How linkage error affects Hidden Markov Models[1]

Paulina Pankowska[2], Bart Bakker[2,3], Daniel Oberski[4] & Dimitris Pavlopoulos[2]

Abstract: *Latent class modeling (LCM) is increasingly used to estimate and correct for classification error in categorical data, without the need for a "gold standard", error-free, data source. To accomplish this, LCMs require multiple indicators of the same phenomenon within one data collection wave ("latent structure model"), or multiple observations over time on a single indicator ("hidden Markov model"), and assume that the errors in these indicators are conditionally independent. Unfortunately, this "local independence" assumption is often unrealistic, untestable, and a source of serious bias.*

Linking independent sources can solve this problem by making the local independence assumption plausible across sources, while potentially allowing for local dependence within sources. For example, linking a labor force survey to administrative employer records yields two indicators per time point that are more likely conditionally independent. However, record linkage introduces a new problem: the records may be erroneously linked—with as yet unknown consequences.

This paper investigates the effects of linkage error on Hidden Markov Model estimates. False negative linkage turns out to be problematic only if it is large and highly correlated with the dependent variable. Perhaps surprisingly, our results suggest that under many conditions false- positive linkage error is simply another source of misclassification that the HMM automatically absorbs in the error rate estimates and corrects in the transition estimates. In these cases, measurement error modeling already accounts for linkage error. Our results also provide an indication as to where these conditions break down and more complex methods are needed.

Keywords: *linkage error; misclassification, classification error, measurement error; latent class models; latent class analysis (LCA); Hidden Markov model (HMM)*

Word count: 6,343

[1] This work was supported by Statistics Netherlands and Vrije Universiteit Amsterdam; the authors also thank Peter Paul de Wolf (Statistics Netherlands) and the members of the SILC research group of the Vrije Universiteit Amsterdam for reviewing the paper and providing valuable and constructive feedback.
[2] Vrije Universiteit Amsterdam
[3] Statistics Netherlands
[4] University of Utrecht

1. INTRODUCTION

Survey data, in spite of survey researchers' best efforts to prevent them, will always contain measurement errors (Alwin 2007, Biemer and Stokes 2004, Kuha and Skinner 1997). Where unaccounted for, such errors severely bias estimates of relationships among variables (Carroll et al. 2006, Fuller 1987, Kuha and Skinner 1997, Saris and Gallhofer 2007). Therefore, it is essential to estimate such errors so their biasing effects can be removed. For categorical variables, an attractive method of doing so (without requiring "gold standard" validation data that are assumed to be perfect) is latent class modeling (LCM) (Biemer 2011, Vermunt and Magidson 2002).

Latent class models use repeated indicators of some categorical phenomenon of interest as input, while their output consists of estimates of the classification error rates of these indicators (the "measurement parameters"). These models also provide estimates of the "structural parameters", which measure quantities of scientific interest, such as prevalence, relationships with external variables, or transitions over time. If the repeated indicators that are used as inputs are part of a set of different survey questions intended to measure a single underlying latent variable, the LCM becomes a "latent structure model"—similar to the linear factor model for continuous data. When the repeated indicators are repetitions of the same question at different time points, the "hidden" (or "latent") Markov model results (HMM) and its continuous data-analog is the quasi-simplex approach (Alwin 2007, Alwin, Baumgartner, and Beattie 2017). In this paper, we focus on the HMM approach that is regularly applied to categorical survey data (Biemer 2011, Biemer et al. 2017, Edwards, Berzofsky, and Biemer 2017). The great advantage of LCMs is that all indicators are allowed to contain errors; thus, LCMs can estimate the quality of a survey indicator without requiring perfect data to compare it to. However, this exciting feature of LCM does not come cheap: a payment in untestable assumptions is required (e.g. Oberski, Hagenaars, and Saris, 2015), namely the

"local independence" assumption that the errors in the repeated indicators occur independently. For example, if a respondent erroneously considers her zero-hours contract "full time" in January, when repeating our question in February, we must assume that this same respondent is just as likely as anybody else (including those who have not made any mistakes earlier) to make a mistake.

The local independence assumption is not only unrealistic, but, with data from a single repeated indicator, also harmful and undetectable. It is unrealistic, because common method variance (i.e. variance attributed to the measurement method as opposed to the constructs the measure represents) is typically found in studies able to detect it (Saris and Gallhofer 2007) and because it seems likely that any personal "style" in answering a survey question will carry over time, as shown, for instance, by Billiet and Davidov (2008). It is harmful because various authors have shown that ignoring local dependence leads to bias in classification error estimates (Georgiadis et al. 2003, Qu and Hagdu 2012, Torrance-Rynard and Walter 1997, Vacek 1985). And, finally, it is undetectable with data from a single repeated indicator, because the local independence assumption is precisely what identifies the model in the first place. While it is possible to detect and model local dependence in LCMs (Hagenaars 1988, Oberski 2016), these parameters are only generally identifiable in HMMs if a second indicator is obtained at each time point (Hagenaars 1990). Such an indicator should then plausibly contain errors independent of the errors present in the first indicator.

An attractive solution to the problem of local independence is therefore to link the survey to external records from administrative registers. Such records can contain considerable errors (Bakker 2012, Oberski 2017, Oberski et al. 2017, Scholtus, Bakker, and Delden 2015), but fortunately, we do not require the linked registers to be perfect, but only that register errors are independent of survey errors. With linked survey-register data, this indeed seems plausible in many cases. For example, linking the respondents' survey answers

to official records obtained through the employer, it appears unlikely that the example "zero-hours" mistake is related to a database error recorded by the tax authority. By combining registers and surveys, it becomes possible to allow for local dependence within each source. Previous studies have done so, and found considerable local dependence (Oberski et al. 2017, Pavlopoulos and Vermunt 2015), confirming the importance of relaxing this assumption and the attractiveness of data linkage.

However, while linkage resolves an important set of problems, it also introduces a new challenge: linkage error. Again, such errors are known to bias estimates of interest when left unaccounted for (Di Consiglio and Tuoto 2014, Lahiri and Larsen 2005). For example, due to the lack of unique identifiers in many datasets, record linkage is often performed probabilistically through a combination of variables such as birth date, gender and address. An interesting case in point is the US Census Bureau's Person Identification Validation System, which assigns Protected Identification Keys based on probabilistic linkage (Bond et al. 2014), and forms the basis of broad host of social science data analyses and official statistics in the United States. Similar systems are used in other countries.

In short, problems of measurement error modelling can be tackled using linkage, but this introduces linkage error. Because both errors affect estimates of interest, it is important to examine the extent to which linkage and measurement error may help or hinder one another. To our knowledge no studies exist that investigate such potential interactions between these two error sources.

In this paper, we investigate the effect of linkage error on the estimates of latent class models using a two-indicator Hidden Markov Model. To mimic a real-life situation, we use a dataset of linked survey-register data from Statistics Netherlands concerning the type of employment contract at different time points. We postulate for the moment that this dataset has no linkage error. We are able to investigate what the effect of linkage error on HMM

estimates is by simulating varying degrees of intensity and dependency of erroneous matches ("false positives") and erroneous non-matches ("false negatives"). Effects of linkage error are examined in both the "measurement part" of the model, and the "structural part".

The remainder of the paper is structured as follows: first the background section discusses the use of HMMs in the context of measurement error correction and the topic of record linkage and linkage error; second, the data sources used are described; third, the methodology applied is discussed; fourth, the results obtained are presented; and fifth, some concluding remarks are provided.

2. BACKGROUND

Measurement error is commonly thought to be a serious threat to using surveys. When the observed answers to survey questions differ from the values the researcher was trying to measure, conclusions will be affected (see e.g. Saris and Gallhofer, 2007). In categorical data, measurement error is known as classification error: there is a hypothetical cross-table between the true values and the observed ones, and this cross-table does not show perfect correspondence. In practice, classification error rates can rarely be directly observed, because even when it is possible to obtain so-called "validation data", such data often turn out to contain errors themselves; in such situations, measurement error modeling is necessary. Measurement error models allow estimation of the error rates, as well as the "structural" parameters of substantive interest, without requiring a gold standard data source; they do this by assuming conditional independence of errors given the true values. The current paper uses a specific measurement error model, the HMM, which can be applied to longitudinal data, such as that obtained from panel data.

2.1 Use of HMMs to estimate classification error

Hidden (or Latent) Markov Models (HMMs) are a group of latent class models increasingly used to estimate and correct for measurement error in longitudinal (or time series) categorical

data (Biemer 2011, 2004). We will discuss two forms of HMMs: the basic form modeling a single source over time, commonly applied across the literature, and the extended form proposed by Pavlopoulos and Vermunt (2015).

The basic HMM operates under the assumption that, at each time point *t*, the observed answer Y_t is generated *independently* with some probability $P(Y_t|X_t)$ from the true, but unobserved, value X_t. Because the generation of Y_t only involves X_t and is independent of all other observed and true values, the observed distribution, $P(Y)$, factorizes as

$$P(Y) = \prod_{t=0}^{T} P(Y_t|X_t)P(X)$$

And because X_t is unobserved, the observed data are marginalized over the true data:

$$P(Y) = \sum_{x=1}^{K} \prod_{t=0}^{T} P(Y_t|X_t)P(X)$$

Unless all probabilities $P(Y_t|X_t)$ equal 1 for a unique category of X_t, classification error occurs. The unobserved true values, meanwhile, follow a Markov or "AR(1)" process, in which each value carries over only partly to the next time point:

$$P(X) = P(X_0)\, P(X_1|\, X_0) \ \ldots\ P(X_T|\, X_{T-1}).$$

The full model (i.e. the probability of following a certain observed path for an individual *i*) then is

$$P(Y) = \sum_{x_0=1}^{k} \sum_{x_1=1}^{k} \ldots \sum_{x_T=1}^{k} P(X_0) \prod_{t=1}^{T} P(X_t|X_{t-1}) \prod_{t=1}^{T} P(Y_t|X_t)$$

The parameters to be estimated of this model are the unknown initial state probabilities - $P(X_0)$, time-specific transition probabilities- $P(X_t|X_{t-1})$, measurement error probabilities- $P(Y_t|X_t)$ or transformations of these probabilities – typically the logit.

The HMM is attractive for two reasons. First, it allows for hidden change over time, $P(X_t|X_{t-1})$, and simultaneously estimates and accounts for classification error. Second, it is possible to estimate the parameters from panel data on single repeated indicators that are

often already collected as part of longitudinal surveys. This identifiability is a consequence of conditional independence: comparing observed dependence among variables *proximal* in time on the one hand, with observed dependence and *distal* in time on the other, yields differing true changes but identical error attenuations. This differential allows error attenuation to be pried apart from true change by observing a single indicator in at least three time-points. To put it simply, the model is identifiable when the measurement error is assumed time homogenous and the indicator is observed for at least three consecutive time-points (An et al. 2013)

However, when conditional independence does not hold, a single indicator at each time point is no longer sufficient. A general model that captures naturally occurring local dependence is not identifiable from a single indicator, and its parameters can therefore not be consistently estimated. Intuitively, the reason for this is that true stability (change) can no longer be pried apart from error attenuation because the local dependence becomes an alternative explanation for stability. Models that rule out this alternative, such as lagged random effects models, are identifiable with more than three time-points, but may still be too restrictive in practice. Alternatively, a model with local dependence can be identified by linking additional data sources, in which the same phenomenon is measured at least twice at each time point. This extended - multiple-indicator - Hidden Markov Model can then, identifiably, account for local dependence among the errors thanks to the presence of an additional indicator.

Pavlopoulos and Vermunt (2015) proposed one such model (see also Manzoni et al. (2010)), which is shown schematically in Figure 1.

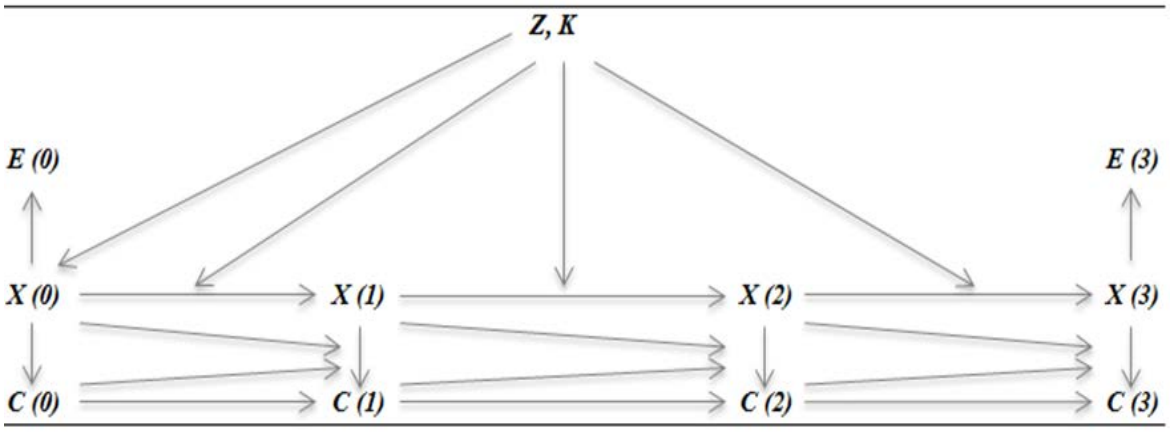


Figure 1- Path diagram for the extended hidden Markov model with two (partially) observed indicators (as in Pavlopoulos and Vermunt, 2015)

In this model, instead of factoring the conditional joint distribution of the observed indicators given the latent states P(Y|X) into separate time points, $\prod_{t=0}^{T} P(Y_t|X_t)$, this distribution is structured by allowing a register error at time $t-1$, i.e. Y_{t-1} not equal to X_{t-1}, to be copied over to time *t* with probability $P(Y_t|X_t, X_{t-1}, Y_{t-1})$. For the survey the local independence assumption is not relaxed. Thus, local dependence can be accounted for in the model by incorporating an additional linked data source. What is more the linkage also allows incorporating dependency of the latent initial state probabilities and transition rates on covariates and latent class membership (which account for observed and unobserved heterogeneity). The initial state probability is then depicted as $P(X_0|Z,k)$ and the transition probabilities as $P(X_t|X_{t-1}, Z, k)$. The following section considers the potential consequences of this linkage.

2.2 Record linkage and linkage error

Record linkage is defined as a process that matches records, usually from two or more distinct data sources, and attempts to select those matches that belong to the same person or unit. The record linkage framework makes use of one or more data fields, which contain the same identifying information in all sources and which are referred to as matching variables (Winkler 1999, Armstrong and Mayda 1993).

There are two main types of linkage methods: deterministic record linkage and probabilistic record linkage.

Deterministic record linkage relies on accepting those pairs as true matches for which there is an exact agreement on the matching variables in all data sources. It usually makes use of a relatively small number of matching variables and is most commonly applied in the presence of the same unique identifier (such as national insurance number or social security number) in the data sources being linked. Deterministic linkage is encouraged primarily when the data sources are of very high quality as coding errors are likely to result in not linking true matches (Blakely and Salmond 2002).

Probabilistic record linkage, on the other hand, tends to make use of a larger number of matching variables and does not require an exact agreement on all of them for a pair to be considered a true match. It is usually applied in the absence of one unique identifier accurately assigned to all records in the datasets under consideration. Probabilistic linkage reviews the extent to which the different matching variables are consistent across the data sources for each matched pair and based on that assigns a weight to it. The weight determines the probability of a match being correct and, as such, whether it should be regarded as a "true" or "false" match (Armstrong and Mayda 1993, Blakely and Salmond 2002, Bohensky et al. 2010, Fellegi and Sunter 1969, Winglee, Valliant, and Scheuren 2005).

While record linkage is undoubtedly an important tool that allows combining information from various sources, it is also associated with different types of errors. The size of these errors depends on the quality of the available unique identifiers, like a personal identification code or combinations of name, address, and birth date. The availability of these identifying information depends on among else privacy laws. Over the last decade more and more data sources have become available with a lack of high quality unique identifiers (see appendix 1). Therefore, probabilistic linkage became more popular.

Linkage errors occur in situations where the data sources do not capture the same cases consistently. Namely, due to missing or inaccurate data, some records that correspond to the same person or unit might not be linked - a phenomenon referred to a false-negative linkage error. Alternatively, as a result of coding or measurement errors, unrelated records can be wrongfully linked – a situation referred to as false-positive linkage error (Winglee, Valliant, and Scheuren 2005, Bohensky et al. 2010).

Record linkage and its associated errors can be formulated using files drawn from two populations[5]- file A containing N_A records and file B containing N_B records, and a set C containing record pairs which are formed as cross-product of files A and B. This set is denoted by $C = \{(a, b); a \in A, b \in B\}$ and its number of records equals to $N = N_A \times N_B$ (Armstrong and Mayda 1993).

The aim of record linkage is to divide set C into two separate sets - one which includes true matches (here denoted by M) and one which includes true non-matches (here denoted by U). This is often done by examining the data contained in files A and B and deciding whether the records certainly belong to the same entity (i.e. are a link, denoted by A_1), possibly belong to the same entity (i.e. are a possible link, denoted by A_2) or certainly belong to different entities (i.e. are a non-link, denoted by A_3) (Armstrong and Mayda 1993, Fellegi and Sunter 1969).

False-positive and false-negative types of error occur respectively when (1) a record pair which belongs to the true non-match set (U) is registered as a link (A_1) and (2) when a record pair belonging to the true match set (M) is registered as a non-link (A_3). Thus, the false-positive linkage error can be denoted by $P(A_1|U)$ and that of false-negative error by $P(A_3|M)$. Possible links (A_2) are considered when the data provide insufficient evidence to categorize the pair as either a link or a non-link and when the overall error levels are below or

[5] Each of those datasets can either contain all entities belonging to its corresponding population or a random sample of those.

equal to the acceptable levels of classification error which are usually specified prior to the linking process (Armstrong and Mayda 1993).

In order to determine which decision rule is most appropriate when classifying matched records, it is essential to estimate the classification error rates associated with the various rules under consideration. There are numerous approaches and frameworks available in the literature that allow approximating those rates. Examples include that of Bartlett et al. (1993) and Belin and Rubin (1995). While the former relies on selecting a random sample of pairs from set *C* and manually determining their true match status, the latter provides a framework for the use of information obtained from a pilot study conducted by the authors. Those methods are, however, often difficult or even impossible to implement in practice due to time constraints and the unavailability of information (primarily regarding the linkage probabilities).

Lahiri and Larsen (2005) proposed the following data-generating process when linkage error occurs (i.e. the probability of linkage error occurring):

$$P(Y) = P(U)\,P(Y|U) + (1 - P(U))\,P(Y|M)$$

Where $P(U)$ is obtainable from the Felligi-Sunter analysis. They then proceeded to consider $P(Y|U)$ as independent and specify a regression model for elements of $P(Y|M)$. In this paper, we suggest that the above model is simply a latent class model with two classes (matches and non-matches), in which part (M) of the data follows the expected data-generating process and part (U) is independent. While we do not explicitly model this we investigate the consequences of this fact when the data-generating process is the "extended" Hidden Markov measurement error model specified in the previous section in a real-data example.

3. DATA

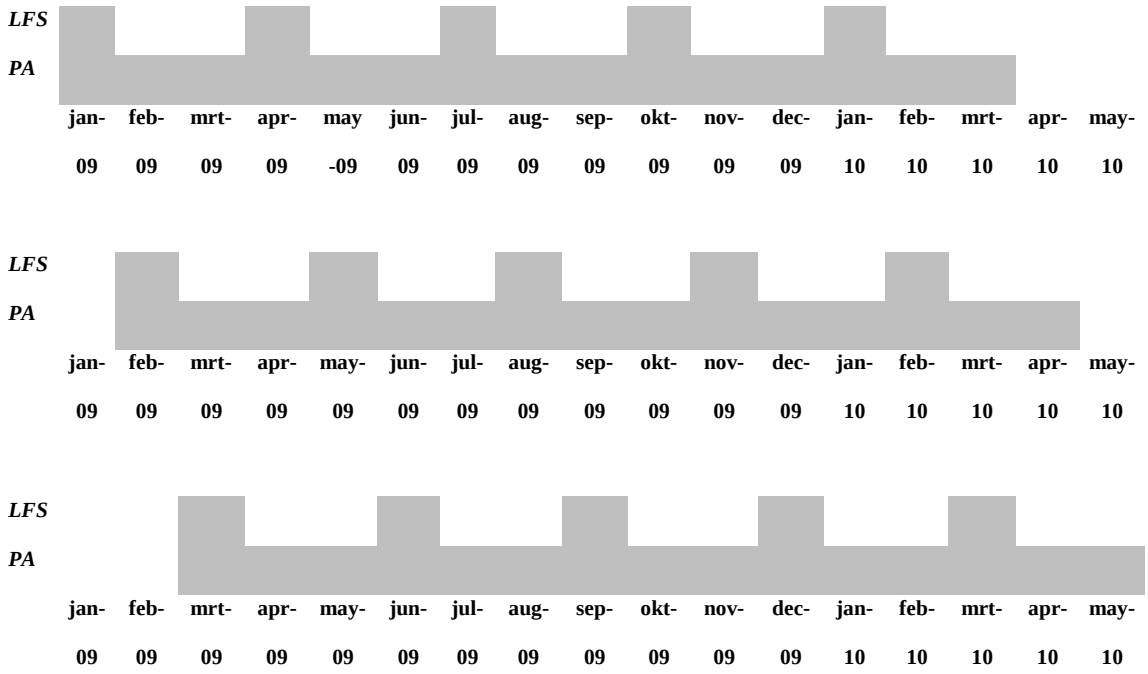
The dataset used in our analysis contains information from the Netherlands' Labour Force Survey (conducted by Statistics Netherlands) and the Employment Register data (in Dutch:

Polisadministratie or in short ER). In this paper, we assume that the resulted dataset contains only correct links or put differently is not subject to linkage error and, thus, the effect of linkage error is determined by simulating false negatives and false positives in this dataset.

The Dutch Labour Force Survey (LFS) is a survey providing information about the relationship between individuals and the labour market based on an address sample. As of the last quarter of 1999, the survey has been a rotating panel survey which consists of five waves conducted every 3 months.[6] The Employment Register data is a register dataset managed by the Dutch Employee Insurance Agency (UWV). It contains monthly information on wages, benefits, and labour relations and covers all insured employees in the Netherlands. While the dataset combines information from various sources, the core information is delivered by employers to the Dutch Tax Authorities (in Dutch: *Belastingdienst*) for tax purposes.[7] The data from both the LFS and the ER are linked at the individual level to the PR. Thus, the target population is restricted to individuals registered in the Netherlands. See Appendix 2 for a detailed description of the record linkage process.

Our sample consists of 8,886 LFS- respondents aged 25 to 55 who have first participated in the survey in the first trimester of 2009. The dataset corresponds to the time period between January 2009 and May 2010 and contains information for a total of 15 months for each individual. The ER variables are observed on a monthly basis and the LFS variables are observed every 3 months. Figure 2 provides an illustration of the sample.

[6] http://www.cbs.nl/en-GB/menu/methoden/dataverzameling/dutch-labour-force-survey-characteristics.htm
[7] http://www.uwv.nl/overuwv/english/about-us-executive-board-organization/detail/organization/data-services



* The figure illustrates how the 3-monthly rotation panel of the LFS corresponds to monthly observations from the ER. A grey shaded cell indicates a valid observation.

Figure 2- An illustration of the sample

The dataset is unbalanced for the LFS as it suffers from attrition and has, for the non-survey months, observations missing completely at random. More specifically, the first wave of the survey included 8,708 individuals , the second 7,458, the third 6,856, the fourth 6,739 and the fifth 6,560. While formally the ER cannot suffer from drop-out (as all employers are obliged to submit their reports), 2,619 observations are missing. Those observations are also assumed to be missing completely at random.

The main variable of interest in our analysis is the individual's employment contract type for his or her main job.[8] This variable takes on three distinctive and mutually exclusive values: 'permanent contract' - i.e. a contract for an unlimited duration of time - 'temporary contract' - i.e. a contract for a fixed/ limited duration of time - and 'other', which contains all other alternatives, such as self-employment, unemployment, unpaid employment and full-time education.

[8] Any secondary jobs are ignored in this analysis

4. METHODOLOGY

4.1. Empirical model

Our empirical strategy consists of a simulation analysis in which we make use of a two-indicator Hidden Markov Model (HMM) where one of the indicators is the individual's contract type according to the ER and the second is the contract type according to the LFS. While the model could be extended further, following Pavlopoulos and Vermunt (2015) and Pankowska et al. (2017), to include covariates and time-dependency of the error, our simulations are based on a simplified model which retains the local independence assumption and the Markov assumption to avoid unnecessary complications. However, the model assumes the latent transition probabilities to be time heterogeneous.

The following equation estimates the probability of following a certain observed path according to our model:

$$P(C_i = c_{i,}\, E_i = e_i) = \sum_{x_0=1}^{3}\sum_{x_1=1}^{3}\dots\sum_{x_T=1}^{3} P(X_{i0} = x_0)\prod_{t=1}^{T} P(X_{it} = x_t | X_{i(t-1)} = x_{t-1})\prod_{t=1}^{T} P(C_{it} = c_t | X_{it} = x_t)\prod_{t=1}^{T} P(E_{it} = e_t | X_{it} = x_t)^{\delta_{it}}$$

Where C_{it} and E_{it} denote the contract type of person i at month t according to the ER and LFS, respectively, with i = 1, ..., N and t = 1, ..., 17.[9] To account for the fact that the contract type according to the survey (E_{it}) can only be observed every third month the indicator δ_{it} is included in the model; δ_{it} equals 1 if the survey information is available for a given month and 0 if it is missing. The model also includes a latent (unobserved) variable (X_{it}) which represents the individual's actual contract type at time t. Both the observed indicators and the latent variable (which are referred to in the model as c_t, e_t, and x_t, respectively) consist of three categories - permanent, temporary and other type of contract. A path diagram for our model is provided in Figure 3.

[9] As described in the Data section, in the analysis, data from January 2009 until March 2010 is used which corresponds to 17 months is used, hence *t* runs from 1 to 17

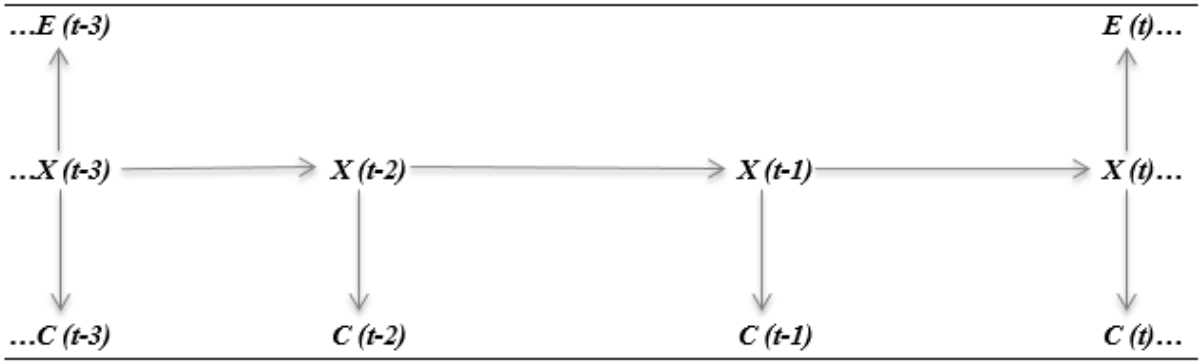


Figure 3- Path diagram for the hidden Markov model with two (partially) observed indicators

We apply the model to different scenarios in which various simulated types of either false-negative or false-positive linkage error are introduced to the original dataset. Each scenario is simulated 200 times and each simulation is based on a different, randomly generated uniform distribution.

We then investigate the bias introduced by the error by comparing the obtained 3-monthly transition rates from temporary to permanent employment to those estimated using the original linked dataset (which is assumed linkage error free).

For the linkage error simulations, we use the R Statistical Software and we employ the Latent GOLD software to estimate the HMM given the various simulated datasets. When estimating HMMs, Latent Gold uses a modified version of the Expectation-Maximization (EM) algorithm called the forward- backward (or Baum-Welch) algorithm. All missing values are treated as missing completely at random (MCAR) and the analysis makes use of the sampling weights of the survey.

4.2. False-negative error simulations

When investigating the effect of false-negative linkage error on the accuracy of our model estimates, we simulate a set of scenarios in which we introduce relatively high, medium and low overall exclusion error to our data which reflect an overall exclusion error of 20%, 10%

and 5% respectively. The error is equal to the proportion of correct links in the data that are erroneously excluded from it - the parameter which is varied in the simulations. [10]

Each of the aforementioned scenarios consists of two sub-scenarios where linkage error is correlated with a variable that is positively related to the model outcome (i.e. transition rates). The exclusion of primarily those individuals who are more likely to transition from temporary to permanent employment should lead to more biased estimates and allows investigating more extreme scenarios. In the first specification, the probability of exclusion for each individual depends on age (i.e. younger individuals have higher probabilities of being excluded). The choice of this covariate is motivated by the fact that young individuals tend to have higher residential and employment mobility and are thus more susceptible to linkage error. Apart from being moderately correlated with the model estimates, age is also relevant as it is a part of the linkage key.

In the second specification, the exclusion probabilities depend on whether at least one 3- monthly transition from temporary to permanent employment occurred according to the register data. This variable was chosen as it is very highly correlated with the model estimates and thus allows modelling an extreme scenario with strong potential biasing effects. In this scenario individuals who have transitioned at least once have a higher exclusion probability.

The simulations are designed in such a way that as we move from scenarios with low overall exclusion errors to scenarios with high overall exclusion errors, the additional linkage error comes increasingly from the values of the variable that corresponds to groups where we expect more linkage error. In the case of age, the exclusion probability of young people (aged 25 to 34) is set to 0.70, 0.30 and 0.15 in the high (20%), medium (10%) and low (5%) overall linkage error scenarios respectively while the exclusion probability of older people (aged 35 to 54) equals to 0.01 in all three cases. For the scenarios where the exclusion probability

[10] When analysing the effect of false-negative linkage error, the scenario in which individuals are selected for exclusion completely randomly (i.e. not based on any characteristics) has been omitted as this case would be equivalent to missingness completely at random which by definition does not bias model estimates.

depends on whether a transition occurred, the exclusion probabilities of those individuals who experienced a transition are 0.90, 0.34 and 0.15 for the scenarios with an overall error of 20%, 10% and 5%, respectively[11].

This approach significantly increases the potential of the scenarios with high overall exclusion error to bias the results. Namely, a higher overall false-negative error does not only indicate that a larger proportion of individuals were excluded from the sample but it also implies that certain individuals (i.e. younger or who transitioned from temporary to permanent employment according to the register data) are more overrepresented in the excluded group. Thus, the remaining (post exclusion) sample is less representative of the population under consideration.

What is more, the covariates that are shown by previous research to have a significant effect on the latent transitions are not included in the second step of the analysis, i.e. when estimating the temporary to permanent employment transition rates using the HMM. Therefore, while these covariates are allowed to determine the level of exclusion error, they are not controlled for when estimating the HMM. Thus, the simulated datasets which contain false-negative linkage error are equivalent to a dataset containing missing not at random data (with respect to the transition rates estimated).

4.3. False-positive error simulations

The analysis of the false-positive error, similarly to that of the false-negative, follows 3 steps. The first step determines the overall level of mislinkage and the individual probabilities of an erroneous link. In the different scenarios considered, those probabilities are either assigned at random or are based on one of the same two variables as in the false-negative linkage error simulations- age and whether a transition occurred at least once according to the register data. Again, for each type of probability (i.e. random, age-dependent and transition- dependent) we

[11] The exclusion probabilities are set in such a way that they overall approx. result in 20%, 10% and 5% of all individuals being excluded.

consider three scenarios: a high (20%), medium (10%) and low (5%) overall mislinkage error rate[12]. Moreover, the probabilities are set in such a way that higher overall error levels also indicate a greater propensity for mislinkage of individuals with certain characteristics (i.e. younger and who have transitioned according to the register).

In the second step, the false-positive error is simulated in the following way: a number of individuals is selected at random according to the aforementioned design. Each of those individuals, here referred to as individual A, is either randomly assigned to another person (in the first set of scenarios) or matched to a similar person based on age, gender, education level and ethnicity (in the second set of scenarios). The person Individual A is matched to acts as a donor and is here referred to as individual B. Then, in the register data, the values of individual A for the contract type variable are replaced those of individual B.

The second set of scenarios, whereby relatively similar individuals (rather than random) are matched, is introduced to simulate a more realistic linkage error scenario which is more reflective of actual potential mismatches. This is particularly true for probabilistic matching where the linkage key often comprises, among others, of individual characteristics such as birthdate or gender.

The third and final step of the simulation analysis is parallel to that of the exclusion error. Each of the simulated datasets is fitted to our HMM and the outcomes of these models are checked for consistency across the simulated scenarios by comparing the results to those obtained when using the original dataset.

5. RESULTS

5.1. The effect of false-negative error

The results obtained from the simulations of 5%, 10% and 20% false-negative error scenarios whereby the exclusion probability depends on age and on whether a transition from

[12] These overall error rates equal the proportion of wrong links (or actually wrongly linked individuals).

temporary to permanent employment occurred in the register data are presented in Table 1 and Figure 4. Table 1 provides the mean estimated 3-months' transition rate from temporary to permanent employment for each of the scenarios as well as the absolute and relative bias introduced by the linkage error. These biases are estimated by comparing the obtained transition rates to those estimated using the original dataset, which is assumed to be linkage error free. Figure 4 provides an illustration of the relationship between the levels of the age-dependent and transition-dependent false-negative linkage errors and the relative bias that they introduce.

Overall, when the exclusion probability depends on a covariate that is weakly or moderately correlated to latent transition rates (i.e. the model estimates), such as age, the bias is negligible irrespective of the level of the overall false-negative linkage error introduced. In more detail, the bias ranges from 0.001 to 0.003 in absolute terms, which translates to a relative bias of 4.35% at most.

A significantly different picture emerges when the exclusion probability depends on a covariate that is strongly correlated with the model outcomes, such as whether at least one 3-monthly transition from temporary to permanent employment occurred according to the register data. In more detail, the employment transition rates in those scenarios are heavily underestimated leading to a substantial, non-negligible bias. In relative terms, the bias ranges from 10.39% for an overall linkage error of 5% to 24.35% when the linkage error amounts to 10% and 84.63% when the linkage error equals to 20%.

The results obtained show that the extended, two-indicator HHM used to correct for measurement error is only sensitive to false-negative linkage error in very extreme situations. That is, we only obtain biased results when the individual level probabilities of being excluded from the sample depend on a covariate that is (extremely) highly correlated with the variable for which measurement error is corrected. In other, less extreme and more realistic

scenarios, as can be seen when the exclusion probabilities depend on age, the bias is incredibly small and thus the results obtained by the HMM can be considered accurate.

Table 1- Simulation results- false-negative linkage error

Error type	Scenario- the probability of being excluded:	Overall error (approx.)	High exclusion probability	Low exclusion probability	Temporary to permanent transition rate		
					Transition rate	Absolute bias	Relative bias (%)
No error	Original HMM	0	-	-	0.069	-	-
False-negative	Depends on age	0.05	0.15 (young)	0.01 (mid/ old age)	0.066	0.003	4.35
		0.1	0.3 (young)	0.01 (mid/ old age)	0.067	0.002	2.98
		0.2	0.7 (young)	0.01 (mid/ old age)	0.068	0.001	1.81
	Depends on transitions occurring	0.05	0.15 (transition occurred)	0.05 (no transition)	0.062	0.007	10.39
		0.1	0.34 (transition occurred)	0.085 (no transition)	0.052	0.017	24.35
		0.2	0.9 (transition occurred)	0.15 (no transition)	0.011	0.058	84.63

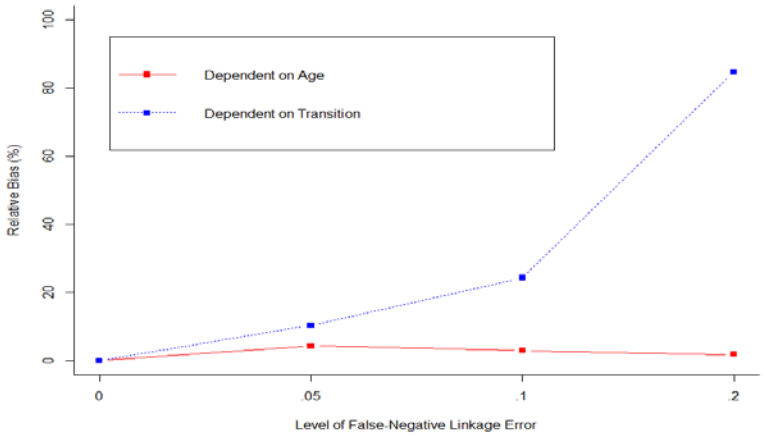


Figure 4- Relative bias by overall level of false-negative linkage error

5.2. The effect of false-positive error

The results obtained in the various false-positive linkage error scenarios are presented in Table 2 and Figure 5.

Overall, the bias introduced by false-positive linkage error is virtually non-existent for all the scenarios where the mislinkage probability is random or depends on age[13]. In contrast, those scenarios for which the probability of mislinkage depends on whether or not the individual experienced a 3-monthly transition from temporary to permanent employment according to the register data are characterized by a very high bias. These findings are consistent for both the scenarios in which an individual is mislinked to a randomly selected individual and where the individual is mislinked to someone who is similar to them with regards to their age, gender, education level and ethnicity.

In more detail, the scenarios in which the mislinkage probability is random or depends on age and the individual is mislinked with a random donor or a similar one in terms of age, gender, education level and ethnicity lead to a relative bias of under 1% in most cases with three exceptions where the bias amounts to just under 3%. It is worthwhile noting that, while there are 3 groups of false-positive error linkage scenarios where an individual is mislinked to a random donor- where the mislinkage probability is random, depends or age and on a transition occurring- there are only two groups of scenarios where the individual is mislinked to a similar donor- where the probability is random and depends on a transition occurring. The set of scenarios where the mislinkage probability depends on age was omitted in the second step as the bias introduced by those scenarios was very similar and equally negligible to one introduced by random scenarios.

Those scenarios in which the mislinkage probability depends on whether a transition occurred in the register data, though, lead to a relative bias of around 10%, 20% and over 60% when the overall mislinkage equal to 5%, 10% and 20%, respectively. The bias substantially increases with the overall transition-dependent mislinkage level, a relationship which cannot be observed for the random or age- dependent mislinkage.

[13] It is worthwhile mentioning that while in absolute terms the bias introduced by age-dependent mislinkage is negligible, in relative terms it does increase considerably as the overall rate of the error goes up. That is, the bias increases from 0.02% to 0.93% and 2.29% for 5%, 10% and 20% level of error, respectively.

Table 2- Simulation results- false-positive linkage error

Error type	Scenario- the probability of being mislinked:	Overall error (approx.)	High exclusion probability	Low exclusion probability	Temporary to permanent transition rate		
					Transition rate	Absolute bias	Relative bias (%)
No error	Original HMM	0	-	-	0.069	-	-
False-positive; mislinkage with random donor	Random	0.05	-	-	0.069	0.000	0.32
		0.1	-	-	0.069	0.000	0.01
		0.2	-	-	0.068	0.000	0.67
	Depends on age	0.05	0.15 (young)	0.01 (mid/ old age)	0.069	0.000	0.02
		0.1	0.3 (young)	0.01 (mid/ old age)	0.068	0.001	0.93
		0.2	0.7 (young)	0.01 (mid/ old age)	0.067	0.002	2.29
	Depends on transition	0.05	0.15 (transition occurred)	0.05 (no transition)	0.064	0.005	7.57
		0.1	0.34 (transition occurred)	0.085 (no transition)	0.055	0.014	20.45
		0.2	0.9 (transition occurred)	0.17 (no transition)	0.024	0.044	64.52
False-positive; mislinkage with similar donor	Random	0.05	-	-	0.067	0.002	2.88
		0.1	-	-	0.067	0.002	2.67
		0.2	-	-	0.069	0.000	0.31
	Depends on transition	0.05	0.15 (transition occurred)	0.05 (no transition)	0.061	0.008	11.28
		0.1	0.34 (transition occurred)	0.085 (no transition)	0.051	0.018	26.42
		0.2	1 (transition occurred)	0.15 (no transition)	0.012	0.057	82.54

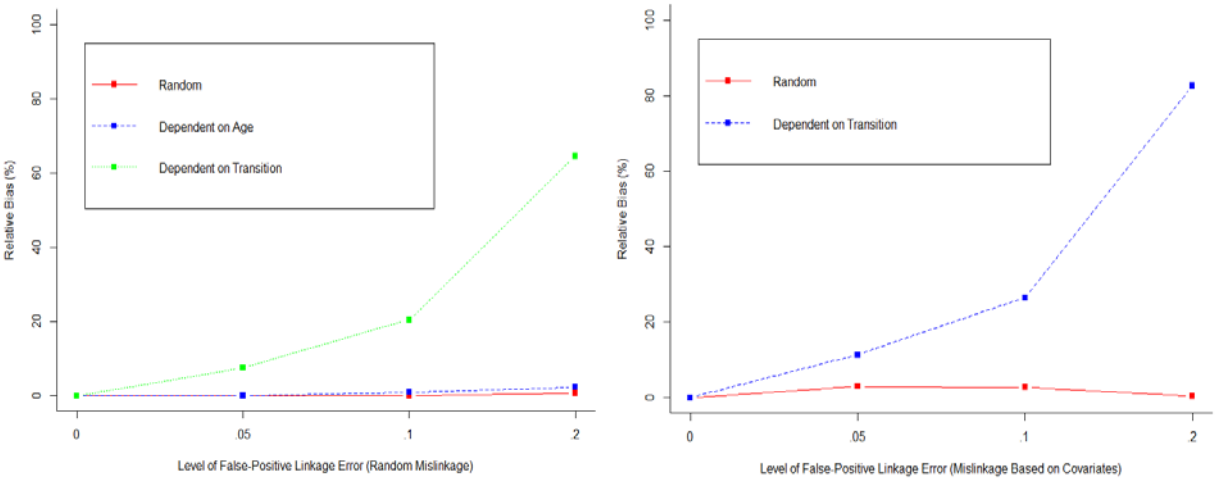


Figure 5- Relative bias by overall level of false-positive linkage error (random mislinkage and based on covariates)

The findings that suggest that there is virtually no effect of random mislinkage or a mislinkage based on a covariate moderately correlated with the model outcomes seem rather puzzling at first glance. That is, while the literature suggests that even a relatively small false-positive linkage error tends to bias model estimates, our results are not biased by very high levels of error. However, if one reconsiders linkage error as a form of systematic measurement error, which occurs consistently for a mislinked individual throughout all time points and for one of the data sources, one can infer that mislinkage can be corrected for by the HMM. This intuition receives some support in the misclassification (measurement error) results obtained. That is, as can be seen in Figure 6, when the overall level of linkage error goes up, the proportion of measurement error increases as well, this is in particular visible for the LFS data[14]. Those results confirm the intuition and suggest that under many conditions false- positive linkage error is simply another source of misclassification that the HMM automatically absorbs in the error rate estimates and corrects in the transition estimates.

Thus, the results suggest that (1) scenarios in which mislinkage is random or depends on a covariate moderately correlated with the HMM estimates do not result in a significant bias while those where mislinkage is based on a covariate highly correlated with the estimates do; (2) the bias is not minimized when similar individuals are mislinked and where the covariates determining the degree of similarity are not strongly correlated with the model estimates; (3) mislinkage appears to be largely detected and corrected for by the HMM, in particular when it is not based on a covariate highly correlated with the model estimates. Such information (regarding the strength of the correlations), however, is often difficult if not impossible to obtain in practice.

[14] This pattern is not observed in the ER data as the simplified HMM does not account for autocorrelation of the error in the register data which is the main factor causing measurement error and so it fails to capture measurement error in this data source altogether and assumes the register data to be virtually error-free.

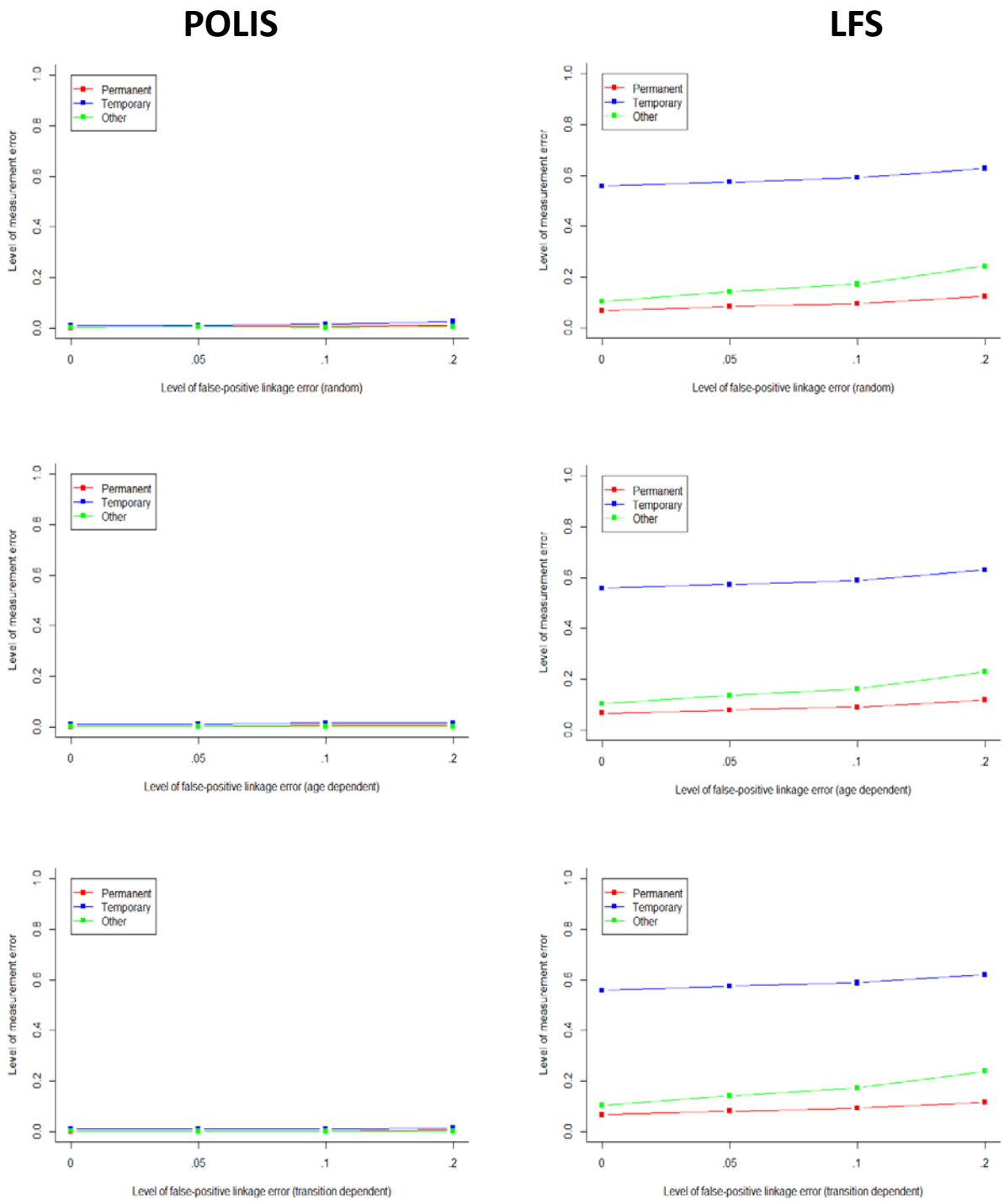


Figure 6- Level of measurement error by type and level of mislinkage

6. CONCLUSIONS

To conclude, latent class models have been gradually replacing more traditional methods in correcting for measurement error in categorical data. The main reason behind this transition is the growing criticism of the availability and accessibility to 'gold standard', error-free data but also the fact that latent class modelling do not rely on the existence of such an error-free data source.

While an appealing alternative tool, latent class models also suffer from a substantial shortcoming. Namely, they rely on an assumption which can rarely be fulfilled or validated in practice: the local independence assumption. This assumption implies that measurement error is independent conditional on the true value only.

An attractive solution that allows modelling more realistic scenarios, whereby the local independence assumption is either fulfilled or can be relaxed, is linking data from various sources. However attractive, this approach introduces a new challenge: linkage error- a phenomenon which can severely bias model estimates.

In this paper we have investigated the feasibility and attractiveness of linking data to enable modelling more realistic measurement error scenarios. In doing so, we have analyzed the bias introduced by various degrees and types of both false-negative and false-positive linkage errors using a two indicator Hidden Markov Model and linked data on employment mobility from the Dutch Labour Force Survey and Employment Register as an illustrative case.

The results suggest that linkage error only leads to (substantial) bias when the individual probability of being subjected to false-negative or false-positive linkage error depends on a covariate that is extremely highly correlated with the transition error estimates. Furthermore, with regards to false-positive linkage error specifically, the bias introduced is not minimized when similar individuals rather than random ones are mislinked. This,

however, might result from the fact that the covariates used to determine the degree of similarity are not strongly correlated with the model estimates. Finally, based on our illustrative example, it can be inferred that latent class models can, to an extent, correct for false-positive linkage error, albeit in some instances (in particular when the mislinkage probability is highly correlated with the model estimates) it does not eliminate it fully.

REFERENCES

Alwin, D. F. (ed.) (2007), *Margins of error: A study of reliability in survey measurement* (Vol. 547), Hoboken, NJ: John Wiley & Sons.

Alwin, D. F., Baumgartner, E. M., and Beattie, B. A. (2017), "Number of response categories and reliability in attitude measurement", *Journal of Survey Statistics and Methodology*, smx025.

An, Y., Hu, Y., Hopkins, J., and Shum, M. (2013) "Identifiability and inference of hidden markov models": Technical report.

Armstrong, J., and Mayda, J. (1993), "Linkage error rates", *Survey Methodology*, 19, 137-147.

Bakker, B. F. (2012), "Estimating the validity of administrative variables", *Statistica Neerlandica*, 66, 8-17.

Bartlett, S., Krewski, D., Wang, Y., and Zielinski, J. (1993), "Evaluation of error rates in large scale computerized record linkage studies", *Survey Methodology*, 19, 3-12.

Belin, T. R., and Rubin, D. B. (1995), "A method for calibrating false-match rates in record linkage", *Journal of the American Statistical Association*, 90, 694-707.

Biemer, P. (2004), "An analysis of classification error for the revised current population survey employment questions", *Survey Methodology*, 30, 127-140.

--------- (ed.) (2011), *Latent class analysis of survey error* (Vol. 571), Hoboken, NJ: John Wiley & Sons.

Biemer, P., De Leeuw, E., Eckman, S., Edwards, B., Kreuter, F., Lyberg, L. E., Tucker, N. C., and West, B. T. (2017), *Total survey error in practice*, Hoboken, NJ: John Wiley & Sons.

Biemer, P., and Stokes, S. L. (2004). "Approaches to the modeling of measurement errors, " In *Measurement errors in surveys*, eds. P. Biemer, R. M. Groves, Lyberg L. E, N. A. Mathiowetz and S. Sudman, Hoboken, NJ: John Wiley & Sons.

Billiet, J. B., and Davidov, E. (2008), "Testing the stability of an acquiescence style factor behind two interrelated substantive variables in a panel design", *Sociological Methods & Research,* 36, 542-562.

Blakely, T., and Salmond, C. (2002), "Probabilistic record linkage and a method to calculate the positive predictive value", *International journal of epidemiology,* 31, 1246-1252.

Bohensky, M. A., Jolley, D., Sundararajan, V., Evans, S., Pilcher, D. V., Scott, I., and Brand, C. A. (2010), "Data linkage: A powerful research tool with potential problems", *BMC health services research,* 10, 346.

Bond, B., Brown, J. D., Luque, A., and O'hara, A. (2014) "The nature of the bias when studying only linkable person records: Evidence from the american community survey": Center for Administrative Records Research and Applications Working Paper.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (ed.) (2006), *Measurement error in nonlinear models: A modern perspective* (2nd ed.), Bocan Raton, FL: CRC press.

Di Consiglio, L., and Tuoto, T. (2014). "When adjusting for bias due to linkage errors: A sensitivity analysis". Paper read at European Conference on Quality in Official Statistics (Q2014), 3-5 June, at Vienna, Austria.

Edwards, S. L., Berzofsky, M. E., and Biemer, P. (2017), "Effect of missing data on classification error in panel surveys", *Journal of Official Statistics,* 33, 551-570.

Fellegi, I. P., and Sunter, A. B. (1969), "A theory for record linkage", *Journal of the American Statistical Association,* 64, 1183-1210.

Fuller, W. A. (ed.) (1987), *Measurement error models*, Hoboken, NJ: John Wiley & Sons.

Georgiadis, M. P., Johnson, W. O., Gardner, I. A., and Singh, R. (2003), "Correlation-adjusted estimation of sensitivity and specificity of two diagnostic tests", *Journal of the Royal Statistical Society: Series C (Applied Statistics),* 52, 63-76.

Hagenaars, J. A. (1988), "Latent structure models with direct effects between indicators: Local dependence models", *Sociological Methods & Research,* 16, 379-405.

--------- (ed.) (1990), *Categorical longitudinal data: Log-linear panel, trend, and cohort analysis*, Newbury Park, CA: Sage Publications.

Kuha, J., and Skinner, C. (1997). "Categorical data analysis and misclassification, " In *Survey measurement and process quality*, eds. L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz and D. Trewin, Hoboken, NJ: John Wiley & Sons.

Lahiri, P., and Larsen, M. D. (2005), "Regression analysis with linked data", *Journal of the American Statistical Association,* 100, 222-230.

Manzoni, A., Vermunt, J. K., Luijkx, R., and Muffels, R. (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), "Beyond the number of classes: Separating substantive from non-substantive dependence in latent class analysis", *Advances in Data Analysis and Classification,* 10, 171-182.

--------- (2017). "Estimating error rates in an administrative register and survey questions using a latent class model, " In *Total survey error in practice,* eds. P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N. C. Tucker and B. T. West, Hoboken, NJ: John Wiley & Sons.

Oberski, D. L., Hagenaars, J. A., and Saris, W. E. (2015), "The latent class multitrait-multimethod model", *Psychological methods,* 20, 422-443.

Oberski, D. L., Kirchner, A., Eckman, S., and Kreuter, F. (2017), "Evaluating the quality of survey and administrative data with generalized multitrait-multimethod models", *Journal of the American Statistical Association*, just-accepted.

Pankowska, P., Bakker, B., Oberski, D. L., and Pavlopoulos, D. (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.

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

Qu, Y., and Hagdu, A. (2012), "Modeling correlations between diagnostic tests in efficacy studies", *Modelling Longitudinal and Spatially Correlated Data*, 122, 363.

Saris, W. E., and Gallhofer, I. N. (2007), *Design, evaluation, and analysis of questionnaires for survey research* (2nd ed.), *Wiley series in survey methodology*, Hoboken, NJ: John Wiley & Sons.

Scholtus, S., Bakker, B. F. M., and Delden, A. V. (2015) "Modelling measurement error to estimate bias in administrative and survey variables": CBS Discussion Paper.

Torrance-Rynard, V. L., and Walter, S. D. (1997), "Effects of dependent errors in the assessment of diagnostic test performance", *Statistics in medicine*, 16, 2157-2175.

Vacek, P. M. (1985), "The effect of conditional dependence on the evaluation of diagnostic tests", *Biometrics*, 959-968.

Vermunt, J. K., and Magidson, J. (2002), "Latent class cluster analysis", *Applied latent class analysis*, 11, 89-106.

Winglee, M., Valliant, R., and Scheuren, F. (2005), "A case study in record linkage", *Survey Methodology*, 31, 3-11.

Winkler, W. E. (1999) "The state of record linkage and current research problems": Statistical Research Division, US Census Bureau.

Appendix 1. Linkage quality

There are several categories of data sources that are difficult to link. The first category is data sources for which it is not allowed by law to store all the identifying variables that could be used as linking variables. This is certainly the case for data sources that comprises sensitive information, like medical information or criminality. Examples of that are bio-banks on genetic information, cancer registrations or police records on criminal suspects. This leads to a situation that the available identifying information is limited, sometimes too limited to use for linkage purposes. Privacy laws differ between countries, but what they have in common that they aim to prevent disclosure of sensitive information.

The second category of data sources that is difficult to link are data sources in which the register keeper or the registered has some interest for not being registered correctly or completely. Examples of such registers are again police records on criminal suspects, in particular if the suspect cannot be identified using a personal identification document. In countries in which students get higher student loans if they do not live with their parents, registrations of the addresses of students are of bad quality. Schools that are funded on the basis of the number of pupils have an interest in keeping pupils in their registers after moving house of the pupils. These pupils usually are still registered on the former addresses.

If address is one of the linkage variables, there are some categories that are difficult to link. The first category is those who move house frequently. This is associated with age (the young adults move house frequently as do the very old), and with life events like getting children (in particular second and third children), or getting a job far away from the place that you reside (associated with education level).

Appendix 2. Record linkage procedure for the combined LFS and ER dataset

The data from both sources are linked at the individual level to the PR. For the LFS, the linkage key is the combination of birth date, gender, postal code, and house number. In the first step, two records are linked if the post code and house number correspond and only one of the other variables of the linkage key differ. In the second step, the remaining, unlinked records are linked on postal code, birth date and gender and no differences on the other variables are allowed. This results in a linkage effectiveness, i.e. the percentage of linked records, of 98.3% for those who had a first interview in 2009.

The ER is linked to the PR in four steps; the procedure is repeated monthly and one-to-one matching is enforced. In the first step, the records are linked on the Citizen Service Number (BSN; a unique personal number allocated to everyone registered in the Netherlands). For those records that are linked in this step, it is verified whether birth date and gender are similar. If not, the records go to the next step together with the records that were not linked on BSN. In the second step, the data are linked using birth date, gender, postal code and house number. In the third step, the remaining records form the first two steps are linked using only the BSN and ignoring differences in the other variables. This procedure is repeated monthly. The linkage effectiveness is approximately 96-97% depending on the chosen month. In the first step already approximately 99,8% of the linked records of all steps are successfully linked.

The linkage to the Population Register results in the assignment of a meaningless linkage number to each linked record of both sources. That linkage number can be used to combine the LFS and ER as well as the data from the successive follow ups. After selection of the persons aged 25-55, the linkage effectiveness of the combined sources is approximately 97%. The records that do not link refer to cross-border workers from Belgium, Germany that belong to the target population of the ER but not the LFS as well as to non-

registered individuals (typically immigrants) that are represented in the LFS but not in the ER. Therefore, when focusing on the population of registered individuals that reside in the Netherlands the linkage of the two data sources of our dataset approaches perfection.