European Conference on Quality in Official Statistics (Q2016)
Madrid, 31 May-3 June 2016

Latent Class Multiple Imputation for multiply observed variables in a combined dataset

L. Boeschoten[1], D. Oberski[2], A.G. de Waal[3]

[1] Tilburg University, Tilburg, Netherlands & Statistics Netherlands, Den Haag, Netherlands; l.boeschoten@tilburguniversity.edu
[2] Tilburg University, Tilburg, Netherlands; d.oberski@tilburguniversity.edu
[3] Statistics Netherlands, Den Haag, Netherlands & Tilburg University, Tilburg, Netherlands; T.dewaal@cbs.nl

Abstract
Both registers and sample surveys can contain measurement error. While some errors are invisibly present, others become visible when logical relations in the data are investigated. When a variable is measured in multiple datasets within a combined dataset, we can get an indication of the errors which are invisibly present within the separate datasets. We propose a new method (MILC) based on latent class modelling that estimates the number of measurement errors in the multiple sources, and simultaneously takes impossible combinations with other variables into account. We then use the latent class model to multiply impute the latent "true" variable. Whether MILC can be applied depends on the entropy R^2 of the LC model and the type of analysis you are interested in.

Keywords: Latent class model; multiple imputation; combined dataset.

1. Introduction

National Statistical Institutes provide statistical information on many different aspects of society. This information is obtained from large datasets. A way to create these rich datasets is by utilizing already available register data and supplement them with survey data (de Waal, 2015). Utilizing already existing register data has a lot of advantages. It saves data collection and processing costs and it reduces the burden on respondents. Combined datasets are used by, among others, the System of Social Statistical Databases and the 2011 Dutch Census (Schulte Nordholt et al., 2014). These datasets often contain categorical variables.

Combined datasets give us new information, based on which we can distinguish and correct for different types of measurement errors. In this paper, we distinguish between visibly and invisibly present measurement errors. It is possible that a dataset contains measurement errors that are not visibly present. When a single dataset is added to a combined dataset, it can be possible to measure errors invisibly present within this single dataset by using a latent variable model. We do this by using multiple indicators from the different datasets within the combined dataset measuring the same latent variable. This is done using structural equation models (Scholtus & Bakker, 2013), latent markov models (Pavlopoulos & Vermunt, 2013), latent class models (Oberski, 2015) and by Guarnera & Varriale (2015). Covariate information can help with detecting the errors and sometimes even make them visibly present. Current solutions for finding and correcting visibly present errors are optimization solutions (Fellegi-Holt method for categorical data; branch-and-bound algorithm; adjusted branch-and-bound algorithm; nearest-neighbour imputation, De Waal et al., 2011, pp. 115-156) and multiple imputation (MI) solutions (latent class MI, Vermunt et al., 2008); (nonparametric Bayesian MI, Si & Reiter, 2013).

Current solutions for measurement errors in register data are tailored to handle either visibly or invisibly present errors. In addition, methods for invisibly present errors do not offer possibilities to take the errors into account in further statistical analyses. Separately solving the invisibly and visibly present errors is undesirable because your end result can be very dependent on the order in which you handle the errors. Therefore, a method is necessary that handles both types of errors simultaneously. Furthermore, uncertainty caused by the errors should be taken into account when performing further statistical analyses.

We propose a new method that simultaneously takes visibly and invisibly present errors into account by combining Multiple Imputation and Latent Class analysis (MILC). With MILC we use multiple sources to estimate measurement error and to correct for it. We start by selecting indicators measuring the same latent variable in a combined dataset. Next, MILC uses the indicators to estimate a latent class model with the number of latent classes equal to the number of categories in the indicators. By fixing covariate information in the latent class model, we can control for impossible combinations within the cases. Lastly, a variable is created taking both the visibly and invisibly present errors into account which can be used for further statistical analyses.

In the following section, we describe the MILC method in more detail. In the third section, the simulation approach and results are briefly discussed. For a more thorough description of the simulation approach and results we refer to the full paper. In the fourth section, we apply the MILC method on a combined dataset from Statistics Netherlands.

2. The MILC method

2.1. Latent Class Analysis

Latent Class (LC) analysis is typically used as a tool for analysing multivariate categorical response data (Vermunt & Magidson, 2013). We use it to estimate both visibly and invisibly present measurement error in categorical variables. We have multiple datasets linked on a unit level, containing the same variable, which we use as indicators measuring one latent variable. This latent variable can be seen as the “true variable” in this situation, and is denoted by *X*. For example, we have *l* dichotomous indicators ($Y_1, \ldots, Y_l$) measuring the variable *home ownership* (*1*= “own”, *2*= “rent”) in multiple datasets linked on person level. Differences between the responses of a person are caused by invisibly present measurement errors in one (or more) of the indicators. Because the indicators all have an equal number of categories ($D_1, \ldots, D_l$), the number of categories in the “true variable” *X*, *C*, is equal to $D_1 = \ldots = D_l$. A specific category is denoted by *x*, where $x=1,\ldots,C$.

The LC model starts with the fact that the probability of obtaining response pattern *y*, $P(Y = y)$, is a weighted average of the *C* class-specific probabilities $P(Y = y|X = x)$. Furthermore, the assumption is being made that the observed indicators are independent of each other given an individual's score on the latent “true variable”. In combination, this yields the following model for $P(Y=y)$:

$$P(Y = y) = \sum_{x=1}^{C} P(X = x) \prod_{l=1}^{L} P(Y_l = y_l \mid X = x). \qquad (1)$$

In equation 1, only the indicators are used to estimate the likelihood of being in a specific latent class. However, it is also possible to use covariate information, and to impose a restriction using a covariate to make sure that we do not create a combination of an LC and a score on the covariate that is in practice impossible. In this paper, we distinguish between three models. In the *unconditional model*, only indicators and covariates without restrictions are used. In the *conditional model*, we take the restriction variable into account as a covariate. In the *restricted conditional model*, we not only take the restriction covariate into account, we also fix the cell of the impossible combination to zero.

The LC model assigns units to latent classes representing scores on the “true” latent variable X. We can distinguish between “true” latent variable X and the estimated class variable W (Bakk et al., 2014). Scores assigned to W can be obtained by estimating the posterior membership probability for each unit. Corresponding to the three models we distinguished for estimating the likelihood, there are three models for obtaining the posterior probability.

2.2. Multiple Imputation

We use multiple imputation (MI) to estimate W. We create m empty variables ($W_1, \ldots, W_m$) in the dataset and impute them by drawing one of the LCs using their posterior membership probabilities. The differences between the m imputations reflect the uncertainty about the latent variable caused by conflicts between the observed indicators. From the m imputed variables we obtain estimates of interest and pool them using Rubin's rules (Rubin, 1987, p.76). By performing MI, we take uncertainty into account caused by missing or conflicting data. Since we want to take parameter uncertainty into account as well, we first take m bootstrap samples from the dataset. Next, we generate one latent class model for each of the m datasets (Van der Palm et al., 2013). Now we impute $W_1, \ldots, W_m$ by using m latent class models.

3. Simulation

3.1. Simulation approach

To empirically evaluate the performance of MILC, we conduct a simulation study using R.We start by creating an infinite population using Latent Gold, containing three dichotomous indicators (*Y1; Y2; Y3*) measuring the latent dichotomous variable (*X*); one dichotomous covariate (*Z*) which has an impossible combination with a score of the latent variable; one other dichotomous covariate (*Q*). Datasets are generated by making use of the restricted conditional model. The following simulation conditions are used:

- Classification probabilities*: 0.70*; *0.80*; *0.90*; *0.95*; *0.99*.
- *P(Z = 2)*: *0.01*; *0.05*; *0.10*; *0.20.*
- Sample size: *1,000*; *10,000*.
- Logit coefficients of *X* regressed on *Q*: *-0.2007*; *0.2007*; *0.6190*.
- Number of imputations: *5*; *10*; *20*.

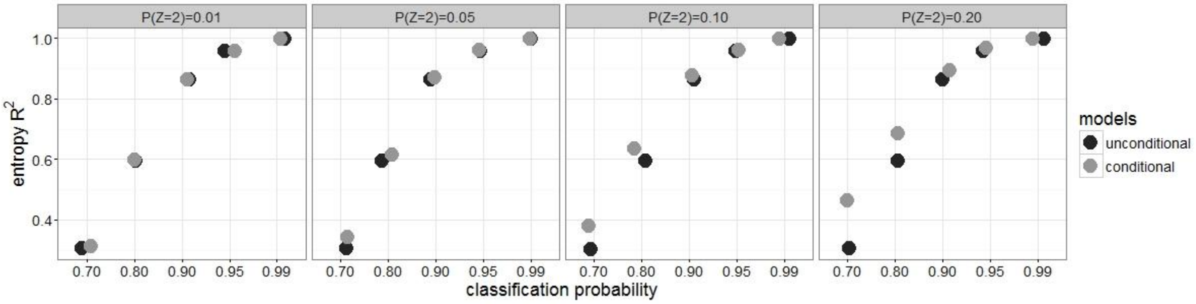

Figure 1 *Entropy R^2 of the unconditional and conditional model with different values for the classification probability and P(Z=2). The restricted conditional model has the same entropy R^2 as the conditional model because the models contain the same variables.*

The entropy R^2 measures how well one can predict class membership based on the observed variables (indicators and covariates). The closer to *1*, the better the predictors (Vermunt & Magidson, 2013). In **Figure 1** we see the entropy R^2 of the models under different values for *P(Z=2)* and classification probabilities. The conditional and the restricted conditional model have the same entropy R^2 because they contain the same variables.

3.2. Simulation conclusions

In the simulation we focused on the relation between imputed latent variable *W* and restriction covariate *Z,* and on the relation between imputed latent variable *W* and covariate *Q*. We investigated the bias of the estimates, the coverage of the 95% confidence intervals of the estimates and the average standard error of the estimates divided by the standard deviation over the estimates. The main conclusion we draw from the simulation results is that low bias, appropriate coverage and correct estimation of standard errors is strongly related to the entropy R^2 of the LC model. The entropy R^2 is influenced by the classification probabilities and the marginal distributions of the variables. For different types of analysis a different entropy R^2 is required. A logistic regression can already be done with an entropy R^2 of *0.60*, while an entropy R^2 of *0.90* is required when you are interested in estimating a cross table containing impossible combinations.

Application

4.1. Data and models

We apply the MILC method on a combined dataset consisting of data from the LISS (Longitudinal Internet Studies for the Social sciences) panel administered by CentERdata (Tilburg University, The Netherlands) and a register from Statistics Netherlands. We use two indicators for the imputed "true" latent variable *home-owner/renter* and a variable measuring whether someone receives *rent benefit*. You can only receive this if you do not own but rent a house. If we want to make a cross-table between the imputed latent variable *home-owner/ renter* and *rent benefit*, there should be zero persons in the cell "home-owner *x* receiving rent benefit". Furthermore, we are also interested in the question whether married individuals more often live in a house they own compared to non-married individuals. Therefore, we need a variable indicating whether a person is married or not in the latent class model as a covariate. We combined variables from three datasets for this study. For a more thorough description of the recoding of the variables, we refer to the full paper.

- **Registration of addresses and buildings (BAG):** This register originates from the municipalities and data is from January 2013 on *3011* individuals. We used a variable indicating whether a person "owns"/ "rents or other" the house he or she lives in.
- **LISS background study:** This survey is from January 2013, here we also have *3011* individuals. We used a variable indicating whether someone is "married" / "not married" and a variable indicating whether someone is an "owner"/"tenant or other".
- **LISS housing study:** This survey is from June 2013 and we use a variable indicating whether someone "receives rent benefit"/ "does not receive rent benefit". Here we only

have *779* observations. This is caused by the fact that another variable, indicating whether someone rents their house, was used as a selection variable.

We apply the MILC method to impute the latent variable home *owner/renter* using the unconditional model, the conditional model and the restricted conditional model. The models have an entropy R^2 of approximately *0.93,* which is comparable to conditions we tested in the simulation study and in which the MILC method appeared to work very well.

4.1. Being a home owner or renter and receiving rent benefit

	P(own | rent benefit)		P(rent | rent benefit)	
	estimate	95% CI	estimate	95% CI
BAG register	0.0051	[0.0001; 0.0102]	0.2953	[0.2632; 0.3273]
LISS background	0.0104	[0.0032; 0.0175]	0.2889	[0.2568; 0.3209]
unconditional	0.0013	[0.0007; 0.0018]	0.2940	[0.2934; 0.2945]
conditional	0.0064	[-0.0263; 0.0391]	0.2888	[0.2561; 0.3215]
restricted conditional	0.0000	-	0.2953	[0.2624; 0.3281]
	P(own | no rent benefit)		P(rent | no rent benefit)	
	estimate	95% CI	estimate	95% CI
BAG register	0.0552	[0.0391; 0.0713]	0.6444	[0.6107; 0.6781]
LISS background	0.0285	[0.0167; 0.0403]	0.6723	[0.6391; 0.7054]
unconditional	0.0154	[0.0149; 0.0159]	0.6842	[0.6837; 0.6848]
conditional	0.0154	[-0.0173; 0.0481]	0.6842	[0.6515; 0.7169]
restricted conditional	0.0205	[-0.0123; 0.0534]	0.6791	[0.6462; 0.7119]

Table 1 *The blocks represents the (pooled) proportions of the variable own/rent for persons (not) receiving rent benefit. Within each block, the first two rows represent the BAG register and the LISS background survey. The last three rows represent the three models used to apply the MILC method. For each proportion a (pooled) estimate and a (pooled) 95% confidence interval is given.*

We can see from the cell totals in **Table 1** whether individuals who say to own their home, also receive rent benefit, something which is not allowed. These discrepancies can be caused by the fact that people make mistakes when filling in a survey, or because people were moving during the period the surveys took place. If we investigate the cell proportions estimated by the MILC method, we see that both the conditional and the unconditional model replicate the structure of the indicators very well, but that individuals are still assigned to the cell of the impossible combination. To estimate this correctly, we need the restricted conditional model.

4.2. The relation between marriage and owning a house

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

Table 2 *The first two rows represent the BAG register and the LISS background survey. The last three rows represent the three models used to apply the MILC method. The columns represent the (pooled) estimate and 95% confidence interval of the intercept and the logit coefficient of the variable owning/renting a house.*

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

4. Discussion

In this paper we introduced the MILC method, which combines latent class analysis and multiple imputation to obtain estimates for variables of which we had multiple indicators in a combined dataset. We distinguished between invisibly present and visibly present errors, and argued the need for a method that takes them into account simultaneously. We evaluated the MILC method in terms of its ability to correctly take impossible combinations and relations with other variables into account. The performance of MILC appeared to be mainly dependent on the entropy R^2. For different types of analysis a different entropy R^2 is required. A logistic regression can already be done with an entropy R^2 of *0.60,* while an entropy R^2 of *0.90* is required when you are interested in estimating a cross table containing impossible combinations. An example of a combined dataset containing data from the LISS panel and the BAG register show to have adequate entropy R^2 and decent results. Furthermore, we investigated the MILC method using three different types of models, the unconditional model, the conditional model and the restricted conditional model. All models can potentially be used when using the MILC method in practice. However, if there are impossible combinations within the data that the researchers needs to be taken into account, only the restricted conditional model is appropriate. In light of our main findings the MILC method can be seen as an appropriate alternative to methods previously used for handling visibly and invisibly present errors, which was done separately.

In an extension, the MILC method could also be used to multiply impute missing values within the combined dataset. In this paper, focus has not been laid on the problem of linkage error as well. The MILC method can probably not prevent linkage errors, but it should be investigated how the MILC method can take errors caused by linkage into account. In another extension, more attention can also be paid to the covariates. In the current approach, we assume that the covariates do not contain measurement error. We could adapt the method in such a way that we assume a specific amount of measurement error in the covariate, or we could use multiple indicators for the “true” latent covariate if available. Furthermore, variables corresponding to all relationships that you want to investigate should be included as covariates in the LC model. By adapting the three step method (Bakk, 2014) to make it applicable to the MILC method, we could also investigate relations of the imputed latent variable and other variables, not taking into account as covariates in the LC model when the MILC method was applied. In conclusion, when researchers have multiple indicators of the same variable in a combined dataset, it is now clear under what conditions the researcher can appropriately use the MILC method to multiply impute the latent variable.

5. References

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

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

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

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

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

Pavlopoulos, D. & Vermunt, J.K. (2015). Measuring temporary employment. Do survey or register data tell the truth? *Survey Methodology, 41* (1), 197-214.

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

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

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

Schulte Nordholt, E., Van Zeijl, J, Hoeksma, L. eds. (2014) *Dutch Census 2011, Analysis and Methodology*, Statistics Netherlands, The Hague/Heerlen

Van der Palm, D. W., Van der Ark, L. A., & Vermunt, J. K. (2016). Divisive latent class modeling as a density estimation method for categorical data. Journal of Classification, 1-21.

Vermunt, J. K., & Magidson, J. (2013). *Lg-syntax user's guide: Manual for latent gold R 5.0 syntax module* October 31, 2013.

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