Department of Methodology and Statistics

Utrecht University, the Netherlands

Research Report:

Comparing MILC and tree-MILC for estimating and correcting multiple types of errors in combined datasets

December 2021

Daniëlle Remmerswaal

4435958

d.m.remmerswaal@uu.nl

Supervisors:

Prof. Dr. Ton de Waal (CBS)

Dr. Laura Boeschoten (UU)

Utrecht University

Centraal Bureau
voor de Statistiek

1 Introduction

National Statistical Institutes (NSIs) provide data and produce official statistics for the public. Whereas surveys have largely replaced the labour-intensive censuses for data-collection, increasingly more administrative sources are now being used for statistical purposes as to reduce costs and the response burden (UNECE, 2011). Registers are often falsely assumed to be error-free, while research has demonstrated this is not the case. See among others the work of Delden et al. (2016); Pankowska et al. (2017). Administrative registers can contain various errors, but NSIs are not always aware of them since registers are generally not created and maintained by statistical agencies themselves, but by external owners for administrative purposes (Reid et al., 2017; UNECE, 2011).

In general, there is a lack of theories to evaluate the errors in administrative registers and the resulting quality of statistics based on them, especially as compared to surveys. Errors in data can lead to biased and inconsistent statistical estimates. Thus to produce statistics of high quality it is important to estimate and correct for errors. In the next sections, we will start by elaborating on different types of errors, and then on methods to correct for them. We first focus on single-type errors in single-source situations, then expand to multiple-source situations, and finally address situations with multiple types of errors.

The Total Survey Error (TSE) framework identifies various types of measurement and representation errors that contribute to the deviation of an estimate from its true parameter value. The TSE framework is originally developed for surveys; see Groves and Lyberg (2010) for an historical overview, and Biemer (2010) for more background information. The TSE framework distinguishes two types of errors: measurement errors of the variables (the discrepancies between what is intended to be measured and what is measured), and representation errors of statistical objects (the discrepancies between the set that is intended to be measured and the set that is measured). Zhang (2012) generalised the TSE framework to be able to use it to assess the quality of administrative sources. The main differences can be found on the 'Representation side': coverage error is replaced with frame error, sampling error with selection error, nonresponse error with missing/redundancy error, and (survey) adjustment error is excluded (see Figure 1).

Numerous methods are being used to estimate and correct for errors. In this section we discuss some of those methods for single-source data. We limit our research to errors in categorical variables, so-called classification errors. Coverage errors can be corrected with capture-recapture methods (Gerritse et al., 2015). Unit nonresponse errors are often corrected with weighting methods (Bethlehem et al., 2011), item nonresponse with multiple imputation (Van Buuren, 2012). Measurement errors are localised by looking at impossible combinations with other variables, and

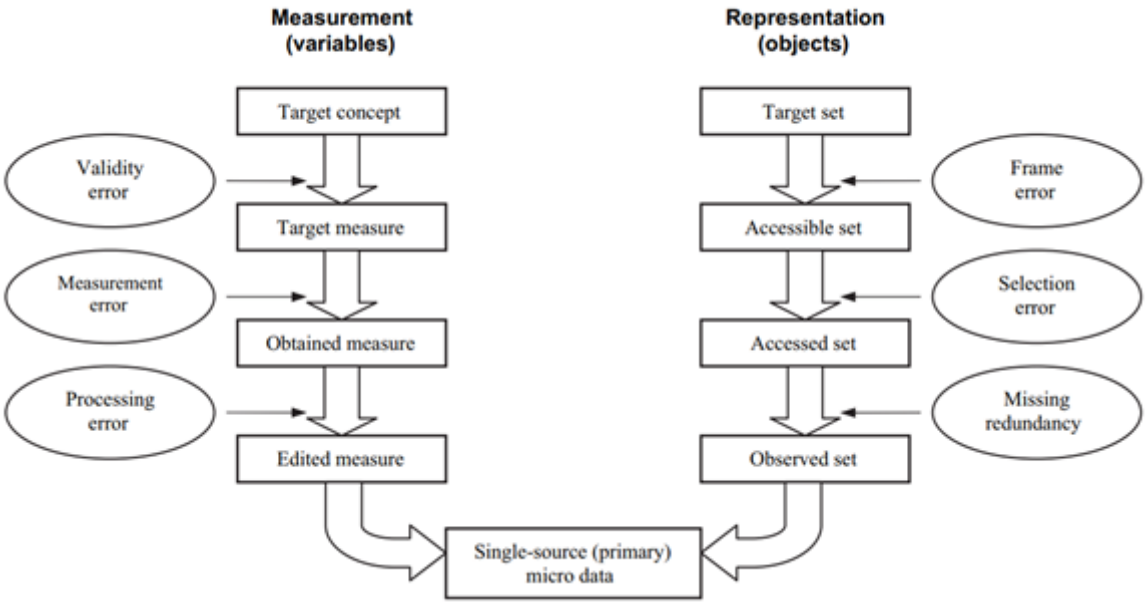

Figure 1: TSE quality assessment framework

corrected with minimal adjustments (based on the Fellegi-Holt paradigm) to comply with edit rules. For more information on statistical data editing we refer to the work of De Waal et al. (2011).

Comparing multiple data sources on the same concept creates more opportunities for error detection and correction. Conveniently, NSIs are often able to link variables from different sources on the unit-level together in a combined dataset. An important method that makes uses of multiple data sources is the latent variable modelling (LVM) approach; e.g. see Biemer (2011). LVM, with latent class analysis (LCA) being specifically for categorical variables, uses repeated measurements of the same variable (Oberski, 2017). Those measurements, either from the same source over time or from several sources at the same time-point, are then used as the 'indicators' of a latent, unobserved, 'true' variable. An advantage is that, in contrary to alternative methods to correct errors, there is no need for error-free validation data. In LVM none of the indicator variables is assumed to be error-free, which resembles reality more. For data measured multiple times, Hidden-Markov Models (HMM), a special case of latent variable modelling, can be used (Pavlopoulos & Vermunt, 2015; Pankowska, 2020). However, in our situation we do not have multiple measurements over time, but at the same time-point. The Multiple Imputing Latent Class (MILC) method developed by Boeschoten et al. (2017) combines a latent class model to estimate the errors with multiple imputation, to also correct the estimands for those errors.

In practice when producing statistics, there are often multiple types of errors, which should be distinguished from each other as to correct for them accordingly. We are interested in situations

with measurement and selection errors occurring simultaneously, a combination of errors from each side of the quality assessment framework (see figure 1). Measurement errors are misalignments between the target measure and the obtained measure, for example caused by people giving wrong answers and computer systems corrupting values or introducing ambiguity (Reid et al., 2017). Selection errors are misalignments between the accessible dataset in theory and the accessed dataset in practice, for example due to delays in reporting situation changes. We will now discuss an example situation where the two types of errors can both be encountered. Businesses are registered at the Chamber of Commerce, in which the address and business activity are recorded. It can happen that a business moved without timely reporting it (selection error). Measurement errors can also occur: a business can be active in multiple regions or multiple business sectors introducing ambiguity and errors. So when producing for example a statistic on the employment rate per sector per region, selection and measurement errors can occur simultaneously.

It has not been researched extensively whether current practices are applicable on situations where misclassifications can be identified as being caused by different error types. In survey research there is an approach for multiple types of measurement errors. The multitrait-multimethod (MTMM) approach (Oberski et al., 2015) studies systematic (caused by survey design) and random measurement error simultaneously. However, this is not applicable to our situation with register data and errors on both sides of the data quality framework (see figure 1).

The MILC method is developed to estimate and correct for one type of error. We are interested in its performance on situations with multiple types of errors. An unexplored potential solution for this situation, is an approach used in Latent Class Tree (LCT) analysis. In LCT analysis, developed by Van den Bergh (2018), LC models are applied sequentially on subsets. In this thesis we expand MILC with an LCT approach, and name it 'tree-MILC'. Tree-MILC is expected to produce more accurate statistical estimates due to its potential ability to identify and correct for multiple error types separately. A more detailed explanation can be found in section 2.2.

This project aims to investigate whether the combination of the MILC-method with a latent class tree approach, tree-MILC, is an appropriate alternative to the existing MILC-method, to measure and correct for a combination of a measurement and a representation error. How well MILC and tree-MILC perform to estimate and correct for various levels of errors, will be investigated with a simulation study, and evaluated with bias .

2 Methods

2.1 MILC

The MILC method, developed by Boeschoten et al. (2017), is used to identify and correct classification errors. MILC is a combination of latent class (LC) analysis, see e.g.Vermunt et al. (2008), and multiple imputation (MI), see Rubin (1987). The typical use of latent variable models is for analysing multivariate response data. In contrast, the LC model in MILC is used to estimate the "true" category of units in combined datasets, that consist of multiple unit-linked categorical variables from different sources measuring the same attribute. While MI is typically used for missing data, in MILC multiple imputation is used to impute the correct class for all observations; see Vidotto et al. (2015) for an overview of the use of LC models for imputation.

Here, an overview of the five MILC steps is given. The steps are visualised in figure 2. In the next sections, each step will be explained more elaborately.

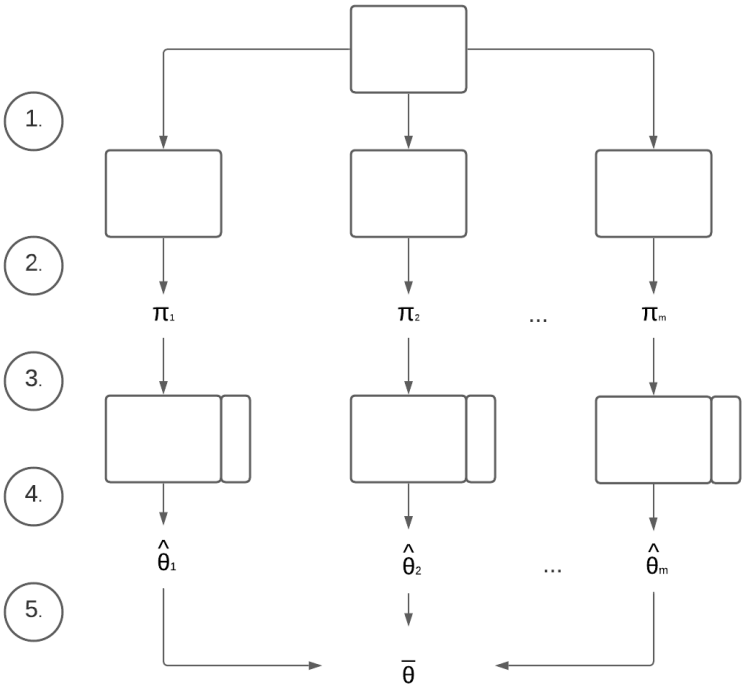

Figure 2: Graphical overview of the MILC method. At step 1 m bootstrap samples are taken from a dataset containing multiple indicator variables. At step 2 an LC model is applied to each bootstrap sample and probabilities (denoted by π) are obtained. At step 3 m imputations (denoted with the vertical bar) for the latent variable are created and placed next to the data. At step 4 the estimates of interest (denoted with θ) are calculated from the imputations. At step 5 the estimates are pooled.

Steps of MILC method

1. m bootstrap samples are taken with replacement from a dataset with unit-linked categorical variables from different sources measuring the same attribute. With m being equal to the number of imputations (see step 3).

2. We apply an LC model with C classes on each of those m bootstrap samples. With C being equal to the number of categories of the variable of interest.

3. By sampling from the obtained posterior membership probabilities we create m imputations of the latent classes and store them in a dataset with the original variables.

4. Calculate the estimate of interest θ for each of the m imputations

5. Pool the m estimates with Rubin's pooling rules

2.1.1 Bootstrap

To reflect parameter uncertainty caused by classification errors of the data in the estimate of the total variance, we take bootstraps from our original data. Appropriate datasets consist of multiple unit-linked categorical variables from different sources measuring the same attribute. We use a multinomial distribution to sample from, with the frequency distributions of each combination of categorical variables, the response patterns, as probabilities. This step results in m bootstrap samples of the same size as the original dataset. With m being the number of imputations we want to create in step 3.

2.1.2 Latent Class Model

On each of the m bootstrap samples we apply an LC model. The basic idea behind latent class analysis (LCA) is to find latent variables based on the correlation structure of observed categorical "indicator" variables. Each of the J categorical variables has C_j possible outcomes for individuals $i = 1, \ldots, N$. In our specific case each source measures the same attribute and has the same number of categories, so $C_j = C$. Each individual has a response on each of the J variables Y_j $(j = 1, \ldots, J)$; for now we consider a situation without missing data. The vector of responses $\mathbf{Y}$ for one individual is called a response pattern. There are $\prod_{j=1}^{J} C_j$, and in our case C^J possible response patterns. In general, the number of latent classes is not known and has to be determined by the user of the LC model. When an LC model is used for correction of classification errors the number of latent classes is also fixed to C.

The two key model assumptions for latent class analysis are the local independence assumption and the mixture assumption. By assufming that responses within one class are independent of each other, we assume the correlation between them was solely caused by their class membership. With this assumption we can calculate the likelihood of a response pattern $\mathbf{Y}$ occurring given the observation is part of class C:

$$P(\mathbf{Y}|X = c) = \prod_{j=1}^{J} P(Y_j|X = c) \tag{1}$$

By assuming that the probability of obtaining a specific response pattern is the weighted sum of all C class-specific conditional response probabilities (mixture assumption) we can calculate the probability density function of a response pattern $\mathbf{Y}$ across all classes:

$$P(\mathbf{Y}) = \sum_{c=1}^{C} P(X = c) P(\mathbf{Y}|X = c) \tag{2}$$

By combining the mixture and local independence assumption and using Bayes' rule, we are able to calculate the posterior membership probabilities of each response pattern, and thus for each individual:

$$P(X = c|\mathbf{Y} = \mathbf{y}) = \frac{P(X = c) P(\mathbf{Y} = \mathbf{y}|X = c)}{P(\mathbf{Y} = \mathbf{y})} \tag{3}$$

2.1.3 Multiple Imputation

While multiple imputation is typically used for imputing missing data, it can also be used to impute all values. This is the case in MILC (Boeschoten et al., 2017), but also in overimputation (Blackwell et al., 2017). In general, there are multiple methods to assign individuals to classes based on the posterior membership probabilities; see for an overview Bakk (2015). The most straightforward method to impute the classes with the posteriors is modal assignment: individuals are assigned to the class with the highest posterior membership probability based on their response pattern. In MILC, a stochastic (proportional) assignment is used; values are assigned to a class by sampling from the posterior membership probabilities obtained by the LC step. We create m empty variables and impute them with one of the classes by sampling from the posteriors of each of the m bootstrap samples. Five imputations are usually sufficient, as demonstrated by Boeschoten et al. (2017).

2.1.4 Calculation of estimates

There are various types of estimates possible to calculate. For categorical variables we can calculate frequency distributions of single variables and of relationships between a variable and a covariate.

We are (for now) interested in the first-mentioned, the relative sizes of each class, calculated as the proportion of units assigned to each class. We calculate this estimate for each of the m imputations, which results in m estimates.

2.1.5 Pooling of estimates

We pool the estimates of the imputations with the standard pooling rules for imputation: Rubin's pooling rules (Rubin, 1987). The pooled proportions are obtained by calculating the mean of the proportions for each imputation:

$$\bar{\theta}_M = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m \tag{4}$$

The variance of the estimate is calculated by adding the within variance to a weighted between variance:

$$T_M = \bar{W}_M + \left(1 + \frac{1}{M}\right) \bar{B}_M \tag{5}$$

With the within variance:

$$\bar{W}_M = \frac{1}{M} \sum_{m=1}^{M} W_m \tag{6}$$

And the between variance:

$$\bar{B}_M = \frac{1}{M-1} \sum_{m=1}^{M} \left(\hat{\theta}_m - \bar{\theta}_M\right)^2 \tag{7}$$

2.2 Tree-MILC

Tree-MILC is potentially a useful method to distinguish the different (selection and measurement) errors in a dataset, and obtain the correct estimates. The idea of implementing a tree-step originates from latent class tree (LCT) analysis, as developed by Van den Bergh (2018).

In LCT analysis, the data is split and structured into classes by method of a sequential comparison of 1- and 2-class models. The top-down approach continues until the information criterion (e.g. BIC) no longer chooses 2-class models over 1-class models. The splits are performed based on the posterior class membership probabilities of each new class (child nodes) conditional on the class before the splitting (parent node). The advantages of this method of latent class analysis are that it provides clear insight in splitting of the classes and the structure of the data, and on how models with different number of classes are related to each other.

LCT analysis is an exploratory method mainly useful for determining the number of classes. In our situation the number of classes is fixed and dependent on the categories in the variables. Therefore we do not use the LCT as intended, but use the idea of a tree-structure to distinguish

between the two types of errors. Tree-MILC has the same procedure as MILC, with an extra LC model and imputation step before the estimates are calculated (step 6). The first LC model (step 2) and imputation (step 3) are applied on the whole dataset to estimate the selection errors. The second LC model (step 4) and imputation (step 5) are applied on a subset (a child node), based on the selection error imputations, to estimate the classification errors. Subsequently, the two LC models extract a different number of classes to impute, as will be elaborated on below. The process is visualised in figure 3.

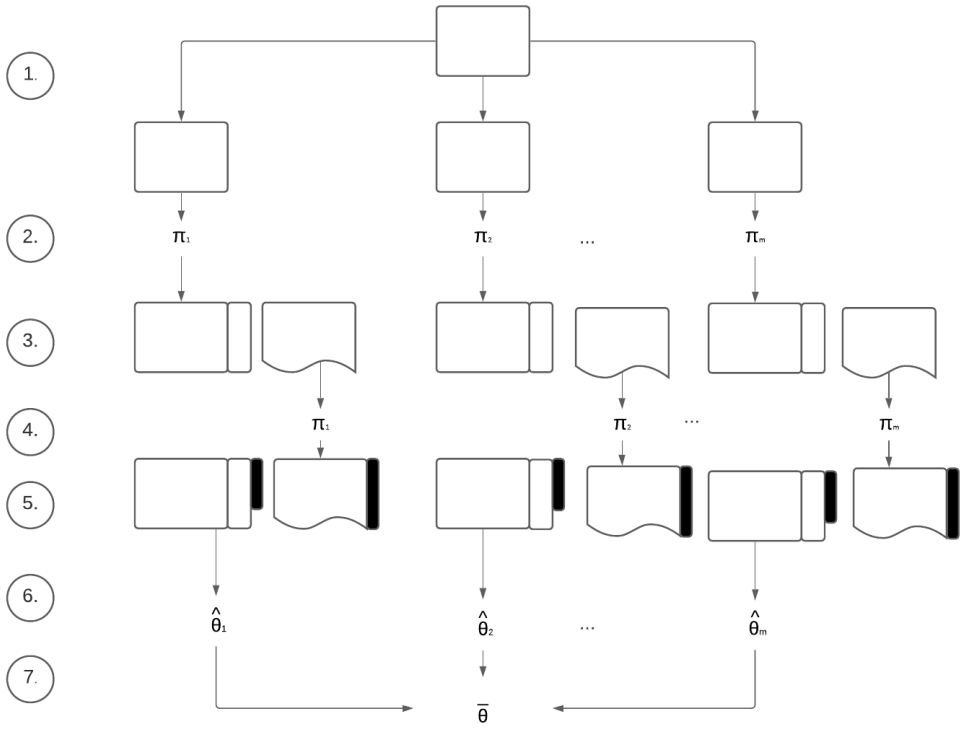

Figure 3: Graphical overview of tree-MILC. At step 1, m bootstrap samples are taken from a dataset containing multiple indicator variables. At step 2, an LC model is applied to each bootstrap sample and probabilities (denoted by π) are obtained. With the imputations (denoted with the vertical bar) created in step 3, subsets are created to which the LC model in step 4 is applied. At step 5 imputations are calculated for the subset. At step 6 estimates of interest (denoted with θ) are calculated for the combined imputations of step 3 and 5. At step 7 the estimates are pooled.

We will now elaborate on some differences with MILC, the procedure concerning the application of the second LC and MI step. In step 2 an LC model is applied with S classes (in our case $S = 2$) to estimate the selection error. By sampling from the obtained posterior membership probabilities from the m LC models, all cases are assigned to one of the S classes. Based on those imputations we create a subset for each of the bootstrap samples that contain the cases of our target set. On those subsets we apply an LC model with C classes, with C being the categories prone to classification error. We then impute the classes by sampling from the obtained posteriors and

place them next to the original dataset. We replace the imputations from step 3 where necessary, as the imputations in step 5 are sub-categories or 'sub-classes' of the 'selected' class.

3 Simulation study

To compare the performance of MILC and tree-MILC to estimate and correct for two types of errors, we perform a simulation study. The simulation study is performed with R (R Core Team, 2021), and in particular, the R package poLCA (Lewis & Linzer, 2011) will be used for the data simulation and the latent class analysis. The simulation study will be performed on simulated datasets with the following features: three indicator variables with four categories each. We will vary the levels of selection error (5% and 20%) and classification error (5% and 20%). The four simulation conditions can be seen in table 1. After MILC and tree-MILC are both applied on the

Table 1: Simulation conditions

		measurement error	
		5%	20%
selection error	5%	condition A	condition B
	20%	condition C	condition D

simulated datasets, the obtained pooled estimates (the proportion of units assigned to each class) are compared to the theoretical estimates from the original simulated datasets. The differences are denoted as the bias, and are used to evaluate the performance of the methods.

References

Bakk, Z. (2015). *Contributions to bias adjusted stepwise latent class modeling* (Doctoral dissertation, Tilburg University). Retrieved from `https://research.tilburguniversity.edu/en/publications/contributions-to-bias-adjusted-stepwise-latent-class-modeling`

Bethlehem, J., Cobben, F., & Schouten, B. (2011). *Handbook of nonresponse in household surveys.* Hoboken, NJ: John Wiley and Sons. (ISBN: 978-0-470-54279-8) doi: 10.1002/9780470891056.ch8.

Biemer, P. (2010). Total survey error: Design, implementation, and evaluation. *The Public Opinion Quarterly*, *74*(5), 817-848. doi: 10.2307/40985407.

Biemer, P. (2011). *Latent class analysis of survey error.* Hoboken, NJ: John Wiley and Sons. (ISBN: 9780470289075)

Blackwell, M., Honaker, J., & King, G. (2017). A unified approach to measurement error and missing data: Overview and applications. *Sociological Methods and Research*, *46*(3), 303-341. doi: 10.1177/0049124115585360.

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*, 921-962. doi: 10.1515/jos-2017-0044.

Delden, A., Scholtus, S., & Burger, J. (2016). Accuracy of mixed-source statistics as affected by classification errors. *Journal of official statistics*, *32*, 619-642. doi: 10.1515/jos-2016-003.

De Waal, T., Pannekoek, J., & Scholtus, S. (2011). *Handbook of statistical data editing and imputation.* Hoboken, NJ: John Wiley and Sons. (ISBN: 978-0-470-54280-4)

Gerritse, S. C., van der Heijden, P. G., & Bakker, B. F. (2015). Sensitivity of population size estimation for violating parametric assumptions in log-linear models. *Journal of Official Statistics*, *31*(3), 357–379. doi: 10.1515/jos-2015-0022.

Groves, R. M., & Lyberg, L. E. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly*, *74*(5), 849-879. doi: 10.1093/poq/nfq065.

Lewis, J., & Linzer, D. (2011). polca: An r package for polytomous variable latent class analysis. *Journal of Statistical Software*, *42*(10), 1–29. doi: 10.18637/jss.v042.i10.

Oberski, D. (2017). Estimating error rates in an administrative register and survey questions using a latent class model. In *Total survey error in practice* (p. 339-358). Hoboken, NJ: John Wiley Sons, Ltd. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119041702.ch16` doi: 10.1002/9781119041702.ch16.

Oberski, D., Kirchner, A., Eckman, S., & Kreuter, F. (2015). Evaluating the quality of survey

and administrative data with generalized multitrait-multimethod models. *Journal of the American Statistical Association*, *112*(2), 1477-1489. doi: 10.1080/01621459.2017.1302338.

Pankowska, P. (2020). *Measurement error: estimation, correction, and analysis of implications* (Doctoral dissertation, Vrije Universiteit Amsterdam). Retrieved from `https://research.vu.nl/ws/portalfiles/portal/111403652/853908.pdf`

Pankowska, P., Bakker, B., Oberski, D., & Pavlopoulos, D. (2017). Reconciliation of inconsistent data sources by correction for measurement error: The feasibility of parameter re-use. *Statistical Journal of the IAOS*, *34*(3), 1-13. doi: 10.3233/SJI-170368.

Pavlopoulos, D., & Vermunt, J. (2015). Measuring temporary employment. do survey or register data tell the truth? *Survey Methodology*, *14*, 197-214. Retrieved from `https://research.vu.nl/ws/portalfiles/portal/1039950/pavlopoulso+vermunt.pdf`

R Core Team. (2021). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Reid, G., Zabala, F., & Holmberg, A. (2017). Extending tse to administrative data: A quality framework and case studies from stats nz. *Journal of Official Statistics*, *33*(2), 477-511. doi: 10.1515/jos-2017-0023.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: John Wiley and Sons. (ISBN: 9780470316696)

UNECE. (2011). Using administrative and secondary sources for official statistics: a handbook of principles and practices. Retrieved from `http://digitallibrary.un.org/record/719971` (United Nations Economic Commission for Europe.)

Van Buuren, S. (2012). *Flexible imputation of missing data. second edition.* Boca Raton, Florida: CRC Press. (ISBN: 9781032178639)

Van den Bergh, M. (2018). *Latent class trees* (Doctoral dissertation, Tilburg University). Retrieved from `https://pure.uvt.nl/ws/portalfiles/portal/20026000/Van_Den_Bergh_Latent_05_01_2018.pdf`

Vermunt, J. K., Van Ginkel, J. R., Van Der Ark, L. A., & Sijtsma, K. (2008). Multiple imputation of incomplete categorical data using latent class analysis. *Sociological Methodology*, *38*(1), 369-397. doi: 10.1111/j.1467-9531.2008.00202.

Vidotto, D., Kaptein, M., & Vermunt, J. (2015). Multiple imputation of missing categorical data using latent class models: State of art. *Psychological Test and Assessment Modeling*, *57*(4), 542–576. Retrieved from `https://pure.uvt.nl/ws/portalfiles/portal/9420410/Vidotto_Vermunt_Multiple_imputation_of_missing.pdf`

Zhang, L.-C. (2012). Topics of statistical theory for register-based statistics and data integration.

Statistica Neerlandica, *66*(1), 41-63. doi: 10.1111/j.1467-9574.2011.00508.