Journal of Survey Statistics and Methodology (2019) ***0**, 1–25*

COMBINING MULTIPLE IMPUTATION AND HIDDEN MARKOV MODELING TO OBTAIN CONSISTENT ESTIMATES OF EMPLOYMENT STATUS

LAURA BOESCHOTEN*
DANILA FILIPPONI
ROBERTA VARRIALE

Recently, a method was proposed that combines multiple imputation and latent class analysis (MILC) to correct for misclassification in combined data sets. A multiply imputed data set is generated which can be used to estimate different statistics of interest in a straightforward manner and can ensure that uncertainty due to misclassification is incorporated in the estimate of the total variance. In this article, MILC is extended by using hidden Markov modeling so that it can handle longitudinal data and correspondingly create multiple imputations for multiple time points. Recently, many researchers have investigated the use of hidden Markov modeling to estimate employment status rates using a combined data set consisting of data originating from the Labor Force Survey (LFS) and register data; this combined data set is used for the setup of the simulation study performed in this article. Furthermore, the proposed method is applied to an Italian combined LFS-register data set. We demonstrate how the MILC method can be extended to create imputations of scores for multiple time points and thereby show how the method can be adapted to practical situations.

KEYWORDS: Combined survey-register data; Employment status; Hidden Markov model; Multiple imputation.

LAURA BOESCHOTEN is with Tilburg University, Warandelaan 2, 5037 AB, Tilburg, The Netherlands and Statistics Netherlands, Henri Faasdreef 312, 2492 JP Den Haag, The Netherlands. DANILA FILIPPONI and ROBERTA VARRIALE are with Istituto Nazionale di Statistica (ISTAT), Via Cesare Balbo, 16 – 00184 Rome, Italy.
*Address correspondence to Laura Boeschoten, Tilburg University, Warandelaan 2, 5037 AB, Tilburg, The Netherlands; E-mail: lauraboeschoten@gmail.com.

doi: 10.1093/jssam/smz052

1. INTRODUCTION

Latent class analysis (LCA) is a statistical method to identify a categorical latent variable using categorical observed variables. Latent class analysis can be used to evaluate measurement errors for categorical response variables when different sources that measure the same phenomenon are available (Biemer 2011, p. 13). In this context, the score of the latent variable is assumed to be the *true* value of the target phenomenon, and the available sources are used as contaminated measures of it. Generally, in such a situation, the number of classes of the latent variable are set to be equal to the number of categories in the observed variables. The model output then provides information on how response patterns of the observed variables are related to scores on the latent variable, which can be interpreted as the measurement error of single categories within the observed variables. It also provides information on the distribution of the latent variable given a response pattern on the observed variables, which can be used to predict the value of the scores of the latent variable for each response pattern profile. Using LCA in such a way enables researchers to perform analyses when an error-free data source is not present. This is particularly attractive within the field of official statistics where often sources of information are used in statistical production that are not directly collected for statistical purposes. An effective use of these sources requires the development of statistical methods for the evaluation and correction of measurement errors to guarantee the highest quality possible of statistical products. As measurement error has the potential to bias both frequency distributions of single observed variables and the strength of relationships between multiple variables, measurement error estimation and correction is crucial.

Latent class analysis for measurement error has been applied on different research topics, such as neighborhood of residence (Oberski 2016), home ownership (Boeschoten, Oberski, and De Waal 2017), and serious road injuries (Boeschoten, Waal, and Vermunt 2019). However, it has emerged substantively within the field of employment research (Biemer 2004; Kreuter, Yan, and Tourangeau 2008; Magidson, Vermunt, and Tran 2009; Manzoni, Vermunt, Luijkx, and Muffels 2010).

The Labor Force Survey (Eurostat 2012) (LFS) is the main source used to estimate the employment rate and its changes over time. However, in many European countries, administrative data on employment status are collected on a regular basis (this can vary from a daily to a yearly collection, depending on the country and the type of employment). A lot of research has been done on integrating surveys and administrative data to estimate not only the employment rates but also more detailed characteristics such as employment contract types (Pavlopoulos and Vermunt 2015). In this context, the use of Hidden Markov models (HMMs), which are a special version of latent variable modeling for longitudinal data, is especially gaining traction for estimating classification errors in panel data. In HMM, the longitudinal target variable is a latent process, and the longitudinal response variables are contaminated measures of it. The HMM can be described as a model that has two parts: the structural part describing the distribution of the latent process, assumed to follow a Markov chain with a certain number of states, and the measurement part that describes the distribution of the response variables given the latent process estimating the emission probabilities. Covariates may be present in this model either in the measurement or in the structural part. For each time occasion, it is possible to have univariate longitudinal data that is one response variable or multivariate longitudinal data that is more than one response variable. Hidden Markov models are more adaptable than LCA models because estimation of parameters does not necessarily require multivariate longitudinal data for identifiability. Under the basic assumptions of HMM, (i.e., the Markov property, conditional independence between classification errors, and time-homogeneous error probabilities), the model is identifiable with one response variable and a minimum of three panel waves. The identifiability of more complex models, such as correlated error models, requires the availability of multiple observed variables (Bassi 1997). Like LCA, the longitudinal model provides an estimate of the classification error of single categories within the observed variables over time together with the joint conditional probabilities of the latent status given the observed data, which can be used to predict the scores of the latent variable and its changes over time.

Before these HMMs can be used to produce official statistics by National Statistical Institutes, thorough investigation on limitations and sensitivity to the various assumptions that are made is essential. Research has already been performed on the estimation of employment rates using combined LFS-administrative data (Pavlopoulos and Vermunt 2015), the re-use of obtained parameter estimates (Pankowska, Bakker, Oberski, and Pavlopoulos 2017), the influence of linkage error (Pankowska, Bakker, Oberski, and Pavlopoulos 2019), and the influence of mixed mode survey designs (Pankowska, Pavlopoulos Oberski, and Bakker 2018). Filipponi, Guarnera, and Varriale (2019) propose the use of HMM to estimate employment status in the Italian employment register. The Italian employment register is realized using information on employment status coming from different administrative sources. However, since administrative data are gathered by organizations for their specific aims, units and variable definitions may not align perfectly with those of the official statistics program. In their work, Filipponi et al. (2019) focus on measurement error correction.

With HMM, measurement error correction can be performed by using the posterior probabilities to predict the scores of a variable that is measured with error and thus is latent. Obviously, the uncertainty linked to this process has to be incorporated, especially when performing further statistical analyses. A way to achieve this is by means of multiple imputation (Rubin 1987). A combination of latent class modeling and multiple imputation has been proposed by Boeschoten et al. (2017) and has been denoted as the multiple imputation of latent classes (MILC) method. More specifically, the MILC method utilizes a latent class model to estimate the conditional probabilities of the observed data given the latent status. Next, the posterior membership probabilities obtained from the LCA model are used to create multiple imputations of the construct under investigation. Assigning values to the latent variable can be beneficial for a number of reasons. First, imputations can also be created for individuals having missing values on either one of the observed variables. Second, as imputations are created for the entire population, it becomes straightforward to produce consistent small-area estimates or to create cross-tables with different covariates. Third, because multiple imputation is used, all results can be supplemented with appropriate variance estimates.

The aim of this article is to investigate how the MILC method can be adapted in such a way that scores are assigned to predicted values of a variable that is measured with error for multiple time points. In section 2, the hidden Markov model developed by Filipponi et al. (2019), used as a starting point for this research, is described in more detail. Next, it is described how multiple imputations can be created using this model and the complete procedure, from creating imputations to obtaining estimates. In section 3, the performance of the imputation procedure proposed is investigated by means of a simulation study. Finally, the imputation procedure is applied to three different regions in Italy using data from 2014 (section 4). Section 5 concludes the work.

2. METHODOLOGY

This section describes in detail how multiple imputations of employment can be made using HMM output. First, the HMM used for this investigation is introduced. Second, it is explained how multiple imputations of employment scores can be created using this model.

2.1 Hidden Markov Model Estimating Employment Status in Italy

The HMM described in this section is developed by Filipponi et al. (2019). The model is applied on a person-linked combined data set containing monthly employment status measured by administrative sources and by the LFS. The administrative data contains individual scores for the complete population for every month, while the LFS is administered twice a year per respondent, with three months in between. Despite that the LFS data are only observed on a sample of the population, Filipponi et al. (2019) showed that the classification errors of this survey are lower than those of the administrative data used. Because the aim is to predict the employment status within the Italian employment register, the choice between competing models has been based on the criteria of lower entropy, which should guarantee a lower classification uncertainty at the aggregate and the individual level. The availability of indicators with low classification error helps the entropy reduction. The fitted HMM considers two indicators, the administrative source and the LFS, and twelve time points—one for every month—spanning a time-frame of one year.

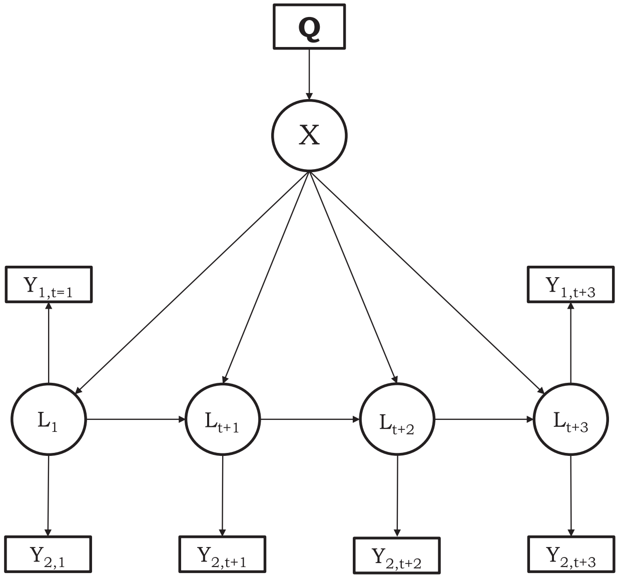

Figure 1. Hidden Markov Model Used to Estimate Employment Status per Month in Italy, as Developed by Filipponi et al. (2019).

Figure 1 shows a graphical overview of the model. Here, let $Y_{1,t}$ denote the response variable originating from the LFS and $Y_{2,\,t}$ denote the response variable originating from the administrative source, where for both response variables $t = 1, \ldots, T$ and $T = 12$. The two vectors with elements $Y_{1,\,1}, \ldots, Y_{1,\,T}$ and $Y_{2,\,1}, \ldots, Y_{2,\,T}$ are respectively denoted by $\mathbf{Y_{(1)}}$ and $\mathbf{Y_{(2)}}$. The numbers of categories of the two response variables are equal, namely two, with (1) unemployed and (2) employed.

The vector $\mathbf{L} = (L_1, \ldots, L_T)$ represents the hidden Markov variable measuring the employment scores over time. Here, the number of latent states is equal to two, with (1) unemployed and (2) employed. A discrete (latent) random effect X is included in the model to account for unobserved heterogeneity, so $\mathbf{L}$ follows a first order Markov chain conditional on X. In particular, the latent variable X identifies three subpopulations of individuals with different trajectories of $\mathbf{L}$ that can be described as never employed ($x = 1$), always employed ($x = 2$), and moving between employment and unemployment ($x = 3$). Because of this mixture component, we can denote the model as a mixture hidden Markov model.

The probability mass function of the latent variable X is affected by a number of time-invariant covariates, denoted by $\mathbf{Q} = (Q_1, Q_2, \ldots, Q_p)$. Multiple administrative sources were linked on person level to construct $Y_{2,t}$, and the covariate Q_1 identifies these different administrative sources, which are strongly related to the different types of employment contracts. The other covariates used in the model are retirement status, education attendance, earnings, age, and gender. The use of the **Q** covariates will help in the identification of the different components of the latent variable X and therefore the different trajectories of **L**. For more information on HMM utilized to predict the Italian employment status, we refer to Filipponi et al. (2019).

For illustration purposes, we consider two covariates, Q_1 and Q_2, with m and n categories respectively. Therefore, corresponding model parameters are specified as,

$$\phi_{x|\mathbf{q}} = \mathrm{P}(X = x|Q_1 = q_1, Q_2 = q_2), \qquad \begin{array}{l} x = 1,2,3, \\ q_1 = 1,\ldots,m, \\ q_2 = 1,\ldots,n. \end{array}$$

The initial probabilities are specified as,

$$\pi_{l_1|x} = \mathrm{P}(L_1 = l_1|X = x), \qquad \begin{array}{l} l_1 = 1,2, \\ x = 1,2,3. \end{array}$$

The latent transition probabilities are specified as,

$$\pi_{l_t|l_{t-1},x} = \mathrm{P}(L_t = l_t|L_{t-1} = l_{t-1}, X = x), \qquad \begin{array}{l} l_t = 1,2, \\ x = 1,2,3, \\ t = 1,\ldots,T, \end{array}$$

and the conditional response probabilities are specified as,

$$\psi_{y_{j,t}|l_t} = \mathrm{P}(Y_{j,t} = y_t|L_t = l_t), \qquad \begin{array}{l} y_t = 1,2, \\ l_t = 1,2, \\ t = 1,\ldots,T, \\ j = 1,2. \end{array}$$

For convenience, all expressions denoting realizations of random variables (e.g., x or l_t) are suppressed, unless for special situations. Note that only the latent variable X depends on Q_1 and Q_2 and that only the initial and latent transition probabilities depend on X. Overall, the distribution of the observed indicators, given the covariates, is

$$\mathrm{P}(\mathbf{Y}_{(1)}\mathbf{Y}_{(2)}|\mathbf{Q}) = \sum_{x=1}^{3}\sum_{l_1=1}^{2}\sum_{l_2=1}^{2}\cdots\sum_{l_T=1}^{2}\phi_{x|\mathbf{q}}\pi_{l_1|x}\prod_{t=2}^{T}\pi_{l_t|l_{t-1},x}\prod_{t=1}^{T}(\psi_{y_{1,t}|l_t})^{\delta_t}\psi_{y_{2,t}|l_t}, \tag{1}$$

where δ_t indicates whether an observation for $Y_{1,t}$ is present at time point t. When fitting the HMM to the person-linked combined data set as described previously, a number of assumptions are made. In defining the probability distribution in (1), different assumptions are made. The first assumptions are made on the structural part of the model. It is assumed that **L** is a homogeneous first order Markov Chain, that is that a person's employment status at time point t given its employment status at t—1, is independent of its employment status on t—2 and the latent transition probabilities do not change over time (Biemer 2011, p. 272). Therefore, the model retains the Markov assumption given the covariates and the latent variable controlling for unobserved heterogeneity. Second, assumptions are made on the measurement model. Here, the classification errors of the indicators are assumed to be locally independent and independent over time. Moreover, it is assumed that the amount of classification error within the indicators does not change over time. Finally, it is assumed that the missing values due to the panel construction are missing completely at random (MCAR, Rubin 1976) and missing values due to attrition are missing at random (MAR, Rubin 1976; Pavlopoulos and Vermunt 2015). Although the administrative data are also potentially incomplete, it is not possible to localize these missing values based on the current data structure because the LFS respondents are sampled from the administrative source and can therefore not be observed here.

The aim of fitting the model to the person-linked data set is to impute the employment status, and this can be carried out by drawing from the posterior probabilities. It is possible to distinguish between conditional imputation, where sequences of latent states are generated from the joint conditional probabilities given the observed data, $\mathrm{P}(L_t|L_{t-1}, \mathbf{Y}_{(\mathbf{1})}, \mathbf{Y}_{(\mathbf{2})}, \mathbf{Q})$, and marginal imputation, where the latent status is generated from the posterior probabilities, $\mathrm{P}(L_t|\mathbf{Y}_{(\mathbf{1})}, \mathbf{Y}_{(\mathbf{2})}, \mathbf{Q})$, for $t = 1, .., 12$. The choice between marginal and conditional imputation depends on the aim of the researcher. If the entire sequence of the latent status is of interest, conditional imputation is more suitable, as an imputation of the latent status at time point t is made conditional on the imputation made for time point $t - 1$. Alternatively, if the imputation of only one time point is of interest, marginal imputation is more straightforward

Since the initial and latent transition probabilities of the latent process depend on latent variable X, the marginal imputation of L_t and conditional imputation of L_t given L_{t-1} are carried out conditionally on X. Therefore, conditional imputation of L_t requires an imputation of X first, which is generated by sampling from the posterior membership probabilities:

$$\mathrm{P}(X = x|\mathbf{Y}_{(\mathbf{1})}, \mathbf{Y}_{(\mathbf{2})}, \mathbf{Q}) = \frac{\mathrm{P}(\mathbf{Y}_{(\mathbf{1})}, \mathbf{Y}_{(\mathbf{2})}|X = x, \mathbf{Q})\mathrm{P}(X = x|\mathbf{Q})}{\sum_{s=1}^{3} \mathrm{P}(\mathbf{Y}_{(\mathbf{1})}, \mathbf{Y}_{(\mathbf{2})}|X = s, \mathbf{Q})\mathrm{P}(X = s|\mathbf{Q})}. \quad (2)$$

The conditional imputation of L_t conditional on the imputation of X (denoted by x^*) is then generated by sampling from the probabilities:

$$P(L_1|X = x^*, \mathbf{Y}_{(1)}, \mathbf{Y}_{(2)}, \mathbf{Q}) = \frac{P(L_1, \mathbf{Y}_{(1)}, \mathbf{Y}_{(2)}, |X = x^*, \mathbf{Q})}{P(\mathbf{Y}_{(1)}, \mathbf{Y}_{(2)}|X = x^*, \mathbf{Q}),} \tag{3}$$

when $t = 1$ and

$$\mathrm{P}(L_t|L_{t-1}, X, \mathbf{Y}_{(1)}, \mathbf{Y}_{(2)}, \mathbf{Q}) = \frac{\mathrm{P}(L_t, L_{t-1} = l^*_{t-1}, \mathbf{Y}_{(1)}, \mathbf{Y}_{(2)}, |X = x^*, \mathbf{Q})}{\mathrm{P}(L_{t-1} = l^*_{t-1}, \mathbf{Y}_{(1)}, \mathbf{Y}_{(2)}|X = x^*, \mathbf{Q}),} \tag{4}$$

when $t > 1$ (which is also conditional on the imputation of L_{t-1} denoted by l^*_{t-1}) and where the distribution defined in (2) to (4) can be obtained by marginalizing

$$\mathrm{P}(\mathbf{Y}_{(1)}, \mathbf{Y}_{(2)}, \mathbf{L}|X, \mathbf{Q}) = \pi_{l_1|x} \prod_{t=2}^{T} \pi_{l_t|l_{t-1},x} \prod_{t=1}^{T} (\psi_{y_{1,t}|l_t})^{\delta_t} \psi_{y_{2,t}|l_t}.$$

Alternatively, the marginal imputation of L_t given X can be generated by sampling from the posterior probabilities regardless of t, x^*, or l^*_{t-1}:

$$\mathrm{P}(L_t|X, \mathbf{Y}_{(1)}, \mathbf{Y}_{(2)}, \mathbf{Q}) = \frac{\mathrm{P}(L_t, \mathbf{Y}_{(1)}, \mathbf{Y}_{(2)}|X, \mathbf{Q})}{\mathrm{P}(\mathbf{Y}_{(1)}, \mathbf{Y}_{(2)}|X, \mathbf{Q})}.$$

It is important to underline that the evaluation of the quantities expressed in (1) to (4) involve sums over a large number of configurations. For example, to compute (1), it is necessary to evaluate a sum over all possible 3×2^T configurations of the vectors $\mathbf{l}$ and X. An efficient way to compute the posterior membership probabilities is the forward recursion algorithm (Baum, Petrie, Soules, and Weiss 1970), which is implemented in Latent GOLD (Vermunt and Magidson 2013), and Latent GOLD version 5.1 is used for the estimation of all HMMs in this article.

2.2 Multiple Imputation Using the Hidden Markov Model

Multiple imputation (MI) is a well-known and attractive method for dealing with missing data problems (Rubin 1987). The basic idea of MI is to construct multiple, say m, data sets, by imputing m times the missing values. This allows to perform the statistical analysis using standard techniques and to obtain the relatively unbiased standard errors. Multiple imputation requires the definition of an imputation model, depending on the measurement level of the variable of interest. Then the imputation should reflect not only the uncertainty of the missing values but also the uncertainty of parameters of the defined imputation model. The parameter uncertainty is guaranteed by a full Bayesian approach, where the random imputations are based on random draws of the parameters (Schafer 1997). Boeschoten et al. (2017) developed the MILC procedure, that is, an MI approach when a latent class model is used to impute a latent construct under evaluation. In this context, the parameter uncertainty is dealt with within a frequentist framework by using a nonparametric bootstrap (King, Honaker, Joseph, and Scheve 2001). In this section, we describe the extension of the MILC procedure to HMM.

The MILC procedure comprises five steps. In the first step, m nonparametric bootstrap samples are generated from the original data set containing the indicators and covariates used to estimate the HMM. A bootstrap sample is obtained by sampling from the observed frequency distribution. In the second step, the HMM described in section 2.1 is fitted on each of the m bootstrap samples. Then in the third step, one imputation for $\mathbf{L}$ is created using the m^{th} HMM obtained using the m^{th} bootstrap sample, resulting in m imputations of the sequence of $\mathbf{L}$. These imputations can be created using either the conditional imputation procedure or the marginal imputation procedure, as described in section 2.1. In the fourth step, estimates of interest can then be obtained from every imputation, and in the fifth step, the estimates obtained for every imputation can be pooled using the pooling rules defined by Rubin (1987, p. 76). For an example of how to apply the pooling rules in the MILC context to obtain pooled estimates of frequency tables, we refer to Boeschoten et al. (2017). It is important to note that drawing bootstrap samples in step one allows us to indirectly take into account parameter uncertainty through the imputations created at a later step.

3. SIMULATION STUDY

3.1 Set-up of the Simulation Study

The performance of the multiple imputation procedure using HMM to obtain employment status is empirically evaluated by a simulation study.

The model used to draw samples has been specified to mimic the combined Italian LFS-administrative data set. Figure 2 graphically represents the model specified. In the simulation study, there are only three time points observed for computation-time reasons.

The mixture variable X incorporates individual heterogeneity. In particular, the groups of people identified by X can be interpreted as people who are "never employed" ($x = 1$), "always employed" ($x = 2$) or "move between being employed and unemployed" ($x = 3$). Two covariates are included in the model: Q_1 with the purpose to mimic the influence of the "source" covariate (described in section 2.1) and therefore has a strong relationship with X; Q_2 which has no relationship with the mixture at all. Table 1 shows the parameters ϕ_x and $\phi_{x|\mathbf{q}}$.

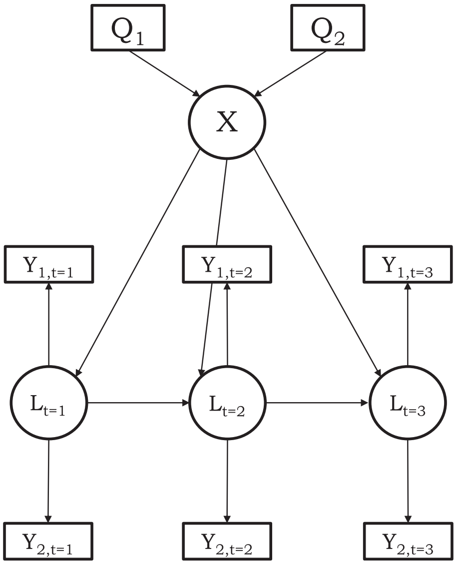

Figure 2. HMM Used in the Simulation Study to Evaluate the Performance of the HMM Multiple Imputation Procedure.

Table 1. Parameters of Latent Variable *X*

ϕ_{x}	x = 1			x = 2			x = 3		
	0.5800			0.3025			0.1175		
$\phi_{x\mid\mathbf{q}}$	$q_2 = 1$	$q_2 = 2$	$q_2 = 3$	$q_2 = 1$	$q_2 = 2$	$q_2 = 3$	$q_2 = 1$	$q_2 = 2$	$q_2 = 3$
$q_1 = 1$	0.95	0.025	0.025	0.95	0.025	0.025	0.95	0.025	0.025
$q_1 = 2$	0.025	0.95	0.025	0.025	0.95	0.025	0.025	0.95	0.025
$q_1 = 3$	0.025	0.025	0.95	0.025	0.025	0.95	0.025	0.025	0.95

The covariate Q_2 is included because it will be used later on for developing a MAR missingness mechanism. Furthermore, it would be interesting to compare the performance of the imputation procedure for estimating the relationship with a strongly related covariate to a weakly related covariate.

The set of three initial probabilities and latent transition probabilities of the Markov chain for each mixture groups are reported in tables 2 and 3.

Here it can be seen that the group "never employed" ($x = 1$) has a strong relationship with $L_1 = 1$. In contrast, $x = 2$ has a strong relationship with $L_1 = 2$, the "employed" group in the Markov chain. As the $x = 3$ group shifts between

Table 2. Parameters of the Hidden Markov Variable L: Initial Probabilities

$\pi_{l_1\|x}$	$l_1 = 1$	$l_1 = 2$
$x = 1$	0.97	0.03
$x = 2$	0.06	0.94
$x = 3$	0.50	0.50

Table 3. Parameters of the Hidden Markov Variable L: Latent Transition Probabilities

$\pi_{l_t\|l_{t-1},x}$	$x = 1$		$x = 2$		$x = 3$	
	$l_t = 1$	$l_t = 2$	$l_t = 1$	$l_t = 2$	$l_t = 1$	$l_t = 2$
$l_{t-=} = 1$	0.97	0.03	0.94	0.06	0.70	0.30
$l_{t-.3} = 2$	0.03	0.97	0.06	0.94	0.30	0.70

Table 4. Parameters of the Measurement Model

$\psi_{y_j\|l}$	Condition 1		Condition 2	
	$l_t = 1$	$l_t = 2$	$l_t = 1$	$l_t = 2$
$y = 1$	0.95	0.05	0.80	0.20
$y = 2$	0.05	0.95	0.20	0.80

being employed and unemployed, their probability of being unemployed or employed on $t = 1$ is 0.5. It can also be seen that for $x = 3$, the probability of changing from being unemployed to employed or the other way around is larger compared with the other groups. Finally, the error probabilities of the two indicator variables are specified in table 4.

These probabilities indicate that for both indicator variables, 95 percent is correctly classified, which can be considered realistic for the LFS indicator but low for the administrative data. To investigate the performance of the procedure with data of a lower quality, indicators are also simulated with 80 percent correctly classified.

In theory, it is possible to evaluate the performance of the imputation procedures by investigating all these parameters. This would, however, result in a large amount of information, which is not all relevant. As previously described,

Table 5. Parameters of Latent Variable $\pi_{\bar{l}}$

	$\bar{l}=1$			$\bar{l}=2$		
$\pi_{\bar{l}}$	0.6388			0.3612		
	$q_2=1$		$q_2=2$		$q_2=3$	
$\pi_{\bar{l}\|\mathbf{q}}$	$\bar{l}=1$	$\bar{l}=2$	$\bar{l}=1$	$\bar{l}=2$	$\bar{l}=1$	$\bar{l}=2$
$q_1=1$	0.9105	0.0895	0.9105	0.0895	0.9105	0.0895
$q_1=2$	0.1412	0.8588	0.1412	0.8588	0.1412	0.8588
$q_1=3$	0.5013	0.4987	0.5013	0.4987	0.5013	0.4987

the aim of this article is to investigate whether the MI is an appropriate method to evaluate the variability when a HMM is used to impute a latent construct. Therefore, it makes sense to evaluate the performance of $\pi_{\bar{l}|\mathbf{q}}$, $\pi_{\bar{l}}$ and ϕ_x. Here, ϕ_x can be found in table 1, while $\pi_{\bar{l}}$ is obtained by marginalizing

$$\pi_{\bar{l}} = \frac{\sum_{q_1=1}^{m}\sum_{q_2=1}^{n}\sum_{x=1}^{3}\sum_{t=1}^{T}\phi_{x|\mathbf{q}}\pi_{l_1|x}\pi_{l_t|t-1,x}N_{q_1q_2}}{\sum_{q_1=1}^{m}\sum_{q_2=1}^{n}N_{q_1q_2}T}$$

and

$$\pi_{\bar{l}|\mathbf{q}} = \frac{\sum_{x=1}^{3}\sum_{t=1}^{T}\phi_{x|\mathbf{q}}\pi_{l_1|x}\pi_{l_t|t-1,x}N}{NT}.$$

The corresponding obtained population values can be found in table 5.

As previously described, two alternative approaches for creating the imputations can be considered, conditional and marginal imputation, and they are both evaluated in this simulation study.

As a reference, the parameters described in section 3.1.2 are obtained from the HMM output directly. The performance of conditional and marginal imputation from the posterior distribution is investigated by means of single and multiple imputation (five and ten imputations). Furthermore, to investigate the extent of parameter uncertainty in the situation under evaluation, MILC is applied both with and without bootstrap for parameter uncertainty. Summarizing, we compare eleven conditions per simulation study:

- HMM: parameters from the HMM directly
- SI: generate a single conditional and marginal imputation.

- MI-5/10-C/M: generate five or ten conditional or marginal imputation.
- MI-B-5/10-C/M: generate five or ten conditional or marginal imputations using five or ten HMMs estimated on five or ten bootstrap samples from the observed data

The parameters under evaluation are investigated using four performance measures. Note that we only provide the following equations for ϕ_x, though they can also be applied to the other parameters under evaluation: $\pi_{\bar{l}}$ and $\pi_{\bar{l}|\mathbf{q}}$. First, the bias of the parameters is investigated, which is equal to the difference between the average estimate over all replications and the value found in the theoretical population:

$$\text{bias}_{\phi_x} = \frac{\sum_{j=1}^{N_{\text{it}}} (\phi_x - \widehat{\phi_x})}{N_{\text{it}}},$$

where N_{it}, stands for the number of simulation replications performed in the simulation study, which is in this case always five hundred. Second, the coverage of the 95 percent confidence interval is under investigation; third, the ratio of the average standard error of the estimate over the standard deviation of the five hundred replication estimates is estimated:

$$\frac{\left[\frac{\sum_{j=1}^{N_{\text{it}}}}{N_{\text{it}}} \text{SE}(\widehat{\phi}_x)\right]}{\text{SD}(\widehat{\phi}_x),}$$

where standard error (SE) is the square root of the estimate of the total variance obtained after applying the pooling rules by Rubin (1976) and

$$\text{SD}(\widehat{\phi}_x) = \sqrt{\frac{\sum_{j=1}^{N_{\text{it}}} (\widehat{\phi_x} - \bar{\phi}_x)^2}{N_{\text{it}}}},$$

which is estimated to confirm that the standard errors of the estimates are properly estimated. Finally, the root mean squared error is estimated:

$$\text{RMSE}_{\phi_x} = \sqrt{\frac{\sum_{j=1}^{N_{\text{it}}} (\widehat{\phi_x} - \phi_x)^2}{N_{\text{it.}}}}$$

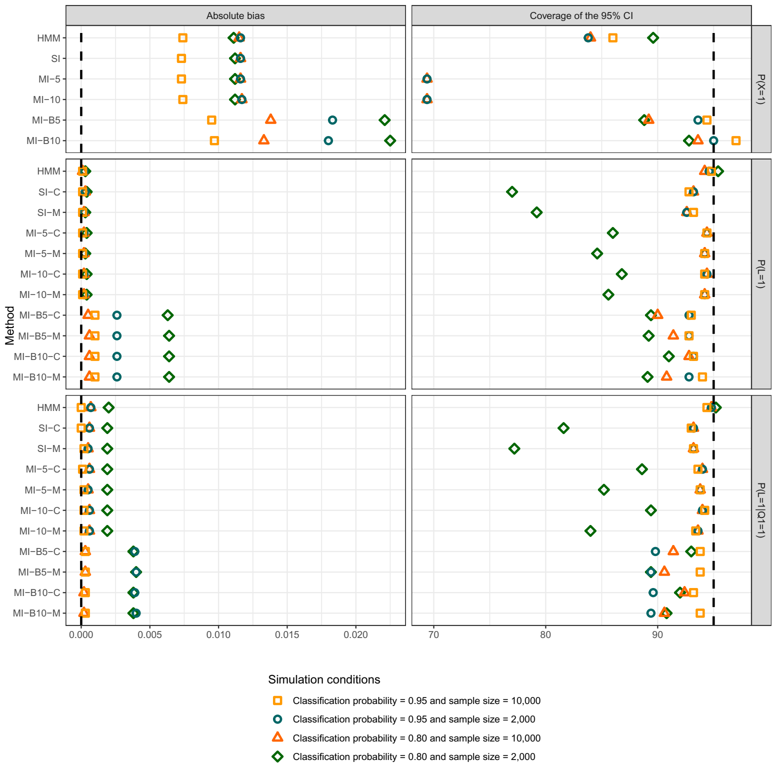

Figure 3. Plot of Results in Terms of Bias and Coverage of the 95 Percent Confidence Interval in the Columns. Note that we removed the following values from the results in terms of coverage: For SI: 42.6, 66.0, 66.0, 56.4 (in order of the conditions as listed in the legend). For MI-5: 45.6 (for $\psi_{1|1}$ =0.80 and ss = 2000) and 61.2 45.0 (for $\psi_{1|1}$ =0.95 and ss = 10.000). For MI-10: 45.6 (for $\psi_{1|1}$ =0.80 and ss = 2000) and 60.4 (for $\psi_{1|1}$ =0.95 and ss = 10.000).

3.2 Simulation Results

In this section, the results of the simulation study are presented. In figures 3 and 4, the different rows of graphs represent three parameters investigated, the different rows within each graph represent different approaches to the longitudinal extension of the HMM, and the four different combinations of color and shape represent the four different simulation conditions concerning classification probability and sample size.

In this section, only the simulation results graphically represented in figures 3 and 4 are discussed. Note that all parameters ϕ_x, $\pi_{\bar{l}}$, and $\pi_{\bar{l}|q_1}$ behave in a similar way, so therefore, only one parameter of each variable is graphically represented and discussed here.

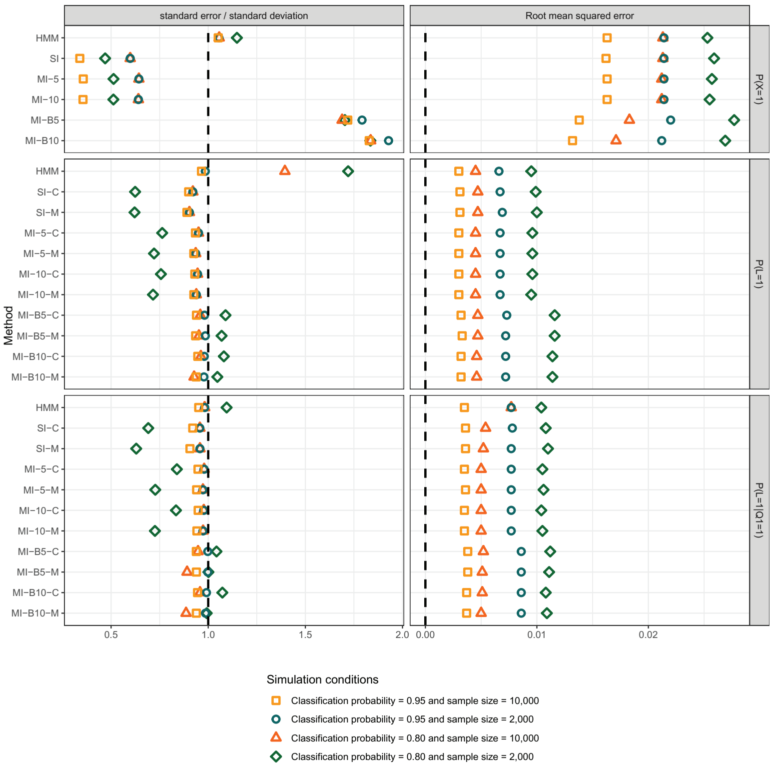

Figure 4. Plot of Results in Terms of SE/SD and RMSE.

When evaluating the results obtained directly from the HMM output, it can be seen that for $\pi_{\bar{l}=1}$ and $\pi_{\bar{l}=1|q_1=1}$ in figures 3 and 4, the model is able to produce estimates almost without bias and with nominal coverage rates equal to the 95 percent confidence interval for all simulated conditions. The confidence interval width becomes wider as the simulation condition becomes more "challenging" (i.e., smaller sample size and/or lower classification probability), and also the average standard error becomes larger in relation to the standard deviation over the estimates in such cases. In contrast, the HMM has more difficulties with estimating the parameters of $\phi_{x=1}$. Here, a small amount of bias and undercoverage can be detected, and these amounts are related to the "difficulty" of the simulation condition.

When comparing the results after a single imputation with the results after multiple imputation, it can be seen that in terms of bias of $\pi_{\bar{l}}$ and $\pi_{\bar{l}|q_1}$, there are no problems for single or multiple imputation. They also seem to perform well on other evaluation criteria for most simulation conditions. Coverage rates however are too low for both single and multiple imputation but are worse for single imputation, and a similar pattern can be seen when evaluating the average standard error divided by the standard deviation over the estimates. For ϕ_x, both single and multiple imputation are not performing well. It is especially outstanding that the "small sample size, large classification probability" and "large sample size, low classification probability" conditions behave very similarly.

In terms of bias, an increase can be observed when the bootstrap is applied in comparison with when it is not applied. However, for the most difficult simulation condition (the condition with undercoverage for $\pi_{\bar{l}}$ and $\pi_{\bar{l}|q_1}$), the results improve when the bootstrap is applied. Another consequence here is a wider confidence interval and a SE/SD that is larger than the nominal value of one. The RMSE also indicates that the bootstrap results in a loss of efficiency. For ϕ_x, the bias and confidence interval width increase even more in comparison with $\pi_{\bar{l}}$ and $\pi_{\bar{l}|q_1}$. However, this also results in coverage rates developing from unacceptable to almost nominal.

Finally, results obtained after creating different numbers of imputations were also investigated. However, the differences between creating five or ten imputations are not noteworthy. Almost no differences can be detected between the two different imputation procedures. The only notable difference is found in the most difficult condition, which had undercoverage for both $\pi_{\bar{l}}$ and $\pi_{\bar{l}|q_1}$. The results obtained after conditional imputation lead to a coverage rate closer to the nominal 95 percent level compared with the results obtained after marginal imputation.

3.3 Missing Values

The aim of this article is to investigate how the MILC method can be extended to a longitudinal context. Because the HMM developed by Filipponi et al. (2019) has been used as a starting point for our research, and LFS only contains a sample of the population (and this subset also contains missing values), it makes sense to investigate if the missing values (both by nonresponse and by design) influence the quality of the estimates obtained when the longitudinal extension of the MILC method is applied.

Because the simulation condition with classification probabilities of 0.95 and sample size of 10,000 is closest to the situation where this model is applied in practice, only this condition will be used to further investigate the influence of missing values. In practice, only a very small subset of the population has observations on the LFS indicator variable measuring employment (see section 4 for some exemplary numbers). Although this situation would ideally be replicated, this would not be feasible for a simulation study for computational reasons. We decided to set the percentage of missing cases for the indicator representing the LFS to 50 percent.

Both a missing completely at random (MCAR) mechanism and a missing at random (MAR) mechanism are investigated. With the MCAR mechanism, the

Table 6. Illustration of the Simulated Data Structure Containing Missing Values

case ID	$Y_{1,\,t=1}$	$Y_{1,\,t=2}$	$Y_{1,\,t=3}$	$Y_{2,\,t=1}$	$Y_{2,\,t=2}$	$Y_{2,\,t=3}$	Q_1	Q_2
1	1	1	1	1	1	NA	1	1
2	1	1	1	1	1	NA	1	1
3	1	1	1	NA	NA	NA	1	1
4	1	1	1	NA	NA	NA	1	1

probability of being missing is equal for all respondents, namely 0.50. With the MAR mechanism, the probability of being missing is not equal for all respondents. Instead, the probability of being missing is related to a respondent's score on another variable. In this case, it is related to the score of the respondent on covariate Q_2. If

$$Q_2 = 1, \mathrm{P}(Y_2 = \mathrm{NA}) = 0.25;$$
$$Q_2 = 2, \mathrm{P}(Y_2 = \mathrm{NA}) = 0.50;$$
$$Q_2 = 3, \mathrm{P}(Y_2 = \mathrm{NA}) = 0.75,$$

where the total number of missing cases also depends on the frequency distribution of Q_2. As the aim is to mimic the structure of the longitudinal combined LFS-administrative data set, the missingness mechanism is used to generate 50 percent missing values on all time points of Y_2. Furthermore, since the LFS is only observed for two time points, the third time point of Y_2 is made missing for all observations. See table 6 for an illustration of this data structure.

In figures 5 and 6, an overview of the results obtained after applying the longitudinal extension of MILC on a data set with MCAR or MAR missingness are compared with results obtained after applying the longitudinal extension of MILC on a fully observed data set. In the figures, the rows of graphs represent three parameters investigated, and the rows within each graph represent different approaches to the longitudinal extension of the HMM. The three combinations of color and shape represent MCAR, MAR, and fully observed simulation conditions.

As in the simulation study conducted in sections 3.1 and 3.2, it can be seen that the results for ϕ_x are more problematic compared with the other results. In general, the results for MCAR or MAR are very similar to those obtained when a fully observed data set is used. Only in terms of bias and in terms of RMSE can some differences be found. Here, it can be seen that the bias and RMSE increase slightly when we shift from fully observed to MCAR and increase more when shifting from MCAR to MAR. In terms of $\pi_{\bar{l}}$ and $\pi_{\bar{l}|q_1}$, the results with MAR and MCAR are even more similar to the fully observed results compared with those obtained for ϕ_x.

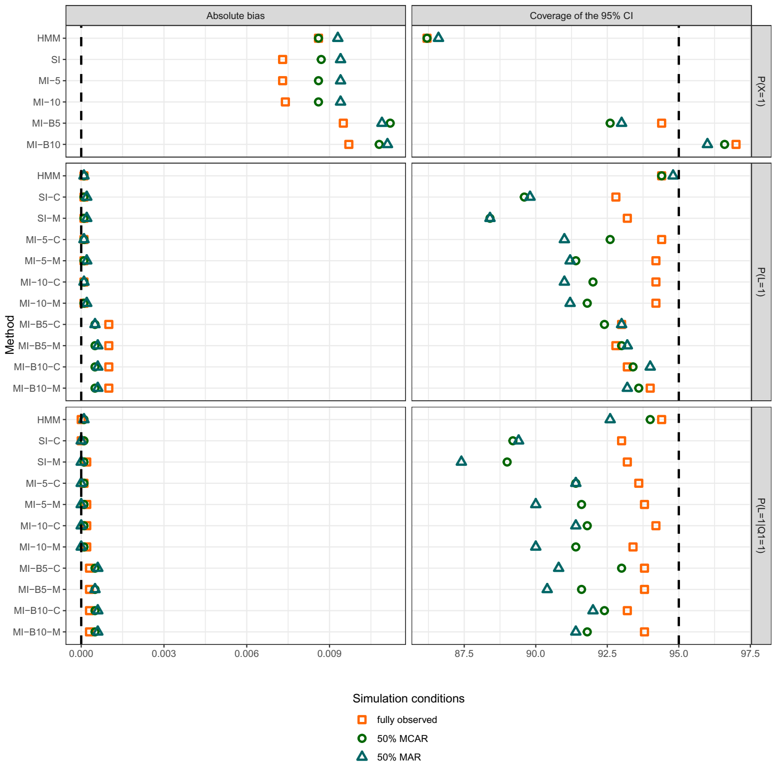

Figure 5. Plot of Results in Terms of Bias and Coverage of the 95 Percent Confidence Interval in the Columns. Note that we removed the following values from the results in terms of coverage: For SI: 42.6, 43.4, 43.2. For MI-5: 45.6, 45.8, 45.8. For MI-10: 45.0, 45.2, 45.8 (in order of the conditions as listed in the legend).

4. APPLICATION

The longitudinal extension of the MILC method is applied separately to data from 2014 from three different regions in Italy: Veneto, Umbria, and Basilicata. As can be seen in figure 4, the regions are spread out over the country, from north to south. Also, the regions differ substantively in the number of inhabitants in the workforce and number of LFS respondents: Veneto has 4,821,983 inhabitants in the workforce and 17,246 LFS respondents; Umbria has 899,366 inhabitants in the workforce and 8,477 LFS respondents; Basilicata has 579,860 inhabitants in the workforce and 10,202 LFS respondents.

In these data sets, the longitudinal extension of the MILC method is applied using the HMM described in figure 1 and section 2.1. As described in section

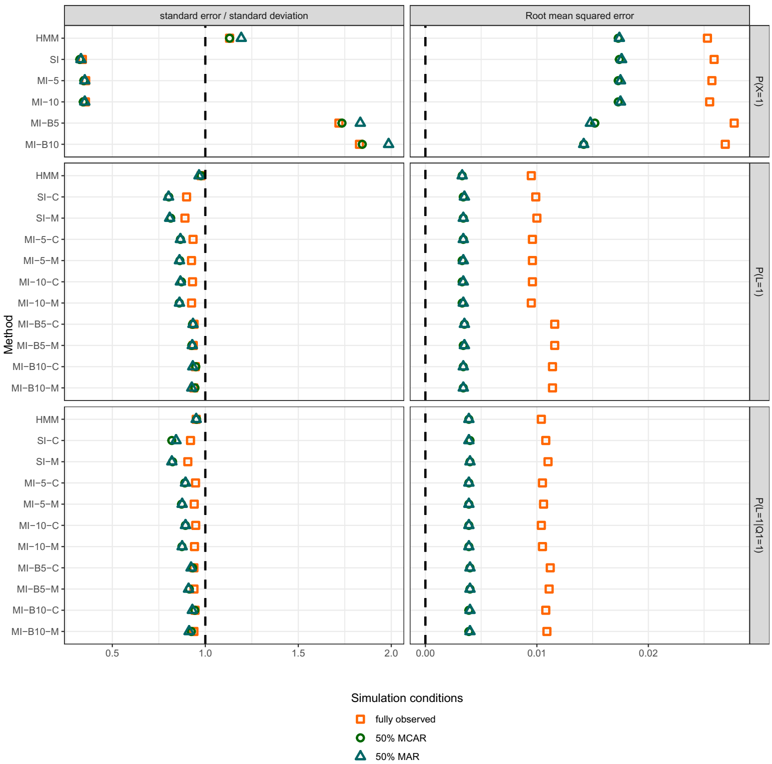

Figure 6. Plot of Results in Terms of SE/SD and RMSE.

2.1, a covariate (Q_1) is included that specifies from which administrative source the $Y_{2,t}$ indicator score originates. Q_1 has four categories: (1) no source, (2) employees, (3) self-employers with time information, (4) self-employers without time information. One other covariate is included in the model: gender (Q_2) with the categories (1) male and (2) female.

Based on the results from the simulation study conducted in sections 3.2 and 3.3, the longitudinal extension of the MILC method is applied using five bootstrap samples. As we concluded from the simulation study that marginal and conditional imputation produce similar results, only the results for conditional imputation are shown.

In table 7, the results in terms of proportions and corresponding standard errors for the different regions are found. By using proportions, it is possible to directly compare the employment rates over the different regions. First, it can be observed that the employment rate ($\pi_{\bar{l}=2}$) decreases as we shift to a more southern region. Similarly, we observe that the proportion of the mixture group representing the "employed trajectory" ($\phi_{x=1}$) also decreases when shifting to a more southern region.

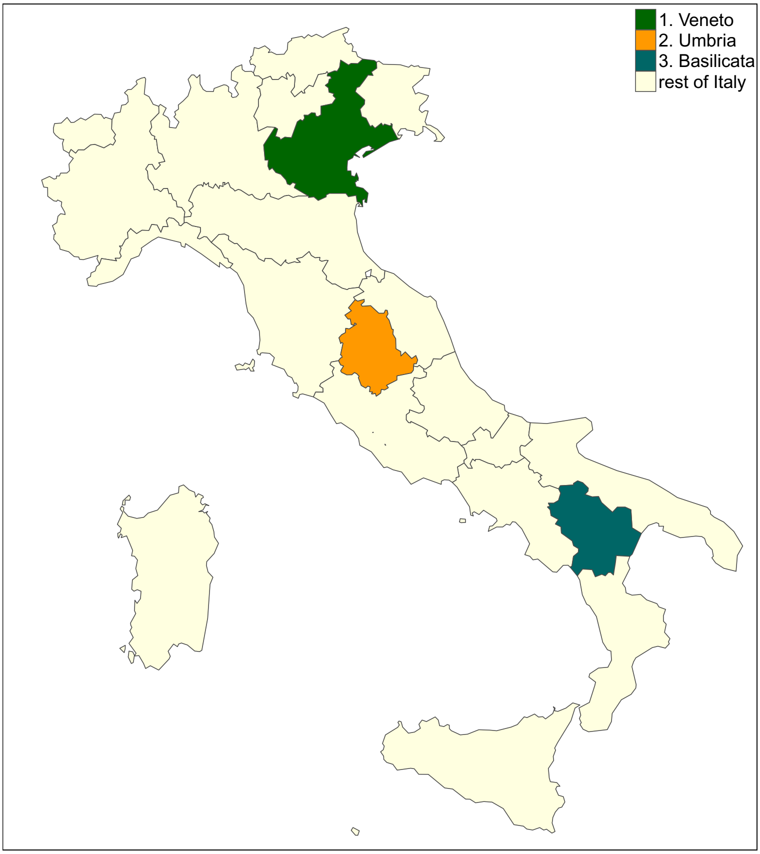

Figure 7. Map of Italy with the Three Regions Highlighted on which the Longitudinal Extension of the MILC Method are Applied. © EuroGeographics for the administrative boundaries.

When we investigate the proportion of being employed conditional on the administrative source where a person's information was obtained, it can be observed that source one does not contain any employed persons ($\pi_{\bar{l}=2|q_1=1}$), while source four does not contain any unemployed persons ($\pi_{\bar{l}=1|q_1=4}$). Furthermore, the proportion of being employed is particularly larger for the Basilicata region compared with the other regions in source two ($\pi_{\bar{l}=1|q_1=2}$).

When investigating the proportion of being employed conditional on gender, differences between north and south are also visible. For example, the proportion of being unemployed conditional on being male ($\pi_{\bar{l}=1|q_2=1}$) shifts from approximately 0.50 in Veneto to approximately 0.60 in Basilicata, while the proportion of being unemployed conditional on being female ($\pi_{\bar{l}=1|q_2=2}$) also increases with approximately 10 percent if we shift from Veneto to Basilicata,

Table 7. Results in Terms of Proportions Obtained after Applying the Longitudinal Extension of the MILC Method to Data from Three Different Regions in Italy.

	Veneto		Umbria		Basilicata	
	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
$\phi_{x=1}$	0.37423	0.00009	0.33869	0.00036	0.26793	0.00028
$\phi_{x=2}$	0.53421	0.00007	0.57020	0.00017	0.61503	0.00022
$\phi_{x=3}$	0.09156	0.00008	0.09111	0.00039	0.11704	0.00027
$\pi_{\bar{l}=1}$	0.57994	0.00007	0.61601	0.00015	0.67801	0.00018
$\pi_{\bar{l}=2}$	0.42006	0.00007	0.38399	0.00015	0.32199	0.00018
$\pi_{\bar{l}=1\|q_1=1}$	1.00000	0.00000	1.00000	0.00000	1.00000	0.00000
$\pi_{\bar{l}=2\|q_1=1}$	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
$\pi_{\bar{l}=1\|q_1=2}$	0.23606	0.00008	0.24065	0.00021	0.42219	0.00033
$\pi_{\bar{l}=2\|q_1=2}$	0.76393	0.00008	0.75934	0.00021	0.57781	0.00033
$\pi_{\bar{l}=1\|q_1=3}$	0.10210	0.00003	0.11137	0.00012	0.17714	0.00017
$\pi_{\bar{l}=2\|q_1=3}$	0.89790	0.00003	0.88863	0.00012	0.82286	0.00017
$\pi_{\bar{l}=1\|q_1=4}$	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
$\pi_{\bar{l}=2\|q_1=4}$	1.00000	0.00000	1.00000	0.00000	1.00000	0.00000
$\pi_{\bar{l}=1\|q_2=1}$	0.50314	0.00003	0.55092	0.00008	0.59548	0.00008
$\pi_{\bar{l}=2\|q_2=1}$	0.49686	0.00003	0.44908	0.00008	0.40452	0.00008
$\pi_{\bar{l}=1\|q_2=2}$	0.65305	0.00003	0.67603	0.00006	0.75725	0.00007
$\pi_{\bar{l}=2\|q_2=2}$	0.34695	0.00003	0.32397	0.00006	0.24275	0.00007

NOTE.—The columns represent the estimates and standard errors of the different regions, and the rows represent the different parameters, which are the same as investigated in the simulation studies.

from approximately 0.65 to approximately 0.75. Therefore, although the probability of being employed is larger for males compared with females, the strength of this relationship does not change over the different regions.

5. DISCUSSION

The MILC is a method that relies on multiple imputation of latent classes. In previous literature, MILC was applied to data sets with a cross-sectional setup; in this article, MILC has been extended to longitudinal data by using hidden Markov models. In the recent literature, hidden Markov Models have been increasingly used in the field of research on employment. In this context, HMMs are applied on unit-linked combined data sets where information comes from different data sources, potentially affected by classification errors. This article presented a method to predict the scores of the latent categorical variable in order to correct for measurement error in all data sources and taking into account the uncertainty of the imputations. This method is an extension of the multiple imputation of latent classes method that has been proposed by Boeschoten et al. (2017) in the context of LC analysis. This article has shown that imputations for different time points can be generated in multiple ways and thereby illustrates the flexibility of the MILC method. In particular, a simulation study highlighted the usability of the MILC method in different conditions. A limitation of the current simulation study is that classification error rates larger than 20 percent were not investigated.

A simplification of the HMM developed by Filipponi et al. (2019) was used in a simulation study where results of multiple alternative strategies that could have been chosen were evaluated. In the first simulation study, these strategies were compared using data of different sample sizes and of different quality. The main conclusion of this simulation study was that the results related to the Markov chain measuring employment status were of a different quality compared with the results related to the mixture measuring different trajectories of employment status over time. From this, we can conclude that if a researcher is interested in evaluating the mixture, it is necessary that the bootstrap is applied to incorporate parameter uncertainty into the estimate of the total variance. When the data is of sufficient quality (which was apparently already the case with 20 percent classification error), parameter uncertainty for the Markov model is at such a low rate that it can be ignored. Note that when multiple imputation is applied in this case without incorporating the bootstrap for parameter uncertainty, reliable results related to the mixture cannot be obtained.

The small differences between the five or ten imputations indicated that a low number such as five was probably enough. Furthermore, two different imputation procedures were evaluated (conditional and marginal), and the results showed minor advantages for conditional imputation. Marginal imputation was, however, much more straightforward to apply since an imputation for every time point could be created unconditionally from the imputation of the other time points. This is something to be taken into consideration when creating imputations of the HMM.

Furthermore, MCAR and MAR missingness mechanisms were investigated and did not show substantive reductions in the quality of the output, so the longitidinal extension of the MILC method should be able to handle the missingness structure that is present in the combined LFS administrative data, where a combination of MAR and MCAR is assumed.

For illustration purposes, the longitudinal extension of the MILC method was applied to three different regions of Italy. Most of the assumptions made when applying multiple imputation of the HMM were related to the HMM itself. This model has been thoroughly investigated by many researchers, both inside and outside the field of official statistics and using data from multiple countries. However, the conclusions described in this article might not hold for when substantive changes within the investigated labor markets happen. For example, the introduction of a basic income could influence the sizes of the mixture groups and possibly also the number and type of mixture groups.

When applying multiple imputation of latent classes, it has been assumed that the covariates are free of classification error. Although this assumption is probably never met in practice, it is not a problem when a simple latent class model is used, as long as the measurement error is random. However, more investigation should be done to see if this holds when the HMM is used, especially in the way the model is currently specified. Here, the mixture groups are determined by the covariates, and classification error in a covariate might result in assignment to the wrong mixture group. In addition, throughout this article the assumption of local independence is made, and a model is used that implies that measurement error in both the survey and administrative data is random. However, in practice it can occur that the measurement error is autocorrelated. In surveys, this is primarily due to individual characteristics and personal responding style; but in administrative data, it might result from the fact that once an error is made, it is likely to be copied on to the following time point. Research is needed to evaluate the performance of combining HMM and multiple imputation when these assumptions are violated.

It should also be noted that the model used for the simulation study was in some ways a simplification of the model developed by Filipponi et al. (2019). The first simplification related to the number of time points. The main reason for reducing this number was because a large number of time points results in a large number of possible profiles, and these can result in parameters not being estimated when applying the bootstrap for parameter uncertainty. It would be interesting to investigate whether HMM with a larger number of time points can be investigated using alternative ways for estimating parameter uncertainty, such as a nonparametric bootstrap or the use of a Gibbs sampler. A second simplification relates to a number of edit restrictions that were specified in the model developed by Filipponi et al. (2019). Although previous research has shown that it is possible to incorporate edit restrictions in multiple imputation of latent classes (Boeschoten et al. 2017), the decision was made to leave them out of this research because the number of different research settings was already very extensive, and the edit restrictions in this setting were very specific for the Italian situation.

In summary, the longitudinal extension of the MILC method presented in this article allows for imputation of scores on multiple time points for longitudinal data structures together with their variability in cases where there is a first order Markov chain, the latent transition probabilities do not change over time, the classification errors are locally independent and independent over time, and missing values follow a MCAR or MAR missingness mechanism. Further research can be performed as to how the methodology is sensitive to violations of these assumptions or how the methodology can be adapted to handle the violations of these assumptions.

REFERENCES

Bassi, F. (1997), "Identification of Latent Class Markov Models with Multiple Indicators and Correlated Measurement Errors," *Journal of the Italian Statistical Society*, 6, 201.

Baum, L. E., T. Petrie, G. Soules, and N. Weiss (1970), "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains," *The Annals of Mathematical Statistics*, 41, 164–171.

Biemer, P. P. (2004), "An Analysis of Classification Error for the Revised Current Population Survey Employment Questions," *Survey Methodology*, 30, 127–140.

———. (2011), *Latent Class Analysis of Survey Error* (Vol. 571), Hoboken, NJ: John Wiley & Sons.

Boeschoten, L., D. Oberski, and T. De Waal (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.

Boeschoten, L., T. D. Waal, and J.K. Vermunt (2019), "Estimating the Number of Serious Road Injuries per Vehicle Type in the Netherlands Using Multiple Imputation of Latent Classes." *Journal of the Royal Statistical Society. Series A*, 182, 1463–1486.

Eurostat (2012), *Labour Force Survey in the EU, Candidate and EFTA Countries: Main Characteristics of the National Surveys 2005*. Eurostat statistical working papers. https://ec.europa.eu/eurostat/documents/3888793/5856085/KS-TC-13-003-EN.PDF/15feeb32-717d-4326-8e7b-da6e2530ed4c.

Filipponi, D., U. Guarnera, and R. Varriale (2019), "Hidden Markov Models to Estimate Italian Employment Status," paper presented at a conference of NTTS 2019 Brussels March 11–13, 2019.

King, G., J. Honaker, A. Joseph, and K. Scheve (2001), "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation," *American Political Science Review*, 95, 49–69.

Kreuter, F., T. Yan, and R. Tourangeau (2008), "Good Item or Bad - Can Latent Class Analysis Tell? The Utility of Latent Class Analysis for the Evaluation of Survey Questions," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171, 723–738.

Magidson, J., J. K. Vermunt, and B. Tran (2009), "Using a Mixture Latent Markov Model to Analyze Longitudinal us Employment Data Involving Measurement Error," in *New Trends in Psychometrics*, eds. K. Shigemasu, A. Okada, T. Imaizumi, and T. Hoshino, pp. 235–242, Universal Academy Press, Inc. 235–242.

Manzoni, A., J. K. Vermunt, R. Luijkx, and R. Muffels (2010), "Memory Bias in Retrospectively Collected Employment Careers: A Model-Based Approach to Correct for Measurement Error," *Sociological Methodology*, 40, 39–73.

Oberski, D. L. (2016), "Estimating Error Rates in an Administrative Register and Survey Questions Using a Latent Class Model," in *Total survey error in practice: improving quality in the era of big data*, eds. P. P. Biemer, E. D. D. Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, C. Tucker, and B. T. West, New York: Wiley.

Pankowska, P., B. Bakker, D. Oberski, and D. Pavlopoulos (2019), "How Linkage Error Affects Hidden Markov Model Estimates: A Sensitivity Analysis," *Journal of Survey Statistics and Methodology*, smz011, https://doi.org/10.1093/jssam/smz011.

Pankowska, P., B. Bakker, D. L. Oberski, and D. Pavlopoulos (2017), "Reconciliation of Inconsistent Data Sources by Correction for Measurement Error: The Feasibility of Parameter Re-Use," *Statistical Journal of the IAOS(Preprint)*, 1–13.

Pankowska, P., D. Pavlopoulos, D. Oberski, and B. Bakker (2018), "Dependent Interviewing: A Remedy or a Curse for Measurement Error in Surveys?" International Total Survey Error Workshop. Available at https://dism.ssri.duke.edu/sites/dism.ssri.duke.edu/files/pdfs/itsew_program_final.pdf.

Pavlopoulos, D., and J. Vermunt (2015), "Measuring Temporary Employment. Do Survey or Register Tell the Truth?," *Survey Methodology*, 41, 197–214.

Rubin, D. B. (1976), "Inference and Missing Data," *Biometrika*, 63, 581–592.

———. (1987), *Multiple Imputation for Nonresponse in Surveys* (Vol. 81), John Wiley & Sons. ISBN 9780471087052, doi:10.1002/9780470316696.

Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data*, Chapman and Hall/CRC.

Vermunt, J. K., and J. Magidson (2013), *Technical Guide for Latent GOLD 5.0: Basic, Advanced, and Syntax*, Belmont, MA: Statistical Innovations Inc.