step versus three-step estimators. *Polit. Anal.*, **12**, 3–27.

Bos, N., Stipdonk, H. and Commandeur, J. (2017) Ernstig verkeersgewonden 2016. Instituut voor Wetenschappelijk Onderzoek Verkeersveiligheid, The Hague. (Available from `https://www.swov.nl/publicatie/ernstig-verkeersgewonden-2016`.)

Dias, J. G. and Vermunt, J. K. (2008) A bootstrap-based aggregate classifier for model-based clustering. *Computnl Statist.*, **23**, 643–659.

Gerritse, S. C., Bakker, B. F., de Wolf, P.-P. and van der Heijden, P. G. (2016) Undercoverage of the population register in the Netherlands, 2010. *Discussion Paper*. Centraal Bureau voor de Statistiek, The Hague.

Graham, J. W., Olchowski, A. E. and Gilreath, T. D. (2007) How many imputations are really needed?: Some practical clarifications of multiple imputation theory. *Prevn Sci.*, **8**, 206–213.

Pankowska, P., Bakker, B., Oberski, D. and Pavlopoulos, D. (2017) Estimating employment mobility using linked data from different sources: does linkage error matter? *Conf. New Techniques and Technologies for Statistics*. (Available from `https://www.conference-service.com/NTTS2017/documents/agenda/abstracts/abstract_138.html`.)

Reiter, J. P. and Raghunathan, T. E. (2007) The multiple adaptations of multiple imputation. *J. Am. Statist. Ass.*, **102**, 1462–1471.

Reurings, M. C. B. and Bos, N. M. (2012) Ernstig verkeersgewonden in de jaren 2009 en 2010: update van de cijfers. *Report*. Stichting Wetenschappelijk Onderzoek Verkeersveiligheid, The Hague. (Available from `https://www.swov.nl/sites/default/files/publicaties/rapport/r-2012-07.pdf`.)

Reurings, M. C. B. and Stipdonk, H. L. (2009) Ernstig gewonde verkeersslachtoffers in Nederland in 1993-2008. Stichting Wetenschappelijk Onderzoek Verkeersveiligheid, The Hague. (Available from `https://www.swov.nl/publicatie/ernstig-gewonde-verkeersslachtoffers-nederland-1993-2008`.)

Reurings, M. C. B. and Stipdonk, H. L. (2011) Estimating the number of serious road injuries in the Netherlands. *Ann. Epidem.*, **21**, 648–653.

Rubin, D. B. (1987) *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Vermunt, J. K. and Magidson, J. (2015) *Upgrade Manual for Latent GOLD 5.1*. Belmont: Statistical Innovations.

Vermunt, J. K., Van Ginkel, J. R., Der Ark, V., Andries, L. and Sijtsma, K. (2008) Multiple imputation of incomplete categorical data using latent class analysis. *Sociol. Methodol.*, **38**, 369–397.
Wang, C.-P., Brown, C. H. and Bandeen-Roche, K. (2005) Residual diagnostics for growth mixture models. *J. Am. Statist. Ass.*, **100**, 1054–1076.
Wong, E. (2011) Abbreviated injury scale. In *Encyclopedia of Clinical Neuropsychology* (eds J. S. Kreutzer, J. DeLuca and B. Caplan), pp. 5–6. New York: Springer.