How linkage error affects hidden Markov models[1]: a sensitivity analysis

Abstract: *Latent class models (LCM) are 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 (HMM)" – 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 data sources can solve this problem by making the local independence assumption plausible across sources, while potentially allowing for local dependence within sources. However, record linkage introduces a new problem: the records may be erroneously linked.*

In this paper we investigate the effects of linkage error on HMM estimates of employment contract types. Our data come from linking a labor force survey to administrative employer records; this linkage yields two indicators per time point that are plausibly conditionally independent. Our results indicate that false-negative linkage error (exclusion) turns out to be problematic only if it is large and highly correlated with the dependent variable. Moreover, under many conditions, false-positive linkage error (mislinkage) turns out to act as another source of misclassification that the HMM can absorb into the error-rate estimates, leaving the latent transition estimates unbiased. In these cases, measurement error modeling already accounts for linkage error. Our results also indicate where these conditions break down and more complex methods would be needed.

Keywords: *linkage error; misclassification; classification error; measurement error; latent class model (LCM); hidden Markov model (HMM)*

Word count: 5,948 (excluding tables, figures, references and appendices)

[1] This work was supported by Statistics Netherlands and Vrije Universiteit Amsterdam; the authors would like to thank Peter Paul de Wolf (Statistics Netherlands), Richard Price (University of Notre Dame) and the members of the SILC research group of the Vrije Universiteit Amsterdam for reviewing the paper and providing valuable and constructive feedback as well as Aleksandra Daniłoś for assisting with the design of the figures.

1. INTRODUCTION

Survey data, despite 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 between 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 - are latent class models (LCMs) (Biemer 2011, Vermunt and Magidson 2002).

LCMs 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, or else the "measurement parameters". These models also provide estimates of the "structural parameters", which measure quantities of scientific interest, such as prevalence of certain groups in the population 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". When the repeated indicators are repetitions of the same question at different time points, a particular variant of an LCM is used: the "hidden" (or "latent") Markov model (HMM) (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 LCMs does not come cheap: a payment in untestable assumptions is required (see e.g. Oberski, Hagenaars, and Saris 2015), and

particularly the "local independence" assumption, which requires that the errors in the repeated indicators occur independently.

This local independence assumption is unrealistic, harmful and, when only one indicator is available, also 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 is likely that any personal "style" in answering a survey question carries over time, as shown, for instance, by Billiet and Davidov (2008). It is harmful because ignoring it leads to bias in classification error estimates (Georgiadis et al. 2003, Qu and Hagdu 2012, Torrance-Rynard and Walter 1997, Vacek 1985). Finally, it is undetectable with data from a single repeated indicator, because the local independence assumption is necessary for model identification in this case. 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 that are 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 as well (Bakker 2012, Oberski 2017, Oberski et al. 2017, Scholtus, Bakker, and Delden 2015). Fortunately, we do not require the linked registers to be error-free, but only that register errors are independent of survey errors. With linked survey-register data, this indeed seems plausible. This means that, by combining registers and surveys, it becomes possible to allow for local dependence within each source. Previous studies have done so, and found indeed considerable local dependence (Bassi et al. 2000, Oberski et al. 2017, Pavlopoulos and Vermunt 2015), confirming both the importance of relaxing this assumption and the attractiveness of data linkage.

However, while linkage allows us to tackle the problems of measurement error modelling, it also introduces a new challenge: linkage error. Again, such errors are known to bias estimates of interest when left unaccounted for (Harron et al. 2017). Several estimators correcting for linkage errors have been suggested (Lahiri and Larsen, 2005, Chambers 2009, Liseo and Tancredi 2011, Goldstein et al. 2012). Some of these estimators assume knowledge, for each case, of the posterior probability of correct linkage for all other cases. This knowledge is unavailable to most analysts in practice. The remaining solutions, thanks to Chambers (2009), do not assume this knowledge, but have only been developed for linear regression models. Existing methods can therefore not be used to correct HMMs for linkage error. At the same time, such errors may introduce bias into HMM estimates, and little is known about the extent of these biases.

In this paper, we study the extent to which linkage error biases HMM parameter estimates. Through a simulation study based on a real data application to linked survey-register employment records at Statistics Netherlands, we demonstrate the sensitivity of the structural (prevalence and transition rate) and measurement (classification error) parameters of the model to linkage error. We find that in certain situations, the HMM can absorb the error into its measurement model, leading to approximately unbiased structural parameter estimates. In other situations, biases in both the measurement and structural parts of the model can occur. A novel geometric representation of the latent class estimation problem demonstrates why this is the case.

Section 2 introduces linkage error and the Hidden Markov Model with multiple indicators for linked data. Section 3 presents the data and Section 4 the methodology; in section 5 we discuss the results of our analysis. Finally, Section 6 concludes.

2. BACKGROUND

2.1 Record linkage and linkage error

Record linkage is a process that matches records and attempts to select those matches that belong to the same person or unit. The process uses one or more data fields (i.e. linkage variables) that contain the same identifying information in all sources (Winkler 1999, Armstrong and Mayda 1993).

There are two main types of record linkage methods- deterministic and probabilistic. Deterministic record linkage defines pairs as true matches if the matching variables agree exactly in all data sources. It usually relies on a relatively small number of matching variables and is most commonly applied in the presence of the same unique identifier in all data sources (Blakely and Salmond 2002). Over the last decades, data sources have been increasingly lacking high-quality unique identifiers and, therefore, deterministic linkage has been gradually replaced by probabilistic linkage (Ariel et al. 2014).

Probabilistic record linkage tends to use 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. Probabilistic linkage 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. In general, linkage errors occur: (I) when due to missing or inaccurate data, some records that correspond to the same person or unit are not linked- a phenomenon referred to a false-negative linkage error, (II) when as a result of coding or measurement errors, unrelated

records are wrongfully linked – a situation referred to as false-positive linkage error (Winglee, Valliant, and Scheuren 2005, Bohensky et al. 2010).

Record linkage and linkage errors can be formulated using files drawn from two populations- file A containing N_A records and file B containing N_B records, and a set C containing record pairs which are the cross-product of files *A* and *B*. This set is denoted by $C = \{(a, b); a \in A, b \in B\}$ and the number of records equals to $N = N_A \times N_B$ (Armstrong and Mayda 1993, Sadinle et al. 2011).

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, Sadinle et al. 2011).

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 false-negative by $P(A_3|M)$ (Armstrong and Mayda 1993, Sadinle et al. 2011).

There are several approaches and frameworks available in the literature to correct for the effects of linkage error. Three prominent approaches are those proposed by Lahiri and Larsen (2005), Chambers (2009), and Liseo and Tancredi (2011). Lahiri and Larsen (2005) propose an M- and U probabilities-weighted linear regression model for linked data, which takes into account linkage uncertainty. However, their method relies on the, often unrealistic, assumption that the linkage/ mislinkage probabilities of all pairs of records are known. Liseo

and Tancredi (2011) propose a Bayesian approach to linkage problems, in which the analysis models and linkage models are subsumed into a single latent variable model estimated via MCMC. A similar approach, implementing Bayesian imputation conditioned on the linkage probabilities, was suggested independently by Goldstein et al. (2012). While the Bayesian approach is in principle comprehensive, it shares with the previous approach the drawback that full knowledge of the linkage process is required by the analyst, a situation that often does not occur in practice for privacy reasons. Finally, Chambers (2009) and Kim and Chambers (2012a, 2012b) introduce a bias-corrected ratio estimator, as well as a class of weighted estimators for linear regression and logistic regression. Moreover, Chambers (2009) suggests replacing the required assumption of perfect information regarding the linkage/mislinkage probabilities with a more practicable approximation based on available aggregate linkage rates. As detailed in Chambers and Kim (2016), since the weighting approach is based on estimating equations, it can in principle be extended to other, more complex, classes of models beyond linear and logistic regression. However, Chambers-type estimators for HMMs are currently not available.

To sum up, available methods to account for linkage error are difficult to implement for HMMs for practical or technical reasons. Therefore, in this paper we do not attempt to introduce a method to correct for such errors in HMMs but focus on the sensitivity, or lack of it, of such models to linkage errors.

2.2 Hidden Markov models: measurement error and linkage error

Hidden Markov Models (HMMs) are a group of latent class models increasingly used to estimate and correct for measurement error in longitudinal categorical data (Biemer 2004, 2011). In this section, we first present the basic single-indicator HMM, commonly applied across the literature; we then extend it by including an additional indicator per time point.

The basic HMM operates under the assumption that, at each time point $t \in \{1, \ldots, T\}$, the observed answer Y_t, assumed to follow a multinomial distribution, is generated *independently* with some probability $P(Y_t|X_t)$ from the true, but unobserved, multinomially distributed variable X_t. Because the generation of Y_t is assumed independent of all other variables, the *T*-dimensional distribution $P(Y|X)$ of observed path *Y* given latent path *X*, factorizes into a product:

$$P(Y|X) = \prod_{t=1}^{T} P(Y_t|X_t) \quad (1)$$

This assumption is known as the "local independence" or "independent categorization error" (ICE) assumption. The latent path *X*, meanwhile, is assumed to follow a Markov or "AR(1)" process,

$$P(X) = \; P(X_0) \prod_{t=1}^{T} P(X_t|\, X_{t-1}) \quad (2)$$

Finally, the observed data distribution $P(Y)$ is assumed to arise by combining the ICE and hidden Markov assumptions above and marginalizing over X, yielding the marginal likelihood

$$P_{\text{HMM}}(Y) = \sum_{X} P(Y|X)P(X) \quad (3)$$

with "structural" parameters $P(X_0)$ and $P(X_T|\, X_{T-1})$ – the initial state and transition probabilities – and "measurement parameters" $P(Y_t|X_t)$ – the probabilities of correct and incorrect classification.

When consistent estimates of $P_{\text{HMM}}(Y)$ are observed (i.e. when Y is ergodic), consistent maximum-likelihood estimates can be obtained by maximizing Equation (3) over the structural and measurement parameters (Leroux 1992). In practice, instead of the exponentially complex summation over all possible latent paths *X* in Equation (3), the more

computationally efficient "forwards-backwards" (Baum-Welch) algorithm is used. This amounts to an adapted expectation-maximization (EM) procedure (McLachlan and Krishnan 2008, pp. 291-2). In the E-step, the posterior P(X|Y) is computed by combining two recursive computational steps that each consider only a single time point at a time together with the result of their respective previous computations. In the M-step the model's parameters are then computed by summation over states at each time point, weighted by the posterior. Thus, the computational complexity of one Baum-Welch iteration is linear in the number of time points, rather than exponential, as when using the marginal likelihood (3). The E- and M-steps are iterated until convergence.

In survey research, the single-indicator HMM is attractive for two reasons. First, and in contrast with standard latent class analysis, it allows for hidden change over time in the true values, $P(X_t|X_{t-1})$, while simultaneously estimating and accounting for classification errors, $P(Y_t \neq x_t|X_t = x_t)$. Second, its parameters can be identified from panel data on single repeated indicators with three or more waves, which are often already collected as part of longitudinal surveys or recorded in administrative databases. This identifiability follows from the model's assumptions, specifically the Markov and conditional independence (ICE) assumptions.

However, conditional independence may in practice be an unrealistic assumption. For example, several studies (e.g. Saris and Gallhofer 2007) have shown that survey respondents have answer tendencies that may persist over time. In administrative databases, records may simply get copied, leading to copying of errors as well (Oberski et al. 2017). To model such error dependencies and simultaneously estimate classification error in both survey and administrative data that measure the same phenomena, Pavlopoulos and Vermunt (2015) suggested to link respondents' survey answers to administrative records. Such linked survey-

administrative data then allow for the relaxation of the ICE assumption, replacing Equation (1) with

$$P(Y|X) = P\left(Y_{\text{survey}}\middle|X\right)P(Y_{\text{admin}}|X) \tag{4}$$

where Y now collects the observed processes for both survey and administrative data. Pavlopoulos and Vermunt (2015) suggest to further specify the conditional dependence as

$$P(Y|X) = \prod_{t=1}^{T} P\left(Y_{t,\text{survey}}\middle|X_t\right) \prod_{t=1}^{T} P\left(Y_{t,\text{admin}}\middle|X_t, X_{t-1}, Y_{t-1,\text{admin}}\right) \tag{5}$$

where $P\left(Y_{t,\text{admin}}\middle|X_t, X_{t-1}, Y_{t-1,\text{admin}}\right)$ is restricted to allow for error copying using a logit formulation. In other words, this model allows for error dependence in the administrative data, while assuming survey and administrative answers to be conditionally independent. Other error dependence structures are also identifiable, including for the survey data. The advantages of linkage are thus that (1) both survey and administrative errors can be modelled simultaneously, and (2) the ICE assumption can be relaxed in a rather flexible way.

The disadvantage of linkage, however, is that incorrect linkages may occur. The effects of linkage error in general circumstances have been considered extensively in the work of Chambers and others, discussed above (Chambers and Kim 2016). This work has shown that linkage error can cause bias in analyses of dependencies such as linear and logistic regression. It therefore seems plausible that bias would also occur in a multivariate method such as HMM, which uses dependencies to estimate its parameters. However, no work to date has examined the precise effects of linkage error for multivariate analysis. This paper does not aim to examine these effects theoretically or solve the problem of linkage error for HMMs. We do, however, note that linkage error can be expected to strongly violate HMM assumptions and cause bias only under certain circumstances. Here we provide an intuitive explanation of this phenomenon. Appendix A.1 provides a geometric representation

of the latent class model (LCM) estimation problem that demonstrates why independent, random, linkage errors will largely be absorbed into the measurement part of such models.

"False-negative" linkage error manifests itself as missing data, and a large literature on the effects of various missingness mechanisms on ML estimates already exists (see e.g. Little and Rubin 1989, Hagenaars and McCutcheon 2002). "False-positive" linkage errors (mislinkages), however, have an entirely different, as yet unstudied, effect on HMMs. We will therefore concentrate on the effect of mislinkages here. The simulation study, however, investigates the effects of both types of errors.

Following Lahiri and Larsen (2005), mislinkage manifests as an additional latent class variable with two categories corresponding to true matches (M) and non-matches (U). Within the class of matches, the HMM holds, while within the class of non-matches an unknown process holds. Lahiri and Larsen (2005) assume non-matches follow a distribution in which all *J* observed variables are independent. The observed data distribution is then a mixture of the true dependence structure and "randomly shuffled" data:

$$P_{\text{linked}}(Y) = P(M)P_{\text{HMM}}(Y|\theta) + [1 - P(M)]\prod_{j=1}^{J} P(Y_j) \qquad (6)$$

where the HMM likelihood has been written $P_{\text{HMM}}(Y|\theta)$ to emphasize its dependence on the model parameters, θ, of interest. Clearly, when fitting the HMM to P_{linked}, asymptotic bias may, in principle, occur whenever there is mislinkage. Intuitively however, unless the mixture P_{linked} induces additional dependence beyond that found in P_{HMM}, its effect is to increase random measurement error in each Y_j. Since the HMM is intended to capture such errors and correct for them, one might expect that the increased error rates are reflected in the measurement part of the model which describes P(Y|X) but not necessarily in the structural model describing P(X). Appendix A.1 argues geometrically that, when the linkage error is independent of Y, this intuition will hold approximately. In particular, we show that the

maximum likelihood solution for the "structural" parameter indicating class size, π, is approximately unaffected by independent linkage error. In the following sections, a simulation study investigates the extent to which this result holds in an HMM.

3. DATA

The dataset used in our analysis contains data from the Dutch Labor Force Survey (LFS) and the Employment Register (ER), which have been linked using each citizen's unique identification number. In this paper, we assume this process does not involve linkage error, and simulate the effect of linkage error by artificially introducing false negative and false positive linkages into this dataset.

Our sample consists of 15 months' observations on 8,886 LFS respondents aged 25 to 55 who first participated in the survey in 2009. This results in a total sample size of 133,290 observations. The employment register is observed on a monthly basis while the LFS is taken every three months and consists of five waves. The main variable of interest in our analysis is an individual's employment contract type for their primary job: {"permanent contract", "temporary contract", "other"}.

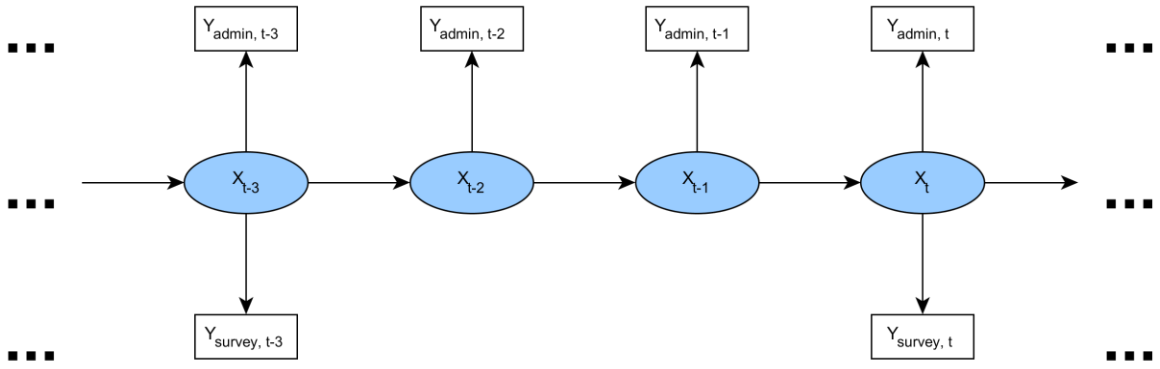

Figure 1. Hidden Markov Model graph. Rectangles are observed variables, while ovals are latent "true" variables. Absence of arrows indicates conditional independence.

4. METHODOLOGY

4.1. Model

Our approach consists of a simulation analysis in which we make use of a two-indicator HMM, where one of the indicators is the individual's contract type according to the LFS and

the second is the contract type according to the ER. While the model could be extended further following Pavlopoulos and Vermunt (2015) and Pankowska et al. (2017), our simulations are based on a simplified model which retains the local independence assumption (see Figure 1), specifying Equation (4) as

$$P(Y_t|X_t) = P\left(Y_{\text{survey,t}}\middle|X_t\right)P\left(Y_{\text{admin,t}}\middle|X_t\right) \tag{7}$$

Figure 1 illustrates our model as a graph. Because the survey has been administered once every quarter, while monthly measures are available from the administrative database, the survey is missing at timepoints *t - 1* and *t – 2*. Estimation with missing data proceeds using full-information maximum likelihood (Vermunt and Magidson 2013).

We apply the model to different conditions in which various types of either false-negative or false-positive linkage errors are introduced into the original dataset. A summary of the simulation setup is provided as a tree graph in Figure 2.

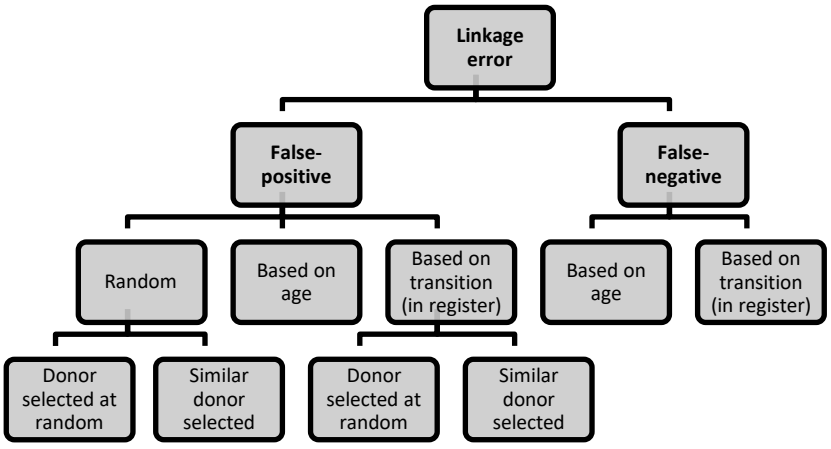

Figure 2. Conditions of the simulation study.

We consider conditions in which individuals are either randomly selected to be mislinked and/or excluded versus conditions in which the probabilities of linkage error depend on covariates mildly or strongly correlated with the model estimates. We also consider different error rates. Our setup allows for the investigation of the biasing effects of the error under varying degrees of severity. Each condition is replicated 200 times. We investigate the bias introduced by the error by comparing the obtained transition rates from temporary to

permanent employment to the transition rates estimated using the original linked dataset. To simulate linkage error, we use the R version 3.2.3. The HMM is estimated using Latent GOLD version 4.5. Our code is available as an online appendix to this paper.

4.2. False-negative error simulations

When investigating the effect of false-negative linkage error on the accuracy of our model estimates, we consider two conditions in which the individuals' probabilities of exclusion are correlated with (I) age[2] and (II) the presence of a (3-monthly) transition from temporary to permanent employment in the register data[3]. A condition in which the missingness is MCAR (missing completely at random) has been omitted. Within each condition, we simulate three sub-conditions in which we introduce high (20%), medium (10%) and low (5%) overall exclusion error into our data; this error is equal to the proportion of correctly linked individuals in the data that are erroneously excluded from it.

For the age-dependent conditions, the correlations are such that the exclusion probabilities of younger individuals are higher than those of older individuals; for the transition-dependent conditions, the probabilities of those individuals who transitioned according to the register data are higher than those of the individuals who did not. These specifications are motivated by the fact that both young individuals and those who transitioned would tend to have higher residential and employment mobility and are thus more susceptible to linkage error.

To assure that the conditions indeed represent varying levels of severity, the simulation is also designed in such a way that as we move from conditions with lower levels of exclusion error to conditions with higher ones, the oversampling of young individuals or

[2] Pankowska et al. (2017) in their analysis of the same data used an extended version of the HMM we use in this paper. Their model, among other things, accounted for effect of age on the latent transitions probabilities. Their results showed that age has a moderate, negative effect on the probability of transitioning from temporary to permanent employment (logit coefficient= -0.3 over the range of the covariate)

[3] According to our model, over 99% of all contracts observed in ER are correctly classified and, therefore, the correlation between the transition covariate we have created and the model estimates is highly reliable.

those who transitioned becomes more extreme (i.e. their individual exclusion probabilities increase). To illustrate, the exclusion probability of young individuals (aged 25 to 34) is set to 0.15, 0.30 and 0.70 when the overall exclusion rate is low (5%), medium (10%) and high (20%), respectively; the exclusion probability of older individuals (aged 35 to 54) equals to 0.01 in all three cases.

Thus, a higher level of false-negative linkage error not only indicates that a larger proportion of individuals is excluded from the sample, but it also implies that the remaining sample is less representative of the overall population in terms of characteristics that are correlated with the transition rates estimated by the model. As those covariates are not controlled for when estimating the HMM, these simulated datasets are equivalent to a dataset containing data missing not at random (MNAR).

Overall, the simulations consist of three steps. First, the exclusion rate and the individual exclusion probabilities are set; then individuals are excluded from the sample with a probability that equals the condition's exclusion probability. Finally, the HMM is fitted to the resultant subsample and the estimates are compared to those obtained when using the full sample. As an illustration, Appendix A.3.1 provides pseudocode generating one condition.

4.3. False-positive error simulations

The analysis of the false-positive linkage error, similarly to that of the false-negative, also follows 3 steps. Note that here, unlike in the false-negative example (whereby individuals are merely excluded from the sample), a proportion of the sample is mislinked with another set of individuals. This adds a further complication to the simulation design, as a donor is required whose ER contract type can be (erroneously) linked to a given individual's LFS contract information. As in the false-negative error conditions, the first step determines the overall level of mislinkage (5%, 10% or 20%) and the individual probabilities of an erroneous link (which are either assigned at random or are age- or transition- dependent).

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 (i) randomly matched to another person or (ii) matched to a similar person based on age, gender, education level and ethnicity. The register values of individual A for the contract type are replaced with those of the matched individual (i.e. the donor), here referred to as individual B.

The second set of conditions, whereby relatively similar individuals are matched, is introduced to approximate a more realistic linkage error condition that is more representative of actual potential mismatches.

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 compared to the results obtained when using the original dataset. Pseudocode illustrating the simulation setup for one of the conditions is included in Appendix A.3.2.

5. RESULTS

5.1. The effect of false-negative error

The simulation results obtained for the various false-negative error conditions are shown in Table 1 and Figure 3. Table 1 presents the mean estimated 3-monthly transition rates as well as the absolute and relative bias introduced by linkage error. These biases are estimated by comparing the obtained transition rates to those calculated using the original dataset. Figure 3 provides an illustration of the relationship between the type (age or transition dependent), level (5%, 10%, 20%) and bias introduced by linkage error.

The results show that when the exclusion probability depends on age, the relative bias introduced by false-negative linkage error does not exceed 5% and, therefore, can be considered negligible. Thus, it appears that when the exclusion probability depends on a

covariate that is weakly or moderately correlated with the model estimates, the bias in the model estimates is marginal, even when the overall exclusion rate is rather high (e.g. 20%).

A largely different picture emerges when the exclusion probability depends on whether a transition occurred. Namely, our results show that the employment transition rates in this set of conditions are heavily underestimated leading to a substantial, non-negligible bias. In relative terms, the bias ranges from 10.6%, for an overall linkage error of 5%, to 25%, when the linkage error amounts to 10%, and to as high as 84.3% when the error rate equals 20%. As this covariate is highly correlated with the model estimates, we can infer from these results that conditions characterized by substantial dependency between the error and model outcomes will result in non-ignorable bias.

Overall, the results obtained suggest that the extended, two-indicator HHM is mainly sensitive to false-negative linkage error in rather extreme situations. That is, we only obtain non-ignorable bias in our (structural) model estimates when the individual-level exclusion probabilities depend on a covariate that is very highly correlated with the latent variable and consequently the model outcomes. In other, less extreme situations, the bias is relatively small and thus the HMM estimates can be considered accurate.

Finally, it is worthwhile noting that our false-negative linkage error analysis can be viewed as a form of complete case analysis with varying degrees of missingness; our two specific sets of conditions mimic MNAR: first, where the exclusion probabilities are dependent on a variable which is moderately correlated with the model estimates; and second, where the probabilities are dependent on a variable exceptionally highly correlated with the model estimates. The former represents a more realistic scenario whereby the latter, while could occur in theory, is considered rather extreme. Our findings confirm this line of thought. More specifically, our results, similarly to the ones reported by studies investigating missingness specifically, show that MNAR leads to substantial bias when the missingness is

highly correlated with model estimates (Bakker and Daas 2012; Galimard et al. 2016; Marshall et al. 2010).

Table 1. Simulation results- false-negative linkage error

Error type	Condition: 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	0.01	0.066	0.003	4.6%
		0.10	0.30	0.01	0.067	0.002	3.2%
		0.20	0.70	0.01	0.066	0.003	3.8%
	Depends on transition	0.05	0.15	0.05	0.062	0.007	10.6%
		0.10	0.34	0.09	0.052	0.017	25.0%
		0.20	0.90	0.17	0.011	0.058	84.3%

[1] In the age- dependent conditions, high exclusion probability was set for young individuals and low for older ones; in the transition-dependent conditions, high exclusion probability was set for individuals who had a transition and low for those who did not.
[2] The transition rates are estimated based on the modal class memberships (i.e. at each time point individuals are assigned the contract type to which they have the highest posterior probability of belonging according to the model); as the entropy R2 is above 0.99 for all conditions, such an assignment is not expected to produce different results from an assignment which takes the uncertainty of class memberships into account.

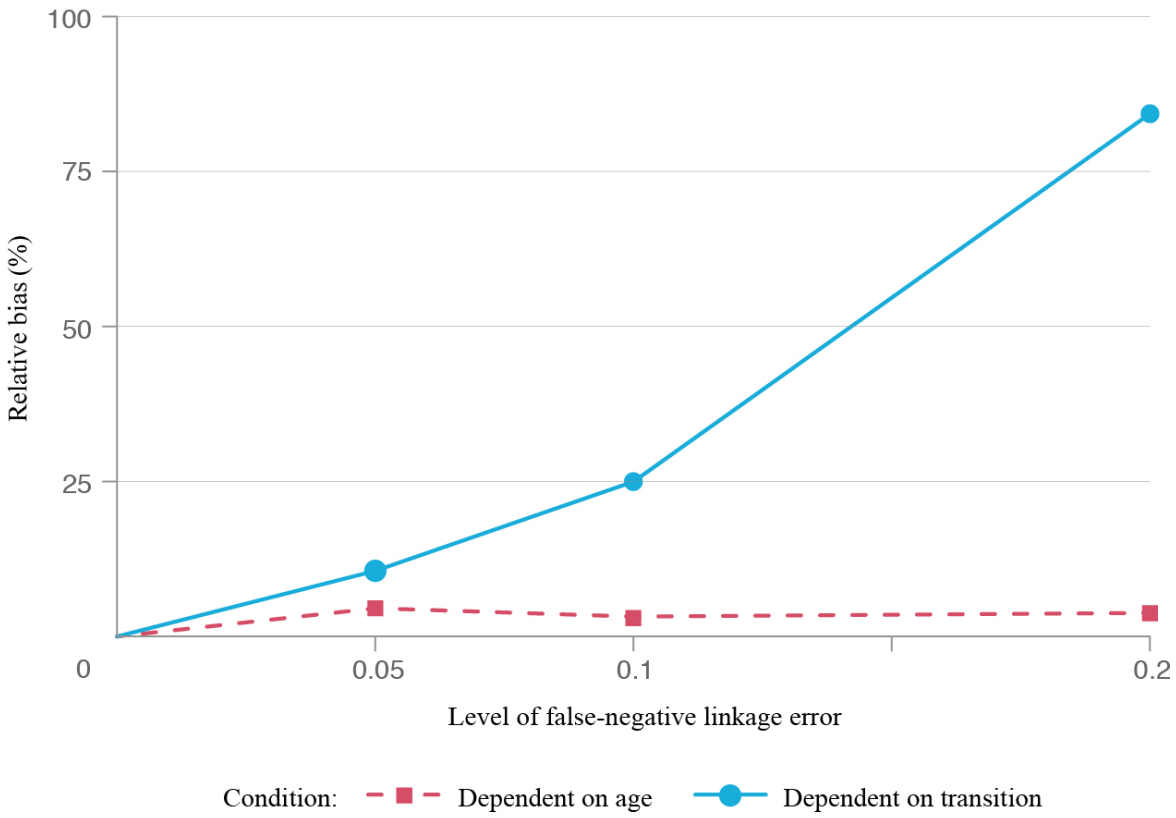

Figure 3. Relative bias by overall level of false-negative linkage error.

5.2. The effect of false-positive error

The results obtained when simulating various levels and types of false-positive linkage error are presented in Table 2 and Figure 4. As can be seen, the bias introduced by false-positive linkage error is rather modest for the conditions where the mislinkage probability is either

random or depends on age. In contrast, those conditions in which the probability of mislinkage depends on whether a transition occurred are characterized by high, non-negligible bias. These findings are consistent for both the conditions in which an individual is mislinked with a randomly selected donor and where the individual is mislinked with a donor similar to them with regards to age, gender, education and ethnicity.

More specifically, the first two sets of conditions, regardless of whether the individual is mislinked with a random or a similar donor, lead to a relative bias of less than 5%. On the other hand, those conditions in which the mislinkage probability depends on the presence of a transition result in a relative bias of around 10%, 20-25% and (well) over 60% when the mislinkage rate is low, medium and high respectively. Figure 4 shows a clear positive relationship between the transition- dependent mislinkage level and the bias in the model estimates. This relationship is not observed for the other two sets of conditions.

Table 2. Simulation results- false-positive linkage error

Error type	Condition: 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	0.0%
		0.10	-	-	0.069	0.000	0.3%
		0.20	-	-	0.068	0.001	0.9%
	Depends on age	0.05	0.15	0.01	0.069	0.000	0.3%
		0.10	0.30	0.01	0.068	0.001	1.2%
		0.20	0.70	0.01	0.067	0.002	2.6%
	Depends on transition	0.05	0.15	0.05	0.064	0.005	7.8%
		0.10	0.34	0.09	0.055	0.014	20.7%
		0.20	0.90	0.17	0.024	0.045	64.6%
False-positive; mislinkage with similar donor	Random	0.05	-	-	0.067	0.002	3.1%
		0.10	-	-	0.067	0.002	3.2%
		0.20	-	-	0.066	0.003	4.9%
	Depends on transition	0.05	0.15	0.05	0.061	0.008	11.5%
		0.10	0.34	0.09	0.051	0.018	26.6%
		0.20	0.90	0.17	0.012	0.057	82.6%

[1] In the age- dependent conditions, high exclusion probability was set for young individuals and low for older ones; in the transition-dependent conditions, high exclusion probability was set for individuals who had a transition and low for those who did not.
[2] The results from the random and age- based mislinkage, when individuals are mislinked with random donors, were very similar and therefore when individuals were mislinked with similar donors the age- based set of conditions was omitted.
[3] The differences in the bias obtained when using random and similar donors might be due to the fact that the StatMatch R package used to match donors does not allow for missing values on the covariates and, thus, the analysis was run on a smaller sample.
[4] The transition rates are estimated based on the modal class memberships (i.e. at each time point individuals are assigned the contract type to which they have the highest posterior probability of belonging according to the model); as the entropy R2 is above 0.99 for all conditions, such an assignment is not expected to produce different results from an assignment which takes the uncertainty of class memberships into account.

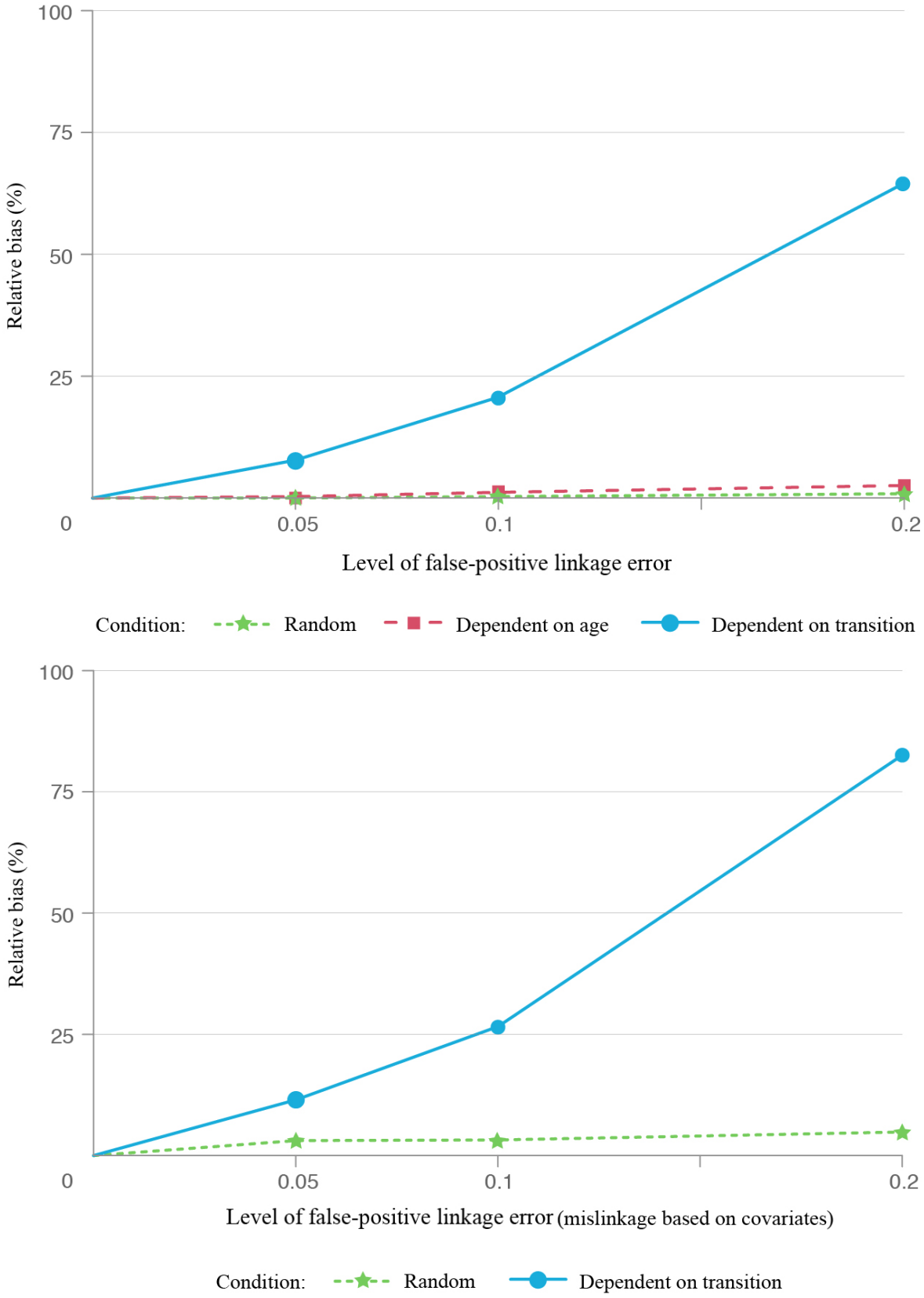

Figure 4. Relative bias by overall level of false-positive linkage error.

Figure 5 examines how mislinkage affects the measurement part of the model; i.e. it shows the effect of linkage error on the proportion of measurement error in our main variable

of interest- i.e. the individual's contract type. As can be seen, as we increase the mislinkage rate, the misclassification rate moves in tandem; this is particularly visible for the LFS data[4]. These results confirm our intuition and suggest that under many conditions false- positive linkage error is simply another source of misclassification that the HMM can absorb into the error rate estimates and corrects for in the transition rates estimates.

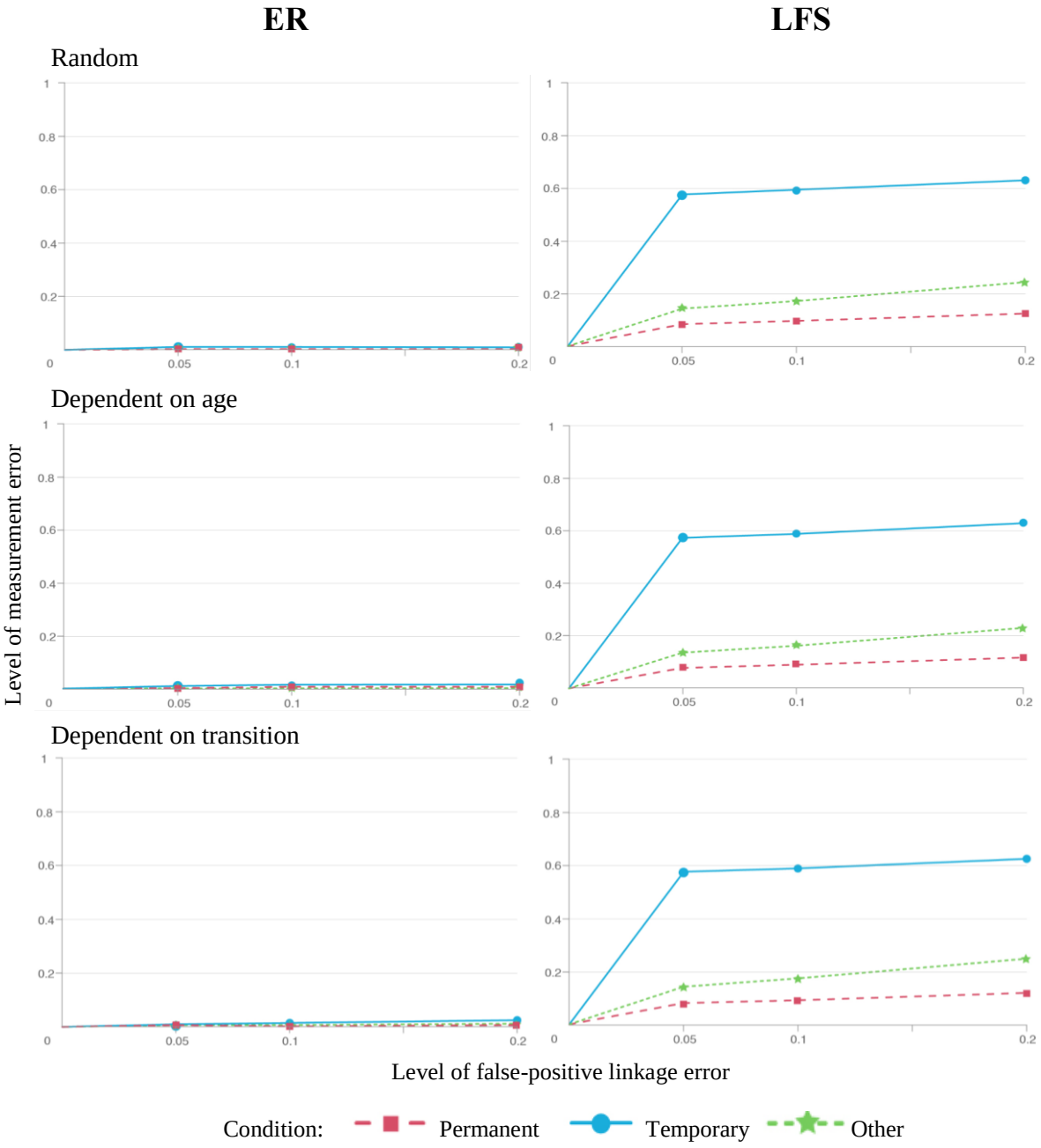

Figure 5. Level of measurement error by type and level of mislinkage.

[4] This pattern is not observed in the ER data as the simplified HMM we use does not account for autocorrelation of the error in these data. As measurement error in the ER is predominantly systematic, the model fails to capture it altogether and assumes the register data to be virtually error-free.

6. CONCLUSION AND DISCUSSION

Latent Class Models (LCMs) have been increasingly used to correct for measurement error in categorical variables. A particularly useful group of LCMs are hidden Markov models (HMMs), as they can be applied to longitudinal data, and thus allow studying transitions and change over time, a quantity which is often of interest in the social sciences. However, while HMMs are an appealing and useful tool, they rely on the (often unrealistic) local independence assumption. An attractive solution that allows the local independence assumption to be relaxed, is linking data from independent sources. Such record linkage identifies HMMs with local dependence within sources while maintaining the independence assumption across sources. However, this approach introduces a new challenge: linkage error.

We investigated the sensitivity of HMM estimates to linkage error. A geometric argument demonstrated that independent (false-positive) linkage error is largely absorbed by measurement parameters of latent class models. Dependent linkage errors, however, can be expected to strongly bias structural model parameters such as the class size in an LCM. Our simulation study further investigated this effect for HMMs based on an existing application to linked data on employment mobility.

Our results suggest that linkage error may not always be a problem for researchers who wish to apply Hidden Markov Models for the purpose of estimating its structural parameters, such as transition rates. When individuals are randomly mislinked or non-linked, the resulting bias in structural parameters was often negligible in our study, a result that confirms the geometric intuition relevant to LCMs. Linkage error led to strong bias only when the individual probability of being erroneously excluded or mislinked depended on the transition rate. The bias was particularly high for high rates of linkage error and when the aforementioned dependency was very strong. The sensitivity of estimates of structural parameters to mislinkage therefore appears relatively low.

Our results show that false-positive linkage error can often be absorbed by the model. In other words, mislinkage often manifests itself as random measurement error that is already corrected for by the model – unless the linkage error probability is strongly dependent. Despite this important caveat, we believe that our findings highlight the attractiveness of HMMs to correct for measurement error in structural parameter estimates, since they allow the use of linked data with relatively low sensitivity to linkage error. This is especially appealing as the methods available to correct for linkage error often cannot be easily applied in this context.

A disadvantage of our findings is that, since linkage error may be absorbed into measurement error parameters, these parameters no longer give "pure" estimates of measurement error. In other words, when the measurement, not structural, parameters are of primary interest (e.g. Biemer 2011), our results suggest linkage and measurement error will be partially conflated. Considering the increasing use of HMMs for this goal, future work should therefore develop methods to correct latent variable model estimates for linkage error, perhaps by extending the estimating equations approach discussed in Chambers and Kim (2016).

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*, 6, 212-239.

Ariel, A., B. Bakker, M. de Groot, G. van Grootheest, J. van der Laan, J. Smit and Verkerk, B. (2014), "Record linkage in health data: a simulation study," Statistics Netherlands Discussion Paper. Available at: https://www.cbs.nl/nl-nl/achtergrond/2014/16/record-linkage-in-health-data-a-simulation-study

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.

Bakker, B. F., and Daas, P. J. (2012), "Methodological challenges of register-based research," *Statistica Neerlandica*, 66, 2–7.

Bassi, F., Hagenaars, J. A., Croon, M. A., and Vermunt, J. K. (2000), "Estimating true changes when categorical panel data are affected by uncorrelated and correlated classification errors: an application to unemployment data," *Sociological Methods & Research*, 29, 230–268.

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. (eds.) (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* (Vol. 173), 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.

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

Chambers, R. (2009), "Regression analysis of probability-linked data," Official Statistics Research Series, 4, Statistics New Zealand. Available at: http://www3.stats.govt.nz/statisphere/Official_Statistics_Research_Series/Regression_Analysis_of_Probability-Linked_Data.pdf

Chambers, R., and Kim, G. (2016), "Secondary analysis of linked data," In *Methodological Developments in Data Linkage*, eds. K. Harron, H. Goldstein and C. Dibben, West Sussex: John Wiley & Sons, Ltd.

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.

Fienberg, S. E., and Gilbert, J. P. (1970), "The geometry of a two by two contingency table," *Journal of the American Statistical Association*, 65, 694–701.

Fuller, W. A. (ed.) (1987), *Measurement error models* (Vol. 204), Hoboken, NJ: John Wiley & Sons.

Galimard, J.-E., Chevret, S., Protopopescu, C., and Resche-Rigon, M. (2016), "A multiple imputation approach for MNAR mechanisms compatible with Heckman's model," *Statistics in medicine*, 35, 2907–2920.

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.

Goldstein, H., Harron, K., and Wade, A. (2012), "The analysis of record-linked data using multiple imputation with data value priors," *Statistics in medicine*, 31, 3481–3493.

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.

Harron, K. L., Doidge, J. C., Knight, H. E., Gilbert, R. E., Goldstein, H., Cromwell, D. A., and van der Meulen, J. H. (2017), "A guide to evaluating linkage quality for the analysis of linked data," *International Journal of Epidemiology*, 46, 1699–1710.

Jones, G., Johnson, W. O., Hanson, T. E., and Christensen, R. (2010), "Identifiability of models for multiple diagnostic testing in the absence of a gold standard," *Biometrics*, 66, 855–863.

Kim, G., and Chambers, R. (2012a), "Regression analysis under incomplete linkage," *Computational Statistics & Data Analysis*, 56, 2756–2770.

Kim, G., and Chambers, R. (2012b), "Regression analysis under probabilistic multi-linkage," *Statistica Neerlandica*, 66, 64–79.

Kuha, J., and Skinner, C. (1997), "Categorical data analysis and misclassification," In *Survey measurement and process quality* (Vol. 324), 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.

Leroux, B. G. (1992), "Maximum-likelihood estimation for hidden Markov models," *Stochastic processes and their applications*, 40, 127–143.

Liseo, B., and Tancredi, A. (2011), "Bayesian estimation of population size via linkage of multivariate Normal data sets," *Journal of Official Statistics*, 27, 491-505.

Little, R. J., and Rubin, D. B. (1989), "The analysis of social science data with missing values," *Sociological Methods & Research*, 18, 292–326.

Marshall, A., Altman, D. G., Royston, P., and Holder, R. L. (2010), "Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study," *BMC medical research methodology,* 10, 7.

McLachlan, G., and Krishnan, T. (2008), *The EM algorithm and extensions* (Vol. 382, 2nd ed.), Hoboken, NJ: John Wiley & Sons.

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*, 112, 1477-1489.

Pankowska, P., Bakker, B., Oberski, D. L., and Pavlopoulos, D. (2018), "Reconciliation of inconsistent data sources by correction for measurement error: The feasibility of parameter re-use," *Statistical Journal of the IAOS*, 34, 317-329.

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," In *Modelling Longitudinal and Spatially Correlated Data* (Vol. 122), eds. T. G. Gregoire, D. R. Brillinger, P. Diggle, E. Russek-Cohen, W. G. Warren and R. D. Wolfinger, Springer Science & Business Media.

Sadinle, M., Hall, R., and Fienberg, S. E. (2011), "Approaches to multiple record linkage,"In *Proceedings of International Statistical Institute*, 260, 1-20.

Saris, W. E., and Gallhofer, I. N. (2007), *Design, evaluation, and analysis of questionnaires for survey research* (2nd ed.), 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," Statistics Netherlands Discussion Paper. Available at: https://www.researchgate.net/profile/Sander_Scholtus/publication/283703551_Modelling_Measurement_Error_to_Estimate_Bias_in_Administrative_and_Survey_Variables/links/5643554f08ae451880a331fb.pdf

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," In *Applied latent class analysis*, eds. J. Hagenaars and A. McCutcheon, Cambridge: Cambridge University Press.

Vermunt, J. K., and Magidson, J. (2013), "Technical guide for Latent GOLD 5.0: Basic, advanced, and syntax," Belmont, MA: Statistical Innovations Inc.

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. Available at: https://www.census.gov/srd/papers/pdf/rr99-04.pdf

Appendix A.1. Fitting of a latent class model to data with independent linkage error- a geometric argument

Jones et al. (2010) adapted the geometric approach of Fienberg and Gilbert (1970) to the analysis of cross-tables, in order to depict maximum likelihood estimation of the measurement parameters and the structural parameter π in a three-indicator LCM. Here we demonstrate how these estimates are affected by independent linkage error. In the Fienberg and Gilbert (1970) approach, all possible normalized 2×2 cross-tables are placed in a tetrahedron representing the simplex $\{x \in R^4 : \sum x_i = 1\}$ (Figure A.1). The four corners of this tetrahedron, A_1, A_2, A_3 and A_4, correspond to cross-tables with all probability mass in a single cell; all other 2×2 cross-tables can be represented as a single point within the tetrahedron. An important subset of tables is the "independence surface" formed by all 2×2 independence tables, which is shown in Figure A.1 as the coloured surface. Points along a line on this surface correspond to all independence tables with constant row or column margins.

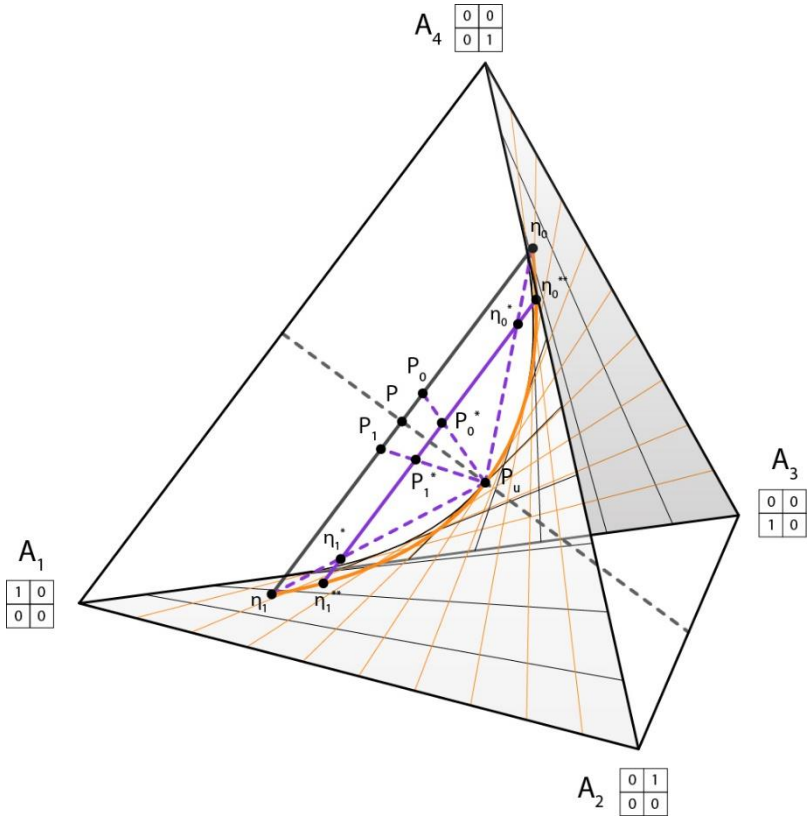

Figure A.1. Geometrical view of fitting of a latent class model to data with independent linkage error.

Following Jones et al. (2010), we consider a binary latent class model with three binary indicators Y_1, Y_2, and Y_3. Without loss of generality, we consider the bivariate cross-table of Y_1 and Y_2 given $Y_3 = 0$ (point p_0) $Y_3 = 1$ (point p_1). The MLE of η_0 and η_1 is then found at the two intersections of the "solution line" $p_1 - p_0$ with the independence surface. This follows from the fact that p_0 and p_1 are both convex combinations of η_0 and η_1, which, by conditional independence given the latent class variable, must lie on the independence surface. The MLE of $\mathrm{P}(\mathrm{X}|Y_3 = 0)$ is then found as $1 - \text{length}(p_0 - \eta_0)/\text{length}(\eta_1 - \eta_0)$ and similarly, $\widehat{\mathrm{P}}(\mathrm{X}|Y_3 = 1) = 1 - \text{length}(p_1 - \eta_1)/\text{length}(\eta_1 - \eta_0)$, implying the MLE for π can be found by applying Bayes' rule (Jones et al. 2010). Note that the length of the line segment $\eta_1 - \eta_0$ indicates the overall accuracy; as η_0 and η_1 differ more, the accuracy increases, with its maximum attained at the corners of the tetrahedron (sensitivity and specificity equal to one).

We now consider how the MLEs are affected by independent linkage error. When linkage error is independent, $\mathrm{P}(\mathrm{Y}|\mathrm{U}) = \mathrm{P}(\mathrm{Y})$, the $P(Y_j)$ in the equation above reduce to the marginals under the model, $P(Y_j) = \sum_{y_{k \neq j}} P_{\mathrm{HMM}}(Y)$. This point, p_U in Figure A.1, can be found by projecting the marginal over Y_3, point p, onto the independence surface along the line perpendicular to A_1A_2 and A_2A_3 (Fienberg and Gilbert 1970, p. 699). The linkage error model adapted from Lahiri & Larsen (2005) in Equation (6) then shows that the joint distribution under linkage error is a convex combination of p_U and the original joint distribution. That is, under independent linkage error, p_0 and p_1 are "shrunk" towards p_U by exactly $P(U)$. Therefore, when linkage error is independent, the observed data points p_0^* and p_1^* lie on a solution line parallel to the original solution line, with length $(p^* - p)/\text{length}\ (p - p_U) = P(U)$.

Similarly, the "true" measurement parameters η_0^* and η_1^* are also convex combinations with p_U, as shown in Figure A.1 by points on the line segments $\eta_0 - p_U$ and

$\eta_1 - p_U$. Thus, under independence, η_0^* and η_1^* must move closer to p_U and away from the corners of the tetrahedron that represent perfect measurement, shortening the overall length of the solution line. In other words, independent linkage error necessarily leads to higher classification errors. The MLEs of these measurement parameters, η_0^{**} and η_0^{**}, meanwhile, are found by projecting the solution line, not onto $\eta_0 - p_U$ and $\eta_1 - p_U$, but rather onto the independence surface. The distances length($\eta_0^* - \eta_0^{**}$) and length($\eta_1^* - \eta_1^{**}$) reflect violations of the LCM's conditional independence assumption. Therefore, linkage error does cause violations of the model's assumptions. However, as can be seen in Figure A.1, these violations will be negligible in practice, and the bias is bounded by a small number (relative to the solution line) that depends on P(U). In short, independent linkage errors are absorbed by the measurement parameters, leaving the structural parameters approximately unaffected.

In contrast, bias will be strong when linkage error is not independent, P(Y|U) $\neq$ P(Y). In this case, the new point may lie anywhere on the independence surface, destroying the parallel property of the new solution line. In this case, none of the previous results apply, and the bias in both measurement and structural parameters can be arbitrarily large.

Finally, we have assumed that the mislinked records have an independent joint distribution. When this assumption does not hold, the projection p_U should be replaced by a projection, $p_{U,\text{dep}}$, say, onto a "dependence surface" defined by a constant odds ratio (Fienberg and Gilbert 1970, pp. 699-701). Because of independence of linkage errors, the projection will still be orthogonal to A_1A_2 and A_2A_3. In this situation, the length of the solution line will still be reduced and classification errors will rise. However, the distance from the "true" interpolation between p_U^* and η to the corresponding projection onto the independence surface may increase. In other words, in this situation, depending on the strength of the dependence p_U^*, some non-negligible bias in the MLE of π may start occur. In

particular, for positive dependence (odds ratio > 1), π will be somewhat underestimated (overestimated for negative dependence).

In this appendix, we have indicated the consequences of linkage error for latent class analysis, and argued that independent linkage errors lead to a relatively small violation of the LCM's assumptions. Although we have not shown this here, we conjecture that the argument extends to higher-dimensional and multiple category problems, such as the HMM. We have also seen that dependence of linkage errors has more potential to cause bias than dependence in the mislinked records. Our paper investigates these conjectures using a simulation study.

Appendix A.2. The combined LFS and ER dataset

A.2.1 Background information on the LFS and ER

The LFS is an address-based sampling survey conducted by Statistics Netherlands which provides information on individuals' labour market position. As of the last quarter of 1999, it has been a rotating panel survey which consists of five waves conducted every 3 months.

The ER is an administrative 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. The data from both the LFS and the ER are linked at the individual level to the Population Register (PR) and so the target population of the data is restricted to individuals registered in the Netherlands.

A.2.2 Missing values

The dataset is unbalanced for the LFS as it suffers from attrition and has, for the non-survey months, observations missing completely at random (MCAR). More specifically, the first wave of the survey includes 8,708 individuals (130,620 observations), the second 7,458 (111,870 observations), the third 6,856 (102,840 observations), the fourth 6,739 (101,085

observations) and the fifth 6,560 (98,400 observations). While ostensibly the ER cannot suffer from drop-out as all employers are obliged by law to submit their reports, 2,619 observations are missing which amounts to just under 2% of the sample. Those observations are also assumed to be MCAR.

A.2.3 Record linkage procedure

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 from both sources 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 consistent in both data sources. If not, the records go to the next step together with those 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 unlinked records from the first two steps are linked using only the BSN and ignoring any differences in the other variables. This procedure is repeated monthly. The overall linkage effectiveness is approximately 96-97%, depending on the chosen month; 99.8% of all linked records are successfully linked in the first step.

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. Having selected only individuals aged 25-55, the linkage effectiveness of the combined sources is approximately 97%. The unlinked records 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.

Appendix A.3. Simulation design

Both pseudocodes illustrate conditions characterized by an overall 5% error rate and in which individuals who have transitioned (from temporary to permanent employment according to the register data) are oversampled.

A.3.1 Pseudocode for a false-negative linkage error condition

Step 1

1. Identify individuals who have had one or more three-monthly transitions:

   $Temp_{t-3} \rightarrow Perm_t$

2. If a given individual has had a transition, set their exclusion threshold t to .15
   a. Else, assign threshold t to .05

Step 2

3. For each individual in the sample, draw a random number from a standard uniform distribution - $U_i \sim \mathrm{U}(0,1)$
4. If $U_i \leq$ t , exclude individual i
   a. Else, do not exclude individual i

Step 3

5. Run the HMM on this new dataset and compare the results to the original ones

A.3.2 Pseudocode for a false-positive linkage error condition

Step 1

1. Identify individuals who have had 1 or more three-monthly transitions:

   $Temp_{t-3} \rightarrow Perm_t$

2. If a given individual has had a transition, assign mislinkage threshold t as .15
   a. Else, assign threshold t as .05

Step 2

3. For each individual in the sample, draw a random number from a standard uniform distribution - $U_i \sim \mathrm{U}(0,1)$
4. If $U_i \leq$ t, mislink individual i
    a. Else, do not mislink individual i

If the donor is random:

5. Assign to the linkage recipient the ER contract type of a randomly chosen individual

If the donor is based on characteristics:

5. a. Use R's *matchit* package to perform statistical matching based on age, gender, nationality, and education

    b. Assign to the linkage recipient the ER contract type of the matched individual

Step 3

6. Run the HMM on this mislinked data and compare the results to the original ones