Measurement error: estimation, correction, and analysis of implications

Pankowska, P.K.

2020

VRIJE UNIVERSITEIT AMSTERDAM

MEASUREMENT ERROR: ESTIMATION, CORRECTION, AND ANALYSIS OF IMPLICATIONS

Paulina Karolina Pankowska

Manuscript Reading Committee
dr. P.P. Biemer (RTI International)

prof.dr. A.G. De Waal (Dept. of Methodology, Tilburg School of Social and Behavioral Science, Tilburg University; Statistics Netherlands)

prof.dr. H.B.G. Ganzeboom (Dept. of Sociology, Vrije Universiteit Amsterdam)

prof.dr. W. Smits (Research Centre for Education and Labour Market, School of Business and Economics, Maastricht University; Statistics Netherlands)

prof.dr. J.K. Vermunt (Dept. of Methodology, Tilburg School of Social and Behavioral Science, Tilburg University)

Artwork by Aleksandra Daniłoś
Layout and design by Stijn Eikenaar| persoonlijkproefschrift.nl
Printed by Ipskamp Printing | proefschriften.net

ISBN: 978-94-6421-035-4

The research documented in this dissertation was funded by Statistics Netherlands (CBS).

VRIJE UNIVERSITEIT

MEASUREMENT ERROR: ESTIMATION, CORRECTION, AND ANALYSIS OF IMPLICATIONS

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor of Philosophy aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. V. Subramaniam,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Sociale Wetenschappen
op dinsdag 17 november 2020 om 15.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Paulina Karolina Pankowska

geboren te Krakau, Polen

promotoren:	prof.dr. B.F.M. Bakker dr. D. Pavlopoulos
copromotor:	dr. D.L. Oberski

Table of contents

Authors' contributions

Chapter 1. Written by Paulina Pankowska. Several rounds of feedback were provided by supervisors Bart Bakker, Daniel Oberski, and Dimitris Pavlopoulos.

Chapter 2. Written by Paulina Pankowska. The research idea was conceived by Paulina Pankowska and Daniel Oberski. Paulina Pankowska conducted the statistical analysis and drafted the chapter. Daniel Oberski critically revised the chapter.

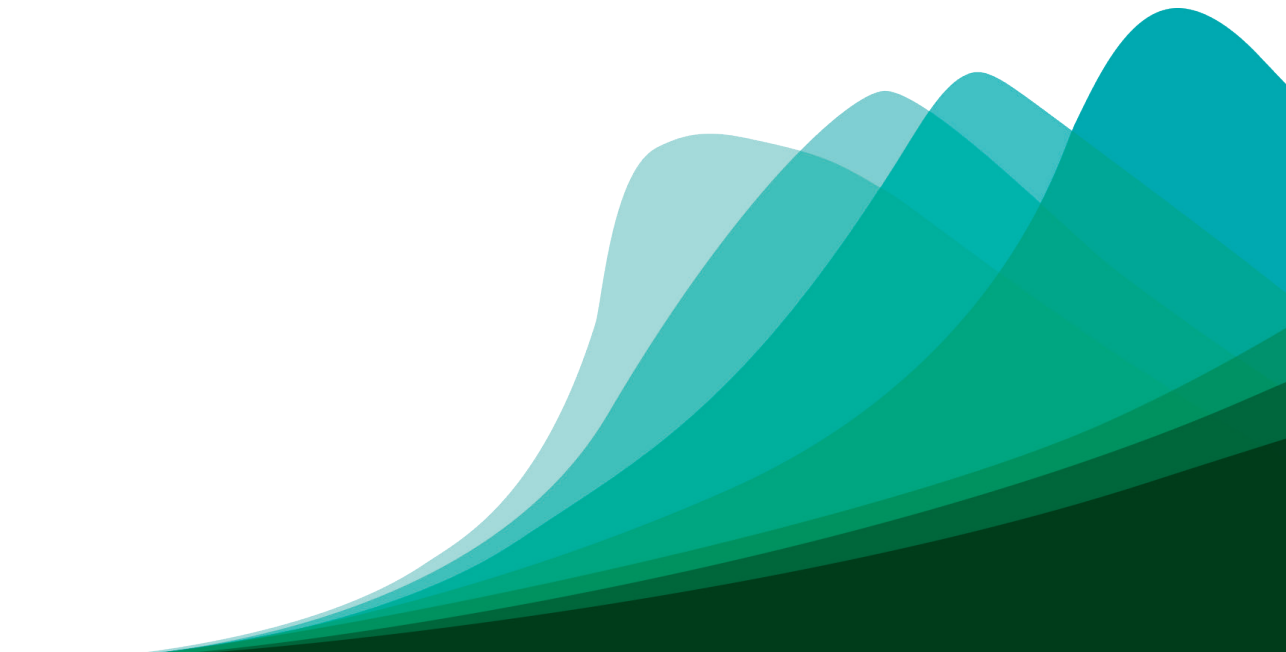
Chapter 3. Written by Paulina Pankowska. The research idea was conceived by the project's Steering Committee of Statistics Netherlands (CBS). Paulina Pankowska conducted the statistical analysis in collaboration with Daniel Oberski and Dimitris Pavlopoulos. Paulina Pankowska drafted the chapter with assistance from Bart Bakker, who provided input regarding the description of the data used. All authors critically revised the chapter. Barry Schouten reviewed and provided feedback on the chapter.

Chapter 4. Written by Paulina Pankowska. The research idea was conceived by Paulina Pankowska and the project's Steering Committee of Statistics Netherlands (CBS). Paulina Pankowska conducted the statistical analysis in collaboration with Daniel Oberski. Paulina Pankowska drafted the chapter with assistance from Daniel Oberski, who provided inputs for the methodology section. Bart Bakker, Daniel Oberski, and Dimitris Pavlopoulos critically revised the chapter. Peter Paul de Wolf provided feedback on the chapter.

Chapter 5. Written by Paulina Pankowska. The research idea was conceived by all authors. Paulina Pankowska conducted the statistical analysis in collaboration with Dimitris Pavlopoulos. Paulina Pankowska drafted the chapter. Bart Bakker, Daniel Oberski, and Dimitris Pavlopoulos were all closely involved in the writing process; all critically commented on and revised several versions of the chapter.

Chapter 6. Written by Paulina Pankowska. Several rounds of feedback were provided by supervisors Bart Bakker, Daniel Oberski, and Dimitris Pavlopoulos.

1



INTRODUCTION

P. Pankowska is the sole author of this chapter but elements of it are based on: Pankowska, P., Bakker, B. F. M., Pavlopoulos, D. & Oberski, D. L. (2020). Reconciliation of inconsistent data sources using hidden Markov models. Manuscript accepted for publication in the Statistical Journal of the IAOS.

Measurement error is a problem inherent to all data sources, in spite of countless attempts to reduce it and address its causes (Alwin, 2007; Biemer et al., 1991; Kuha & Skinner, 1997). Its presence often leads to biased and inconsistent statistical estimates and, as a consequence, to erroneous findings and conclusions. Such errors can also lower the precision of obtained estimates and reduce the power of statistical tests (Biemer & Wiesen, 2002). It is therefore crucial to understand, account, and correct for measurement error to ensure research validity (Fuller, 2009; Grace, 2017; Kuha & Skinner, 1997). What is more, in the context of official statistics, measurement error is particularly problematic as it has the potential to cause bias in both the descriptive statistics of a single variable and in the estimates of the relationships between multiple variables. This in turn hinders the production of reliable and accurate estimates by National Statistics Institutes (NSIs) (Boeschoten, 2019).

To provide a few illustrative examples, Pavlopoulos and Vermunt (2015) show that, due to measurement error, survey and register data for the same sample provide substantially different estimates of the distribution of employment contract types in the Netherlands. After correcting for this error, the authors find that permanent contracts are significantly overestimated in the survey data, while they are underestimated in the register data; however, the situation is reversed for temporary contracts (which are underestimated in the survey and overestimated in the register). The authors further demonstrate that "dynamic", over-time statistics are also affected by measurement error (and this effect is likely to be more severe than for "static", cross-sectional estimates) by showing that the 3-monthly transition rates from temporary to permanent employment are substantially inflated in both data sources. That is, according to the survey data, the transition rate is equal to 0.057, while according to the register records it is 0.085. When correcting for measurement error, the authors estimate the transition probability to be much lower than both data sources, amounting to just 0.032. This in turn implies that approximately half of the observed transitions are not in fact true transitions (Pavlopoulos & Vermunt, 2015). In a more general setup, Pavlopoulos et al. (2012) show that when a dichotomous random variable X is measured with a constant 0.05 misclassification rate across two time points (t_1 and t_2), the observed transition between t_1 and t_2 is overestimated by as much as a factor of 2.73. More specifically, the authors demonstrate that, if the true, error-free transition probability is set to 0.05, the transition probability estimated from data containing measurement error is inflated and amounts to 0.135.

Studies correcting for measurement error in continuous variables reach equally striking conclusions. For instance, Gottschalk (2005) reveals that the observed downward adjustments of nominal wages in the absence of a job change are highly overestimated due to measurement error. That is, the authors show that the fraction of individuals who reported a lower wage in t than in t, as observed in the cross-tabulation of the two surveys considered, is overstated by a factor of 3, compared to the error-corrected estimates (i.e. 17 vs. at most 5 percent). Using US and Irish data O'Neill and Sweetman (2013) show that the relationship between self-reported BMI and income

(calculated using a least squares estimator) is severely overestimated due to (non-random) measurement error in the provided BMI values. The reported biases range from 6 to 20 percent depending on the models and data used. In the context of official statistics, Scholtus et al. (2015) use structural equation modelling to demonstrate that measurement error in the administrative VAT data leads to a relative bias of 3 to 19 percent in the estimates of the 2012/2013 annual turnover levels (by NACE group) for the Netherlands. As a consequence, the uncorrected estimates substantially under- or overestimate the contributions of several NACE groups to the Dutch economy.

In general, measurement error occurs when the observed value of a random variable differs from its true (unobserved) value. For continuous data, this means that we do not observe a true random variable X directly, but rather we observe its measurement Y, which is the sum of the true value X and random noise variable ε (Gustafson, 2003):

$$Y = X + \varepsilon \tag{1.1}$$

where ε represents the error term (i.e. the noise), and in the absence of measurement error $Y = X$. In the case of categorical data, measurement error, which is referred to as misclassification, occurs when the actual and recorded categories of a multinomial variable differ. In this context, measurement error can be characterized in terms of misclassification probabilities, i.e. how likely is a wrong classification, given the true classification (Gustafson, 2003). To illustrate, consider a multinomial random variable X, with K categories, which cannot be observed directly but rather is measured in a survey by a variable Y, also with K categories. The misclassification probability can be then defined as follows:

$$P(Y = y|X = x), \text{ where } y, x \in \{0, \dots, K\} \text{ and } y \neq x \tag{1.2}$$

In surveys, measurement error is a well-studied phenomenon that is caused primarily by inadequate questionnaire design, incorrect data collection procedures, interviewer effects (Alwin, 2007; Biemer et al., 1991; Saris & Gallhofer, 2007) or respondent effects (Sudman et al., 1997; Tourangeau et al., 2000). In contrast, research on measurement error in register data is scarce. Despite this, however, it is well-known that administrative register data often contain errors (Bakker, 2012; Oberski et al., 2017; Oberski, 2015; Scholtus et al., 2015; De Waal et al., 2011). These errors can mirror those observed in surveys, in particular when they occur during data entry. However, some types of error are unique to registers. These include, administrative delay, definition error, and errors caused by administrative incentives (Bakker & Daas, 2012; Huynh et al., 2002; Zhang, 2012). “Big data” sources, such as data derived from sensors, can also be subject to measurement error (which is often referred to as “noise” in the data science literature), due to changes in the operating conditions, ageing of the sensors, and problems related to inaccuracies and imprecision (Elnahrawy & Nath, 2003). Depending on the mechanism causing it, the error can be either random i.e. without any specific

patterns, or systematic, i.e. occurring consistently over-time or with a probability of occurring that depends on covariates (Crocker & Algina, 1986; Scholtus, 2018).

The biasing effects of measurement error on model estimates specifically depend on a number of factors, including: (i) the complexity of the model under consideration, (ii) whether the estimates are cross-sectional or longitudinal, and (iii) the joint distribution of the error and the variables included in the model. Overall, the magnitude of these effects is particularly high when estimates are based on complex rather than simple (e.g. linear) models, when they are longitudinal, and when the errors are correlated with the variables included in the model. Depending on its severity, measurement error can reduce the efficiency of estimates, over- or under-estimate the relationships between variables and even reverse the signs of these relationships, obscure real effects, and/or lead to the emergence of spurious effects (Bound et al., 2001).

An increasingly popular approach that allows for the estimation of and correction for measurement error (without the need for gold standard data) relies on the use of latent variable modelling. This method has been applied broadly in a variety of settings, for instance, in a cross-sectional setup it was applied to categorical data by Biemer and Wiesen (2002), Flaherty (2002) and Pickles et al. (1995); it was also applied to continuous data by Bakker (2012) and to mixed type data by Oberski et al. (2017). In a longitudinal context, it was also used by Lugtig and Lensvelt-Mulders (2014) for continuous data and by Biemer and Bushery (2000) and Pavlopoulos and Vermunt (2015) for categorical data. Latent variable models (LVMs), unlike alternative measurement-error-correction techniques, do not make use of error-free validation data that are rarely available in practice. Instead LVMs make use of the availability of repeated indicators of the same variable, either cross-sectionally from various sources or over time from the same source, to extract information about measurement error directly from the data (Biemer & Bushery, 2000).

A group of LVMs that are applied to categorical, longitudinal data specifically (with $t \geq 3$), and which are the main focus of this thesis, are hidden Markov models (HMMs) (Biemer, 2004, 2011; Oberski et al., 2017; Pavlopoulos & Vermunt, 2015). The basic HMM operates under the assumption that, at each time point *t*, the observed data Y_t is generated independently with some probability $P(Y_t|X_t)$ from the true, but unobserved, value X_t, where both X and Y have L categories. Assuming the generation of Y_t to only involve X_t and to be independent of all other observed and true values, the observed distribution factorizes as:

$$P(Y) = \prod_{t=0}^{T} P(Y_t|X_t)P(X) \tag{1.3}$$

where, $P(Y)$ and $P(X)$ denote the observed path and the true, latent path, respectively.

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

$$P(Y) = \sum_{k=1}^{K} \prod_{t=0}^{T} P(Y_t|X_t)P(X = x_k) \tag{1.4}$$

where $K = L^T$ enumerates all possible patterns of X over the entire time period and x_k denotes a realized unobserved path. Classification error occurs when for any of the categories of the observed variable — Y_t — the response probability — $P(Y_t|X_t)$ — does not equal 1 for a unique category of X.

The unobserved true values (*latent states*) are assumed to follow a (first-order) Markov process, in which each value carries over partly to the next time point:

$$P(X) = P(X_0, \dots, X_T) = P(X_0)P(X_1|X_0) \dots P(X_T|X_{T-1}) \tag{1.5}$$

The full model then is

$$P(Y) = \sum_{x_0=1}^{L} \sum_{x_1=1}^{L} \dots \sum_{x_T=1}^{L} P(X_0) \prod_{t=1}^{T} P(X_t|X_{t-1}) \prod_{t=0}^{T} P(Y_t|X_t) \tag{1.6}$$

The parameters to be estimated for this model, typically in the form of a logit, are the structural parameters, i.e. the initial state probabilities — $P(X_0)$, the latent transition probabilities —$P(X_t|X_{t-1})$, and the classification or measurement error probabilities (often referred to as the measurement error probabilities) — $P(Y_t|X_t)$. A one indicator HMM, where Y denotes the observed variable and X refers to the "true" (or latent) state, is illustrated in Figure 1.1.

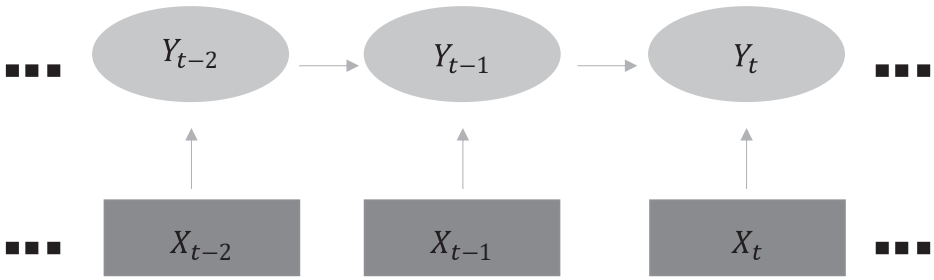


Figure 1.1 - Standard one- indicator hidden Markov model graph
Note: Rectangles denote the true variable X and ovals the observed variable Y; absence of arrows indicate independence and their presence dependence

The standard, single-indicator HMM relies on the local independence assumption for identifiability, which requires that the errors in the repeated measures occur independently. While necessary, this assumption is often viewed as highly restrictive and unrealistic, as it does not allow for the modelling of the presence of systematic errors without risking poor model identifiability. To overcome this challenge, it is possible to use extended, multiple-indicator versions of HMMs. The use of such models, in particular when the indicators come from different (independent) sources, makes the local independence assumption plausible across sources, while allowing for local dependence within sources (Bassi et al., 2000). Pavlopoulos and Vermunt (2015), for instance, use such an extended, two-indicator HMM to correct for measurement error in both survey and register data on the employment contract type in the Netherlands. Formally, the single-indicator HMM can be extended to multiple indicators by replacing $P(Y|X)$ above with $P(Y_1,Y_2|X) = P(Y_1|X)P(Y_2|X)$. While using multiple indicator extensions of HMMs is an attractive solution for the aforementioned problem, it also introduces some new challenges. Most importantly, record linkage might lead to linkage error – a new potential source of bias. Furthermore, the implementation of such extended models tends to be complex and time-consuming.

Given the potentially strong, adverse effects of measurement error and the possibility of minimizing these using HMMs, the aim of this thesis is twofold: first to understand in more detail the problem of measurement error and second to investigate whether it can be resolved using hidden Markov models. More specifically, the thesis examines whether and under what circumstances measurement error causes non-negligible bias (using clustering as an illustrative example), and whether error dependency (i.e. the presence of systematic error) is an important factor that needs to be considered. Building on from this foundation, the focus of the thesis then turns to whether extended HMMs that are applied to linked data can be used for error correction and whether this method can be feasibly implemented.

Thesis outline

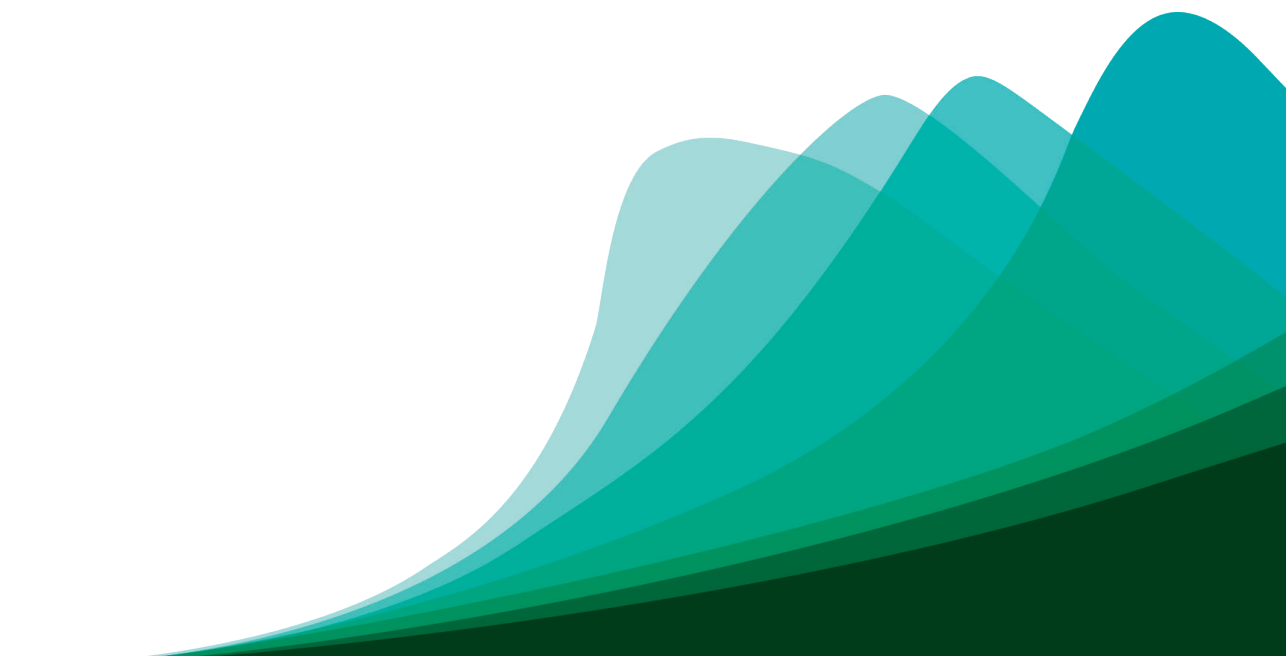
The remainder of the thesis is structured as follows. Chapters 2 and 3 illustrate the problem introduced by measurement error by showing its potential biasing effects and demonstrating its presence in different data sources. Chapter 2 investigates the bias introduced by measurement error in the context of clustering using a simulation study. In other words, it examines the extent to which error impacts the estimation of quantities of interest, and, in a related way, underlines the importance of correcting for it. In doing so, we test the sensitivity of the results of two commonly used model- and density-based clustering algorithms (i.e. GMMs and DBSCAN) to the presence of various degrees of random and systematic measurement error. More specifically, we assess the similarity of the clusters obtained using an error-free dataset to those obtained once error is introduced. We also compare the number of clusters found in the presence and

absence of error to determine whether measurement error obscures clusters and/or leads to the emergence of spurious clusters.

Chapter 3 then examines the effect of a major change in the data collection process (in this case a switch in the interviewing regime) on the nature and magnitude of measurement error. In this analysis we apply a two-indicator HMM to a linked survey and administrative data, which enables the modelling of error dependency in both sources. This paper provides a more detailed understanding of the problem of error dependency, as it assesses the extent to which dependent (systematic) errors are present in survey and administrative data and, what is more, whether error dependency is a by-product of specific data collection processes. This, in turn, goes some way towards answering the question of whether such error needs to be specifically accounted for when correcting for measurement error using HMMs.

Chapters 4 and 5 focus on the applicability of the discussed method, i.e. on the feasibility of using extended HMMs to correct for measurement error. As mentioned above, the use of multiple-indicator HMMs, which allow for the relaxing of the local independence assumption within sources, requires linking data sources at the micro level. However, such record linkage might lead to linkage error and, consequently, to biased estimates. Therefore, chapter 4 directly tests the sensitivity of the structural parameter estimates of a two-indicator HMM to varying degrees of false-positive and false-negative linkage error. Chapter 5 then attempts to resolve a more practical problem, that is the complexity of the implementation of multiple-indicator HMMs. The use of these models requires performing record linkage followed by model re-estimation for each new survey wave or administrative time period. While it is theoretically possible to run the analysis periodically and use the obtained error parameter estimates as a correction factor for a number of years, this practice is conditioned on the assumption that the size and structure of the error parameters are constant for the relevant time period. In Chapter 5 we therefore examine whether parameter estimates can be carried forward for a number of years, provided that no major changes in the data collection processes occurred.

2



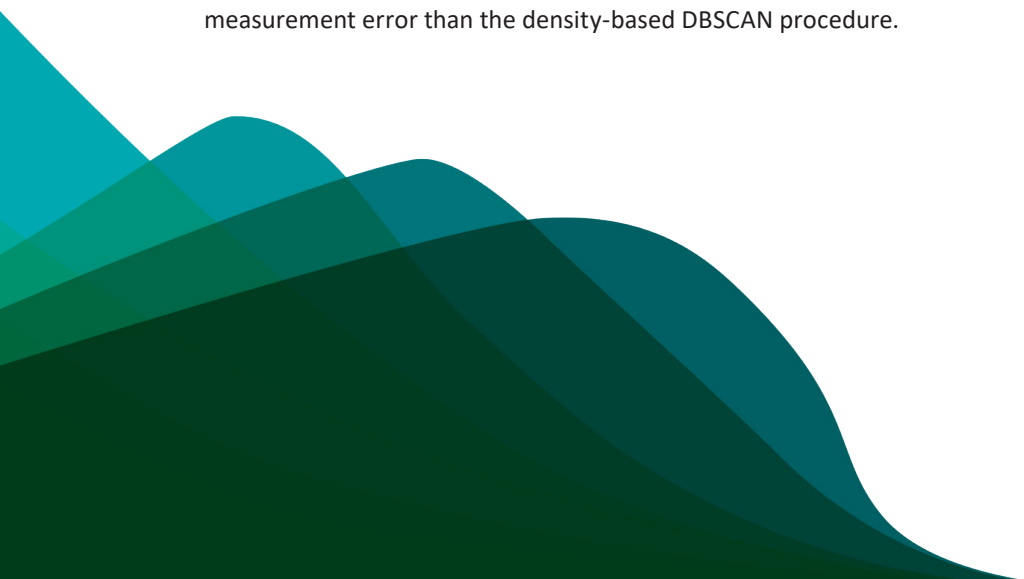
THE EFFECT OF MEASUREMENT ERROR ON CLUSTERING ALGORITHMS

This chapter is based on: Pankowska, P. & Oberski, D. L. (2020). The effect of measurement error on clustering algorithms. Unpublished manuscript.

Abstract

Clustering consists of a popular set of techniques used across the sciences to separate data into interesting groups for further analysis or interpretation. Many data sources on which clustering is performed are well-known to contain random and systematic measurement errors. Such errors may in turn adversely affect clustering – for instance, by producing spurious clusters or by obscuring useful clusters. Several techniques have been developed to deal with this problem, for example by merging partition-based mixture components, or by including a noise component in density-based clustering. However, little is known about the effectiveness of these commonly used solutions across a range of measurement error conditions. Moreover, no work to-date has examined the effect of systematic (non-independent and non-identically distributed (i.i.d.) and non-centered) errors on clustering solutions.

In this paper, we perform a Monte Carlo study to investigate the sensitivity of two commonly used model- and density-based clustering algorithms, GMMs with merging and DBSCAN, to differing magnitudes of both random and systematic error. We evaluate the number of clusters found, cluster stability, and the similarity of obtained clusters to the ones obtained in the absence of measurement error. We find that measurement error is particularly problematic when it is systematic as opposed to random, and when it affects all variables in the dataset. For the conditions considered here, we also find that the partition-based GMM with merged components is less sensitive to measurement error than the density-based DBSCAN procedure.



2.1 Introduction

Clustering is a popular set of statistical techniques widely applied in various scientific disciplines that allows for the separation of data into interesting groups for further analysis or interpretation (Aldenderfer & Blashfield, 1984; Dave, 1991; Tan et al., 2019). Its main goal is to divide observations, according to their degree of similarity, into a small number of relatively homogenous groups (Bailey, 1975). To illustrate, sociologists and economists often use clustering to group career paths and family trajectories, while in psychology and medicine it is commonly applied to identify different variations of an illness or to detect patterns in the spatial or temporal distribution of a disease (McVicar & Anyadike-Danes, 2002; Piccarreta & Billari, 2007; Tan et al., 2019). In the business world, clustering is performed in the context of customer/market segmentation, a process that divides the market into groups of customers with distinct needs, characteristics, and/or behaviors (Goyat, 2011; Tan et al., 2019). In addition, clustering is also frequently used in the fields of pattern recognition, information retrieval, machine learning, and data mining (Tan et al., 2019).

While clustering overall is an important and useful tool (Bailey, 1975), traditional clustering algorithms tend to assume the data are free from measurement error (Kumar & Patel, 2007). However, as is well-known, this is an unrealistic assumption. For example, surveys and registers are acknowledged to contain nonnegligible measurement error (Kumar & Patel, 2007; Pankowska et al., 2018, 2020). In surveys, measurement error is known to result from flaws in the survey response process, the process of data collection, processing, and editing, and from interviewer or respondent effects (Biemer, 2004; Sudman et al., 1997). Errors in register data can be caused by similar factors, but additionally suffer from administrative delay, definition error, and errors caused by administrative incentives (Bakker & Daas, 2012; Zhang, 2012). Other data sources, such as for instance, weblog data, also contain measurement errors (which are often referred to as “noise”) due to the presence of, among other things, online advertisements, navigation panels, copyrights notices, or webpage links from external websites (Onyancha et al., 2017). All such errors can be considered to have a random (centered i.i.d.) component, as well as a systematic component (location shift and dependence). For example, survey respondents tend to make the same (dependent) errors over time when answering questions (Pankowska et al., 2020).

How do random and systematic measurement error distort conclusions derived from data analysis? For regression and classification, it is well-known how errors bias parameter estimates of interest (see Carroll et al., 2006; Fuller, 2009; Gustafson, 2003). For example, Pavlopoulos and Vermunt (2015) and Pankowska et al. (2018) demonstrate that estimates of longitudinal turnover in people’s employment contracts differ by more than 300 percent —depending on whether measurement error is accounted for or not (estimated turnover proportion decreased from 0.07 to 0.02). However, in the context of clustering, little is known about such effects. On the one hand, errors have the potential to obscure existing clusters, or to produce spurious clusters. On the other,

clusters found may still be useful for the purposes at hand – for example interpretation, or relations to external covariates. Indeed, it is difficult to apply the concept of "bias" to the idea of clustering, since this method does not have a universally accepted single purpose (Hennig, 2015). In short, while it is clear that data used for clustering have errors, it is not obvious how these errors affect clustering results.

Among the few studies that have investigated the relationship between measurement error and clustering are Dave (1991), which demonstrated the impact of outliers on clustering, and Milligan (1980), which examined the effect of outliers, random error, and nonlinear distortion on clustering. Both concluded that cluster solutions were severely affected, although systematic error was not included in their studies. The effect of systematic error has been investigated in one very specific case, namely in medical diagnostic testing without a gold standard. This field has applied the two-class confirmatory latent class model, in which cluster interpretability is not explored, but assumed (Oberski, 2016). In the case in question here, the biasing effects of systematic error on model parameters of interest are well-documented (Hadgu et al., 2012; Torrance-Rynard & Walter, 1997; Vacek, 1985; Van Smeden et al., 2016). However, this work does not extend to more exploratory techniques, which may be focused on interpreting clusters and/or employing them for further analysis.

The observation that errors may affect clustering motivated the development of new techniques, including fuzzy *c*-means clustering (Bezdek et al., 1984), noise clustering (Banfield & Raftery, 1993; Dave, 1991), outlier-robust partition-based clustering (Davé & Krishnapuram, 1997; Gallegos & Ritter, 2005; García-Escudero et al., 2008), noise-robust density-based clustering (Ester et al., 1996), and other "noise-aware" clustering algorithms (see Aggarwal and Reddy, 2013, Ch. 18, for a review). In recent years, the application of (semi-) supervised and unsupervised deep neural networks to noisy data has sparked a literature on general noise-aware learning algorithms (e.g. Goldberger & Ben-Reuven, 2016; Malach & Shalev-Shwartz, 2017); these methods have been adapted to clustering as well, with a focus on improving classification performance after clustering (see Jindal et al., 2019, for an overview). Currently, however, we still lack an understanding of (i) the degree to which systematic — i.e. non-i.i.d. and/or uncentered — error affects traditional clustering techniques, and (ii) the degree to which interpretation-oriented purposes of clustering are affected.

In this paper, we perform a Monte Carlo study to investigate the sensitivity of two commonly used clustering algorithms, the Gaussian mixture model (GMM) and DBSCAN, to differing magnitudes and types of random and systematic measurement errors. These techniques were selected because GMMs are a key member of the model-based clustering family (Bouveyron et al., 2019), and DBSCAN was motivated specifically by the desire to handle noise (Ester et al., 1996), and therefore provides an interesting comparison. Additionally, DBSCAN, unlike GMMs, can handle non- spherical clusters such as moon-shaped clusters. We describe how measurement error affects the number of clusters found and the stability of the clusters, two criteria that lie at the basis of cluster interpretation (Hennig, 2015). We also evaluate the similarity to clusters

obtained in the absence of measurement error, a measure that can be conceived of as similar to that of "bias" in other techniques.

The remainder of the paper is structured as follows: section 2.2 first provides some background information on clustering techniques in general and on the GMM and DBSCAN algorithms in particular; it then discusses the topic of measurement error and its potential implications for clustering results. Section 2.3 explains the simulation setup and section 2.4 discusses the results of the analysis. Finally, section 2.5 offers some concluding remarks.

2.2 Background

2.2.1 Clustering

Cluster analysis is an umbrella term for a variety of algorithms and methods that are used to discover which observations in a dataset are similar and which dissimilar, given a combination of (measured) characteristics (Romensburg, 2004). Thus, the aim of clustering is to group cases such that observations belonging to the same cluster are more alike than those belonging to different clusters (Figueiredo Filho et al., 2014; Hair et al., 2014). As clustering can be seen as a classification problem with unobserved outcomes, it is an "unsupervised" learning problem (Jain, 2010; Bouveyron et al. 2019). Other applications include the use of clustering to help generate interesting research questions or hypotheses, as well as for strategic decision making in the management field (Romensburg, 2004).

There are numerous clustering algorithms available in the literature. Two commonly used approaches are density-based and model-based clustering (Maimon & Rokach, 2005). **Model-based clustering** is a probabilistic approach that assumes that the observed data was generated from a mixture of component models, where each of these component models is a probability distribution (Bouveyron et al., 2019). This clustering method requires predefining the number of clusters (Sammut & Webb, 2011). On the other hand, **density-based clustering** is a deterministic method that defines clusters in a data space as contiguous regions with high point density. Clusters are separated from each other by regions of low point density and data points lying in these low-density regions may be classified as outliers or noise. In the density-based clustering literature, the mixture models described above are also known as partition-based clustering. Unlike model-based methods, density-based clustering algorithms do not require the number of clusters as an input parameter (Kriegel et al., 2011), nor do they require the clusters to have a parametrically specified, usually convex, shape – they can therefore be seen as "nonparametric" clustering techniques (Kriegel et al., 2011; Maimon & Rokach, 2005).

The following two subsections provide an overview of the GMM and DBSCAN algorithms, two highly popular model- and density- based clustering algorithms that we use in our study.

Gaussian mixture models

Gaussian mixture models (GMMs) are among the most commonly used model-based clustering algorithms (Yeung et al., 2001). They belong to the family of latent variable models and can be defined as a parametric probability density function consisting of a weighted sum of Gaussian component densities (Reynolds, 2009). In other words, GMMs assume that data points are generated from a mixture of a finite (predetermined) number, K, of Gaussian distributions with unknown mean parameters, variance-covariance matrices, and cluster sizes (weights).

GMM seek to estimate a vector of parameters $\theta_k = \{\mu_k, \Sigma_k, w_k\}$ for each of the K d-dimensional multivariate Gaussian distributions that correspond to the clusters of interest, z. Conditional on the component $z = k$, the observed data vector $\mathbf{x}$ is assumed to follow the multivariate normal distribution,

$$f(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi)^{\frac{D}{2}}|\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k{}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)\right\} \quad (2.1)$$

The marginal density of the observed variables is simply a weighted sum of these K densities:

$$p(\mathbf{x} \mid \boldsymbol{\theta}) = \Sigma_{k=1}^{K} w_k f(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (2.2)$$

Where x is a d-dimensional vector of continuous data, w_k is a weight parameter for distribution k ($\Sigma_{k=1}^{K} w_k = 1$), μ_k is a d-length vector of means and Σ_k is a $d \times d$ variance-covariance matrix. Constraints can be imposed on this variance-covariance matrix. Common choices are to restrict it to a diagonal matrix (spherical components), to set all within-component covariance matrices equal, $\Sigma_k = \Sigma$ (equal shapes), or to specify a reduced-rank decomposition $\Sigma_k = \Lambda\Lambda' + \Psi$ (mixture of factor analyzers) (Bouveyron et al., 2019; McLachlan & Peel, 2004).

GMM parameters are estimated by fitting a pre-specified number of multivariate normal distributions to the data using the EM algorithm, iterating between estimating the posterior

$$\hat{p}^{(t)}(z = k \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \widehat{\boldsymbol{\theta}}^{(\boldsymbol{t-1})}, \ z = k)}{p(\mathbf{x} \mid \widehat{\boldsymbol{\theta}}^{(\boldsymbol{t-1})})} \quad (2.3)$$

and maximizing the expected likelihood

$$\widehat{\boldsymbol{\theta}}^{(\boldsymbol{t})} = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{\hat{p}^{(t)}(z=k|\mathbf{x})}[p(z, \mathbf{x} \mid \boldsymbol{\theta})] \quad (2.4)$$

These two steps are iterated until convergence of — $\widehat{\boldsymbol{\theta}}^{(t)}$ — the marginal likelihood (McLachlan & Peel, 2004; Reynolds, 2009). Note that the posterior estimates $\hat{p}^{(t)}(z = k \mid \mathbf{x})$ produced as a by-product of this procedure form a soft ("fuzzy") classification procedure for the discrete latent components variable, z. Direct optimization of the marginal likelihood $p(x|\theta)$ is possible as well, although usually avoided for stability reasons. Bayesian solutions to the estimation problem can be found in Frühwirth-Schnatter (2006).

The Gaussian parametric form restricts within-component shapes to "fuzzy" ellipses, whose contours decline exponentially. Fuzziness in clustering has been suggested in the literature to deal with random noise (e.g. Bezdek et al., 1984). In cases where the original clusters are elliptical, one might therefore expect that GMMs should be robust to (Gaussian) random errors. However, even in such ideal cases, systematic errors can easily distort their shape. For example, mean-regressive measurement error will create nonconvex clusters, which cannot be accounted for.

DBSCAN

DBSCAN – "Density-Based Spatial Clustering of Applications with Noise" – is a nonparametric, deterministic clustering algorithm which groups together points that are close to each other. The algorithm, developed by Ester, Kriegel, Sander and Xu (1996) requires two hyperparameters:

(i) ε (Eps)- the maximum distance between two points for them to be considered neighbors;

(ii) (*minPoints)*- the minimum number of neighboring points required to form a so- called dense region.

Using these two hyperparameters, the algorithm identifies the following:

a. ***ε*-neighborhood.** The ε-neighborhood of point p consists of all points q in the dataset D which are within an ε distance from p, which is determined using a distance function such as the Manhattan Distance or the Euclidean Distance; formally this can be defined by $\{q \in \mathcal{D} \mid \text{dist}(p, q) \leq \varepsilon\}$;

b. ***Core object/point.*** A core point is one that contains a number of points equal to or greater than *minPoints* in its *ε-neighborhood;*

c. ***Directly density- reachable points.*** Point q is defined as directly density-reachable if it is within the *ε-neighborhood* of p, and p is a core point;

d. ***Density-reachable points.*** Point q is density reachable from point p if for a chain of objects $p_1,...,p_n$, where $p_1 = p$ and $p_n = q$, p_{i+1} is directly density-reachable from p_i, given ε and *minPoints*, for $1 \leq i \leq n$. If q is density- reachable for a core point p but is not itself a core point, it is defined as a ***border point***;

e. ***Density connected points.*** Points p and q are density connected if there exists an object $o \in D$ which is a density-reachable point, given ε and *minPoints*, for both p and q;

f. ***(Density-based) cluster.*** A cluster C is a non-empty sub-set of D that satisfies the following conditions:
   - $\forall\, p, q$: if $p \in C$ and q is density-reachable from p, given ε and *minPoints, then* $q \in C$ (the so called "maximality" requirement)
   - $\forall\, p, q \in C$: *p is density-connected to q*, given ε and *minPoints.*

g. ***Noise.*** The noise cluster contains the set of points in dataset D that do not belong to any of the clusters $\{c_1, \ldots, c_i\}$; noise = $\{p \in \mathcal{D} \mid \forall\, i : p \notin C_i\}$;

Put simply, given the above, the algorithm starts by randomly selecting a core point $p \in D$ as a seed. It then finds all points in the dataset that are density-reachable from that seed and forms a cluster from a combination of the seed and these points. This process is repeated until all points in the dataset are assigned to a cluster or are classified as noise. DBSCAN is widely used, particularly in the data mining community, due to its flexibility, as it does not require the clusters to be of any specific shape or form (Birant & Kut, 2007; Ester et al., 1996).

2.2.2 Measurement error and its impact on clustering

Measurement error, which is often referred to as "noise" in the data science literature, occurs when the measured or observed value of a variable differs from its true value (Everitt & Skrondal, 2002). Thus, in the context of continuous variables, measurement error can be defined as the difference between the true and measured/observed value of a variable. The error can be either random, i.e. occurring by chance without a specific pattern, or systematic, e.g. such that either consistently under- or overestimates the values of a variable, is dependent on certain characteristics, or is subject to autocorrelation. Overall, measurement error has been shown to severely affect model estimates and lead to biased results (Crocker & Algina, 1986; Pankowska et al., 2018, 2020).

Formally, for a given random variable X and its observed counterpart Y, e.g. an individual characteristic such as income that is measured using a survey question, measurement error can be conceptualized in the following way:

$$Y = X + \varepsilon \tag{2.5}$$

Where ε is the measurement error term and, thus, in the absence of measurement error $Y = X$. When measurement error is random, we can think of ε as a normally distributed random quantity that is uncorrelated with X and is i.i.d, i.e. $\varepsilon \sim N(0,\sigma)$ and so $E[Y] = E[X]$. This is to say that in the presence of random measurement error, the observed value of random variable X differs from its true value in a way that is uncorrelated with X and which does not exhibit any specific patterns. In the survey context, such error occurs, for instance, when some individuals due to chance only either over- or underreport their income.

Systematic error (also referred to as systematic bias) can occur for a number of reasons. To illustrate, some survey respondents might systematically overreport their

income due to social desirability bias (Hariri & Lassen, 2017). In this case ε can be defined as a normally distributed random variable that is independent of X and i.i.d but such that $E[Y] > E[X]$; that is $\varepsilon \sim N(0,\sigma)$, where $\mu \neq 0$.

When the probability of making an error depends on a covariate Z that is uncorrelated with X, e.g. when the likelihood of overreporting one's income depends on whether the interview was conducted by proxy, we can think of ε as no longer an i.i.d random variable but rather one whose distribution parameters are some function of Z. In other words, while ε remains independent of X it is only i.i.d conditional on Z and can be defined as follows:

$$\varepsilon \sim \begin{cases} N(\mu_0, \sigma_0) \; if \; Z = 0 \\ N(\mu_1, \sigma_1) \; if \; Z = 1 \end{cases}, \text{ where } \mu_1 > \mu_0 \tag{2.6}$$

Finally, if the probability of misreporting income depends on the level of income itself, then ε is both no longer independent of X nor is it i.i.d. In this case, the relationship $\varepsilon \sim N(0,\sigma)$ still holds, but is extended in such a way that μ could be some monotonic function of X, with the substantive implication being that higher income individuals are more likely to misreport their income:

$$\varepsilon \sim N(\mu, \sigma), \text{ where } \mu = f(X) \tag{2.7}$$

The impact of the aforementioned types of measurement error on clustering specifically has not been studied extensively. Although the overall research on the topic is scarce, the literature available (which concerns solely random types of errors), does argue that clustering algorithms are likely to be (substantially) affected by measurement error (Dave, 1991; Frigui & Krishnapuram, 1996; Kumar & Patel, 2007). One of the few papers actually examining this impact is by Milligan (1980). The author investigates the effects of different types of error perturbation on the results of two types of clustering (hierarchical and k-means) and concludes that in many cases the presence of error in the data leads to a degradation in cluster recovery. This analysis, however, focuses predominantly on random error/noise and does not investigate the impact of systematic errors.

As mentioned above, given the lack of comprehensive evidence regarding the effects of measurement error on clustering, our simulation study looks at how different types and magnitudes of both random and systematic errors affect two aspects of clustering results. More specifically, we look at the number of clusters, as well as the similarity of the clusters to the "original" ones (i.e. those obtained in the absence of error). The choice of the GMM and DBSCAN algorithms (in addition to being driven by their popularity and wide application) is motivated by their potential to mitigate some of the effects of measurement error. More specifically, the attractiveness of GMMs is linked to the fact that they are probabilistic models and so can account for some of the uncertainty introduced by measurement error. The DBSCAN algorithm is used in

our analysis as it includes a noise cluster, which might potentially capture (some of the) observations that contain measurement error and leave the substantial clusters, to an extent, intact. The setup of the simulation study is discussed in detail in the next subsection.

2.3 Simulation setup

As stated above, we use a simulation analysis to demonstrate the effect of different degrees and types of measurement error on DBSCAN and GMM estimates. In more detail, our approach is to first generate a "baseline" dataset containing no measurement error, and then to compare model outcomes on that and error-induced datasets. These steps will be explained in more detail below. Also, as an illustration, Appendix 2.A provides pseudocode for generating the "baseline" dataset and introducing measurement error according to one condition.

Step I: Simulating the "baseline" dataset and performing clustering

First, we generated the initial/original, error-free dataset. In essence, our aim was to create a simple dataset consisting of a mixture of multivariate Gaussians, which will ensure strong internal cohesion (homogeneity) and external isolation (separation) of estimated clusters. A more complex data structure could negatively affect cluster recovery and lead to a situation wherein the algorithms produce different results for the same dataset, even in the absence of measurement error, due to random model variability. In this case, it would be difficult if not impossible for us to separate the effect of the data structure from that of introducing measurement error on the clustering results. As such, we drew $n = 1000$ observations from a mixture of three multivariate normal (MVN) distributions, with deterministic proportions of 0.4, 0.35, 0.25. To rephrase, this is to say that the first 400 observations were drawn from $MVN\ A$, the next 350 from $MVN\ B$, and the final 250 were taken from $MVN\ C$. As each MVN had dimensionality of three (corresponding to variables X_1, X_2 and X_3), the end result was a (1000, 3) matrix of random variables. To ensure the aforementioned separation of sample clusters, we used the following population parameters for our simulation[1]:

1 It is worthwhile noting that these population parameters were selected at random; the only consideration was obtaining spherical, fully separated clusters in the absence of measurement error.

$$G_1 \sim \boldsymbol{N}(\boldsymbol{\mu_1}, \boldsymbol{\Sigma_1}) \text{ where } \boldsymbol{\mu_1} = \begin{bmatrix} -2 \\ 9 \\ 12 \end{bmatrix} \text{ and } \boldsymbol{\Sigma_1} = \begin{bmatrix} 1.50 & 0.30 & 0.20 \\ 0.30 & 0.80 & 0.15 \\ 0.20 & 0.15 & 1.30 \end{bmatrix}; n_1 = 400$$

$$G_2 \sim \boldsymbol{N}(\boldsymbol{\mu_2}, \boldsymbol{\Sigma_2}) \text{ where } \boldsymbol{\mu_2} = \begin{bmatrix} 5 \\ 11 \\ 18 \end{bmatrix} \text{ and } \boldsymbol{\Sigma_2} = \begin{bmatrix} 2.00 & 0.40 & 0.15 \\ 0.40 & 1.60 & 0.25 \\ 0.15 & 0.25 & 1.00 \end{bmatrix}; n_2 = 350$$

$$G_3 \sim \boldsymbol{N}(\boldsymbol{\mu_3}, \boldsymbol{\Sigma_3}) \text{ where } \boldsymbol{\mu_3} = \begin{bmatrix} 4 \\ 4 \\ 5 \end{bmatrix} \text{ and } \boldsymbol{\Sigma_3} = \begin{bmatrix} 1.70 & 0.60 & 0.30 \\ 0.60 & 1.50 & 0.40 \\ 0.30 & 0.40 & 1.45 \end{bmatrix}; n_3 = 250$$

The draws from these multivariate Gaussians (constituting the dataset) correspond to the visualization depicted in Figure 2.1.

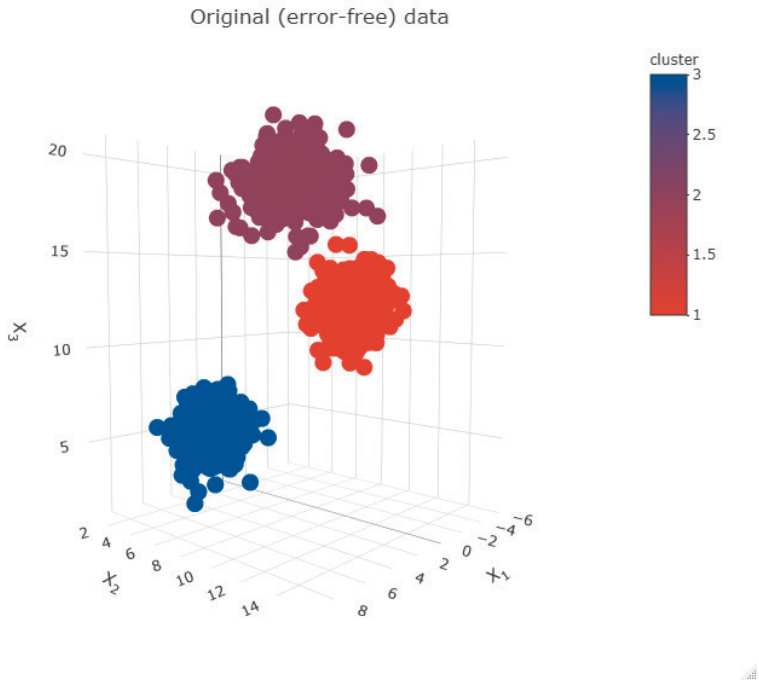


Figure 2.1- 3D scatterplot of the ""baseline" simulated dataset

Next, we performed clustering on the simulated dataset (i.e. the" baseline" dataset) using a Gaussian mixture model (GMM) and DBSCAN clustering. When fitting GMMs, we fit several models with the number of clusters, k, varying from 1 to 10 and chose the model with the best model fit, i.e. the lowest BIC. When using DBSCAN, we set the minimum number of neighboring points to be four for all conditions (as it is recommended that $minPoints = no.\ of\ dimensions + 1$), while we allowed ε to vary per condition and chose the appropriate distance based on a visual inspection of the k-nearest neighbor distance plot.

Step II: Introducing measurement error into the "baseline" dataset and performing clustering

Having fit the models to the "baseline" dataset, we then introduced various types and levels of measurement error into this data. In doing so, we considered a total of 36 conditions, each bootstrapped 100 times.

As illustrated in Figure 2.2, the following factors were varied in the conditions considered:

- Measurement error rate/ proportion of observations subject to error: 0.1, 0.2 vs. 0.4 (3 levels);
- Number of variables containing measurement error: 1 vs. 3 (i.e. all) (2 levels);
- Type of measurement error: random vs. systematic (2 levels);
- The magnitude of measurement error: low, medium vs. high (3 levels).

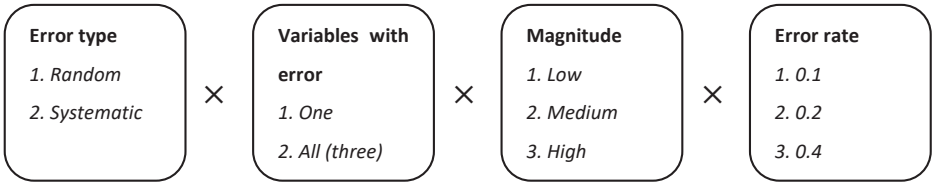


Figure 2.2- Outline of simulation setup/ simulation conditions

In more detail, for each condition we first randomly selected 0.1, 0.2, or 0.4 of the observations in the dataset. For these observations, we then introduced errors in either one or all three variables. When simulating random error in one of the variables (i.e. only in X_1), we added a draw from a normal distribution with $\mu = 0$, $\sigma = \{4,8,16\}$. The different σ's represent varying degrees of error severity (i.e. low, medium, and high). When introducing systematic error to X_1, we added a draw with $\mu = 2.5,5,10$,$\sigma = \{2\}$ wherein the $\mu's$ represent the three different error magnitudes. This is equivalent to the first type of systematic error which is discussed in section 2.2 (i.e. where the error term can be defined as follows: $\varepsilon \sim N(0,\sigma)$, where $\mu \neq 0$). For the conditions where random error affects all three variables, we added draws normal distributions where:

$$\mu_{x1} = 0,\ \sigma_{x1} = \{4, 8, 16\}$$
$$\mu_{x2} = 0,\ \sigma_{x2} = \{2, 4, 16\}$$
$$\mu_{x2} = 0,\ \sigma_{x2} = \{6, 12, 24\}$$

For systematic error we used the following:

$$\mu_{x1} = \{2.5,5,10\},\ \sigma_{x1} = \{2\}$$
$$\mu_{x2} = \{-2.5,-5,-10\},\ \sigma_{x2} = \{2\}$$
$$\mu_{x3} = \{1.25,2.5,5\},\ \sigma_{x3} = \{2\}$$

Figure 2.3 visually shows how the introduction of measurement error affects the simulated dataset, using the following four conditions as illustrative examples: (i) ***random*** error affecting **one variable** with rate of 0.4 and high magnitude; (ii) **random**

error affecting **three variables** with rate of 0.4 and high magnitude; (iii) **systematic** error affecting **one variable** with rate of 0.4 and high magnitude; (iv) **systematic** error affecting **three variables** with rate of 0.4 and high magnitude.[2]

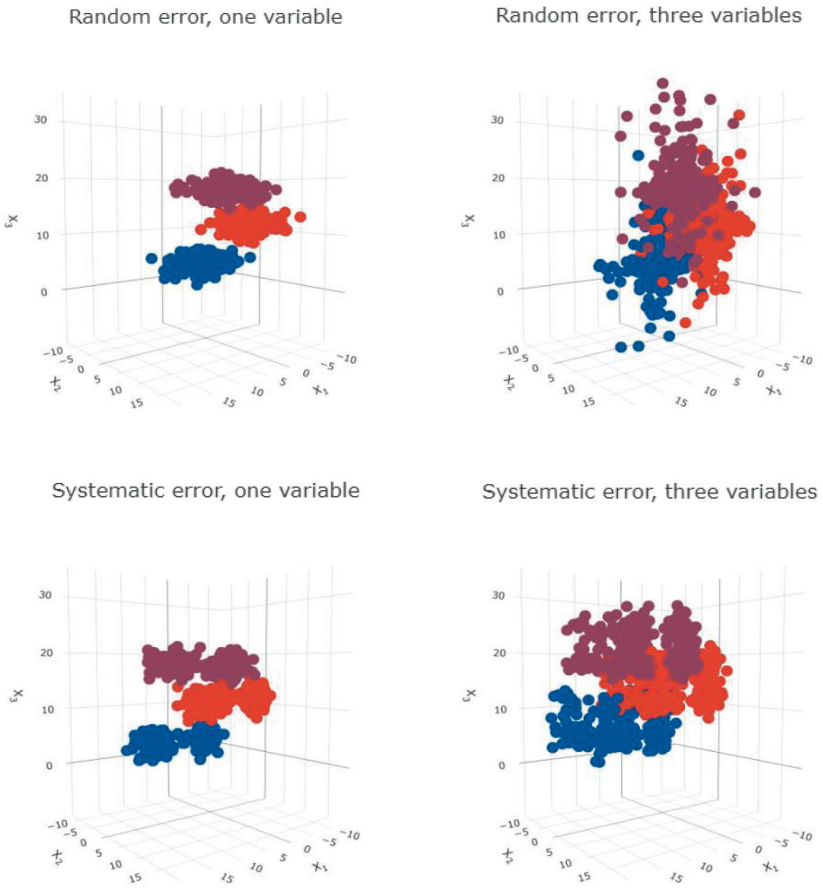


Figure 2.3- 3D scatterplot of datasets containing measurement error

As can be seen from the figure, when random measurement error affects one of the variables in the dataset, even though it is characterized by high error rate and severe magnitude, the resultant data structure still largely corresponds to the one in the absence of error. That is, even though the groups present in the data appear more "stretched out" and there are some outliers, the data is still characterized by the presence of three, clearly separated groups, which overall seem similar to the original

2 The selection of the more extreme conditions (i.e. high error rate and magnitude) was motivated by the fact that their influence on the dataset and subsequently the clustering results is expected to be substantial and highly visible.

ones. When the random error affects all three variables, on the other hand, the original, three fully separated groups largely overlap; this overlap is likely to impede the recovery of the original clusters. With regards to the condition wherein systematic error affects one variable, as can be deduced from Figure 2.3, the algorithms are likely to return results that also include spurious clusters. However, as these additional error-driven clusters appear very similar to the original ones, which can still be largely observed in the data, merging the clusters based on similarity (a procedure which is explained in more detail below) might mitigate the impact of this error. Finally, when systematic error affects all three variables, the clusters overlap to such an extent that the clustering algorithms are likely to produce highly dissimilar results to those obtained using the error-free dataset.

Having introduced the error, we then performed clustering on the resultant datasets, using GMM and DBSCAN. In doing so, we followed the same steps as when performing clustering on the "baseline" dataset. Finally, we compared the results to those obtained when no measurement error was introduced (i.e. when using the "baseline" dataset).

Step III: Comparing clustering results in the absence and presence of measurement error

When comparing the results, we focused on two specific metrics: the number of clusters obtained and the similarity of the clusters. While we consider similarity to be of much greater substantive importance, we also look at the number of clusters to understand whether different types of measurement error either obscure clusters or lead to spurious clusters. The importance of the similarity criterion stems from the fact that when the clusters obtained in the presence of measurement error are largely similar to those obtained in its absence, regardless of the number of clusters returned the results can be used for further research or interpretation, and the inferences made should be largely unbiased.

The examination of the number of clusters was relatively straightforward and involved simply comparing the number of the clusters obtained in the absence and presence of measurement error. The evaluation of cluster similarity was carried out based on the adjusted Rand index. The Rand index is a commonly used measure of the similarity between two clusters which varies from 0 to 1, where 0 implies perfect dissimilarity and 1 perfect match (Rand, 1971).

The Rand Index can be formalized as follows:

$$\text{Rand Index} = \frac{a+b}{\binom{n}{2}} \tag{2.8}$$

Where a is the sum of the number of paired observations that are grouped together in the same cluster for both clustering results and b is the number of paired observations that are ungrouped and belong to different clusters for both clustering results. $\binom{n}{2}$ represents the sum of all possible unordered pairs (Rand, 1971). The adjusted Rand

index is in principle similar to the original index but in addition it accounts for the fact that pairs of observations can be correctly grouped or ungrouped due to chance; it is bounded between ±1 (Hubert & Arabie, 1985).

In addition to considering the number and similarity of the clusters obtained directly from the fitted GMM (the so-called mixture components), we also examined the clusters obtained by merging the mixture components. Such merging is a common practice that is applied when the resultant mixture components are not separated sufficiently from one another for them to be interpretable and meaningful. The process is performed in a hierarchical order, whereby the value of a given merging criterion is computed for all pairs of components and the pair with the highest value is merged. Criterion values are then recomputed for the resultant clusters and the merging process continues until the highest criterion value obtained is below a predefined cut-off value. In our application, we use the Bhattacharyya distance as our criterion value and apply a threshold of 0.1. For further details regarding the merging process and the criterion used we refer to Hennig (2010).

For DBSCAN, we also compare the numbers and similarity of the "baseline" clusters with a subset of the clusters obtained in the presence of error that includes only stable clusters. To calculate stability, we resampled the datasets for each condition using bootstrapping (50 iterations) and compared the clustering results of the bootstrapped samples to those obtained on the original erroneous datasets. In doing so, we used the Jaccard similarity coefficient, which is defined as the size of the intersection of two clusters divided by the size of the union of these clusters. We considered a cluster to be stable if on average, given the 50 bootstraps we run, the Jaccard coefficient was higher than 0.7. For further details regarding the calculations of cluster stability and the criterion used we refer to Hennig (2007).

The analysis was carried out using the R environment for statistical computing (version 3.4.4). When fitting the algorithms to the datasets as well as when merging the GMM components and checking for cluster stability for the DBSCAN results, we used predominantly the *Flexible Procedures for Clustering (fpc)* package (Hennig, 2015).

2.4 Results

2.4.1 Clustering in the absence of measurement error

The clustering results obtained in the absence of measurement error (i.e. using the "baseline" dataset) for both GMM and DBSCAN almost perfectly recover the population parameters used to simulate the data. However, as the population parameters were set to ensure the emergence of three distinct, perfectly separated clusters, this was to be expected. More specifically, for GMM, the model that fits the data best (i.e. has the lowest BIC) correctly classifies all 1,000 observations in the dataset and returns the following three clusters (which are extremely similar to the Gaussian distributions used to simulate the data):

$$G_1 \sim \boldsymbol{N}(\widehat{\boldsymbol{\mu_1}}, \widehat{\boldsymbol{\Sigma_1}}) \text{ where } \widehat{\boldsymbol{\mu_1}} = \begin{bmatrix} -1.93 \\ 8.99 \\ 11.99 \end{bmatrix} \text{ and } \widehat{\boldsymbol{\Sigma_1}} = \begin{bmatrix} 1.63 & 0.36 & 0.36 \\ 0.36 & 0.99 & 0.14 \\ 0.36 & 0.14 & 1.52 \end{bmatrix}; n_1 = 400$$

$$G_2 \sim \boldsymbol{N}(\widehat{\boldsymbol{\mu_2}}, \widehat{\boldsymbol{\Sigma_2}}) \text{ where } \widehat{\boldsymbol{\mu_2}} = \begin{bmatrix} 4.95 \\ 10.98 \\ 17.99 \end{bmatrix} \text{ and } \widehat{\boldsymbol{\Sigma_2}} = \begin{bmatrix} 1.78 & 0.32 & 0.22 \\ 0.32 & 1.41 & 0.26 \\ 0.22 & 0.26 & 0.95 \end{bmatrix}; n_2 = 350$$

$$G_3 \sim \boldsymbol{N}(\widehat{\boldsymbol{\mu_3}}, \widehat{\boldsymbol{\Sigma_3}}) \text{ where } \widehat{\boldsymbol{\mu_3}} = \begin{bmatrix} 4.12 \\ 4.01 \\ 5.06 \end{bmatrix} \text{ and } \widehat{\boldsymbol{\Sigma_3}} = \begin{bmatrix} 1.21 & 0.47 & 0.26 \\ 0.47 & 1.40 & 0.28 \\ 0.26 & 0.28 & 1.53 \end{bmatrix}; n_3 = 250$$

The DBSCAN results return four clusters: three substantive ones and a noise cluster. The centroids of the substantive clusters (calculated based on cluster membership) are as follows:

$$C_1 = \begin{bmatrix} -1.91 \\ 9.02 \\ 12.01 \end{bmatrix}; n_1 = 391$$

$$C_2 = \begin{bmatrix} 4.91 \\ 11.02 \\ 18.00 \end{bmatrix}; n_2 = 331$$

$$C_3 = \begin{bmatrix} 4.10 \\ 3.98 \\ 5.00 \end{bmatrix}; n_3 = 242$$

The algorithm assigns 36 observations to the *noise cluster*, all remaining observations (96.4 percent) are classified correctly.

2.4.2 Clustering in the presence of measurement error

GMM estimates

The results obtained when fitting the GMM algorithm to the datasets containing measurement error (for all 36 conditions) are displayed in Table 2.1 and Figures 2.4 and 2.5. As can be seen, overall, the number of clusters as well as cluster similarity (calculated based on the Adjusted Rand Index) remain largely unaffected by random measurement error, provided that only one of the variables in the dataset is subject to error and that magnitude of this error is relatively low or of medium magnitude. The effect of the error rate appears negligible in this case. When error severity is high, on the other hand, we can observe the emergence of spurious clusters, although cluster similarity remains relatively high. Random measurement error also leads to spurious clusters when it affects all three variables, regardless of its magnitude and the error rate. The similarity between the clusters obtained for these conditions and the "original" ones is inversely related to the error rate and its magnitude (i.e. as the error rate and/or magnitude increase, the similarity between the aforementioned clusters decreases).

The effect of systematic error on the clustering results appears significantly more severe. Namely, virtually all 18 conditions can be characterized by the emergence of spurious clusters. What is more, the similarity measure is only truly high when the error affects one variable and is low in magnitude (regardless of the error rate). The remaining conditions return clusters that are substantially different from those in the "baseline" dataset.

When merging the obtained mixture components into more meaningful clusters, a highly optimistic picture regarding the robustness of GMMs to random measurement error emerges. More specifically, the number of clusters appears largely unaffected by measurement error with the exception of two rather extreme conditions, i.e. when the error affects all three variables, its rate is 0.4, and it is either medium or large in magnitude. In the case of these two scenarios measurement error obscures clusters. The resultant clusters are also in most cases highly similar to those obtained in the absence of error. Again, the two above specified conditions are an exception and lead to the emergence of dissimilar clusters. The clusters obtained under the condition wherein a high in magnitude random error affects all three variables and 0.2 of the observations are also dissimilar to the "original" ones, albeit to a lesser extent.

While merging also improves the clustering results for datasets that contain systematic error, it does so to a lesser extent. That is, systematic error distorts the number of clusters for half of the conditions considered, i.e. when the error affects one variable and is large in magnitude or when it affects all three variables and its severity is either medium or high. Likewise, cluster similarity can also be considered dissatisfactory for these conditions. It is worth mentioning that the adjusted Rand index is particularly low when three variables are subject to medium systematic error and 40 percent of observations are affected, or when the error is large and 20 or 40 percent of the cases are affected.

Table 2.1- GMM clustering results by simulation condition

Error type	Variables incl. error	Magnitude	Error rate	Number of clusters	Adjusted R Index	Number of merged clusters	Adjusted R Index of merged clusters
Random	one	low	0.1	3.0	0.999	3.0	0.999
			0.2	3.0	0.997	3.0	0.999
			0.4	3.0	0.999	3.0	0.999
		medium	0.1	3.2	0.986	3.0	0.998
			0.2	3.5	0.952	3.0	0.997
			0.4	3.2	0.979	3.0	0.998
		high	0.1	4.5	0.914	3.1	0.993
			0.2	5.9	0.775	3.1	0.995
			0.4	5.1	0.725	3.0	0.996
	three	low	0.1	4.5	0.908	3.0	0.991
			0.2	5.1	0.817	3.0	0.984
			0.4	4.8	0.739	3.0	0.965
		medium	0.1	4.2	0.892	3.1	0.973
			0.2	5.8	0.742	3.2	0.952
			0.4	5.8	0.600	2.1	0.514
		high	0.1	4.1	0.839	3.5	0.937
			0.2	6.3	0.632	3.7	0.777
			0.4	7.1	0.460	1.8	0.314

Table 2.1- GMM clustering results by simulation condition (continued)

Error type	Variables incl. error	Magnitude	Error rate	Number of clusters	Adjusted R Index	Number of merged clusters	Adjusted R Index of merged clusters
Systematic	one	low	0.1	3.7	0.950	3.0	0.999
			0.2	3.8	0.941	3.0	0.999
			0.4	3.1	0.992	3.0	0.999
		medium	0.1	7.1	0.752	3.4	0.984
			0.2	7.5	0.643	3.2	0.987
			0.4	6.7	0.573	3.0	0.996
		high	0.1	6.8	0.794	4.8	0.834
			0.2	7.0	0.655	5.9	0.733
			0.4	6.0	0.586	6.0	0.587
	three	low	0.1	5.6	0.800	3.1	0.995
			0.2	6.3	0.685	3.0	0.991
			0.4	5.6	0.683	3.0	0.983
		medium	0.1	7.8	0.654	5.2	0.888
			0.2	8.8	0.517	4.4	0.848
			0.4	8.5	0.406	2.1	0.475
		high	0.1	8.7	0.599	6.6	0.678
			0.2	9.0	0.380	5.2	0.373
			0.4	8.9	0.274	3.4	0.204

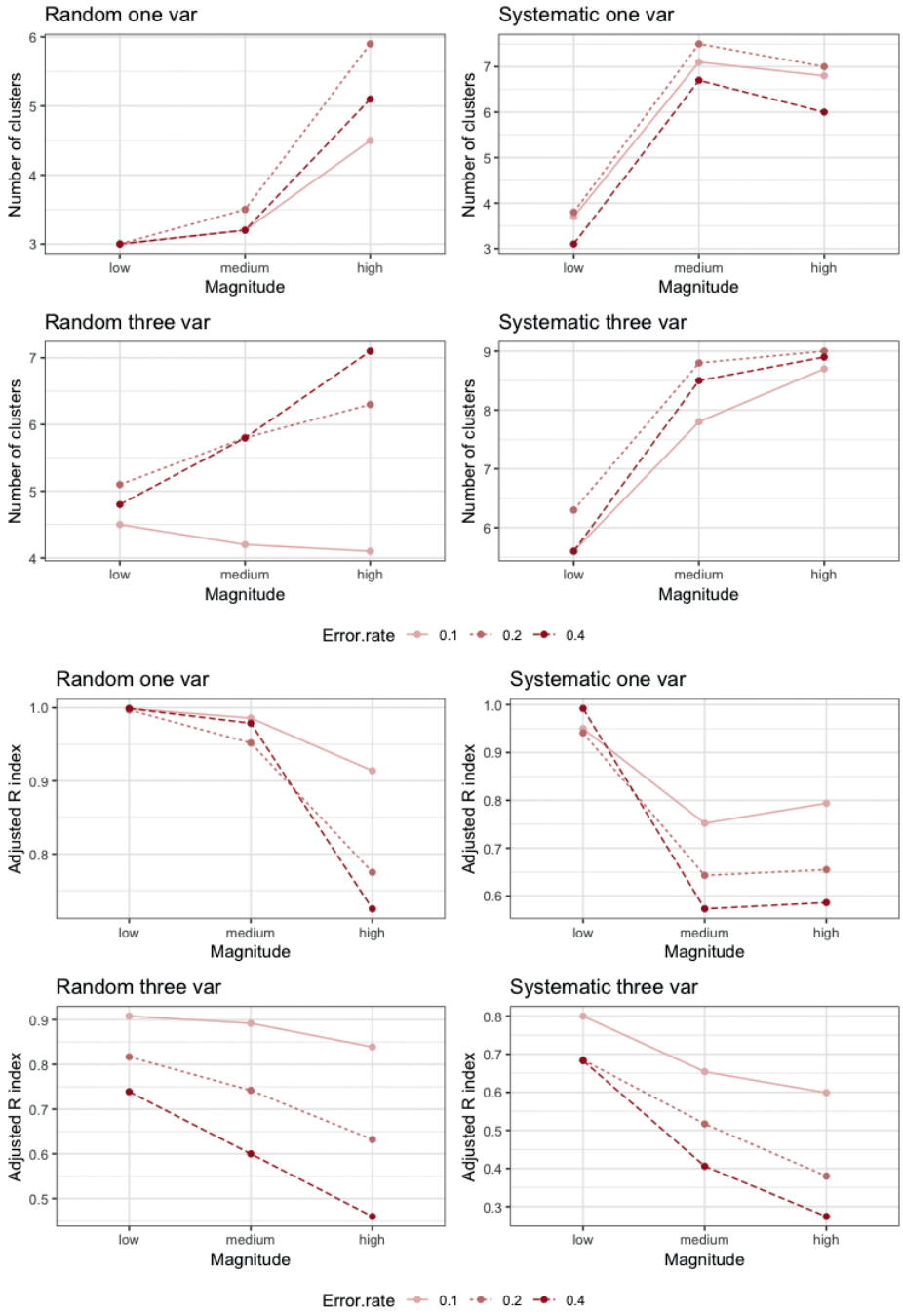


Figure 2.4- (top) Number of clusters and (bottom) cluster similarity for GMM mixture components

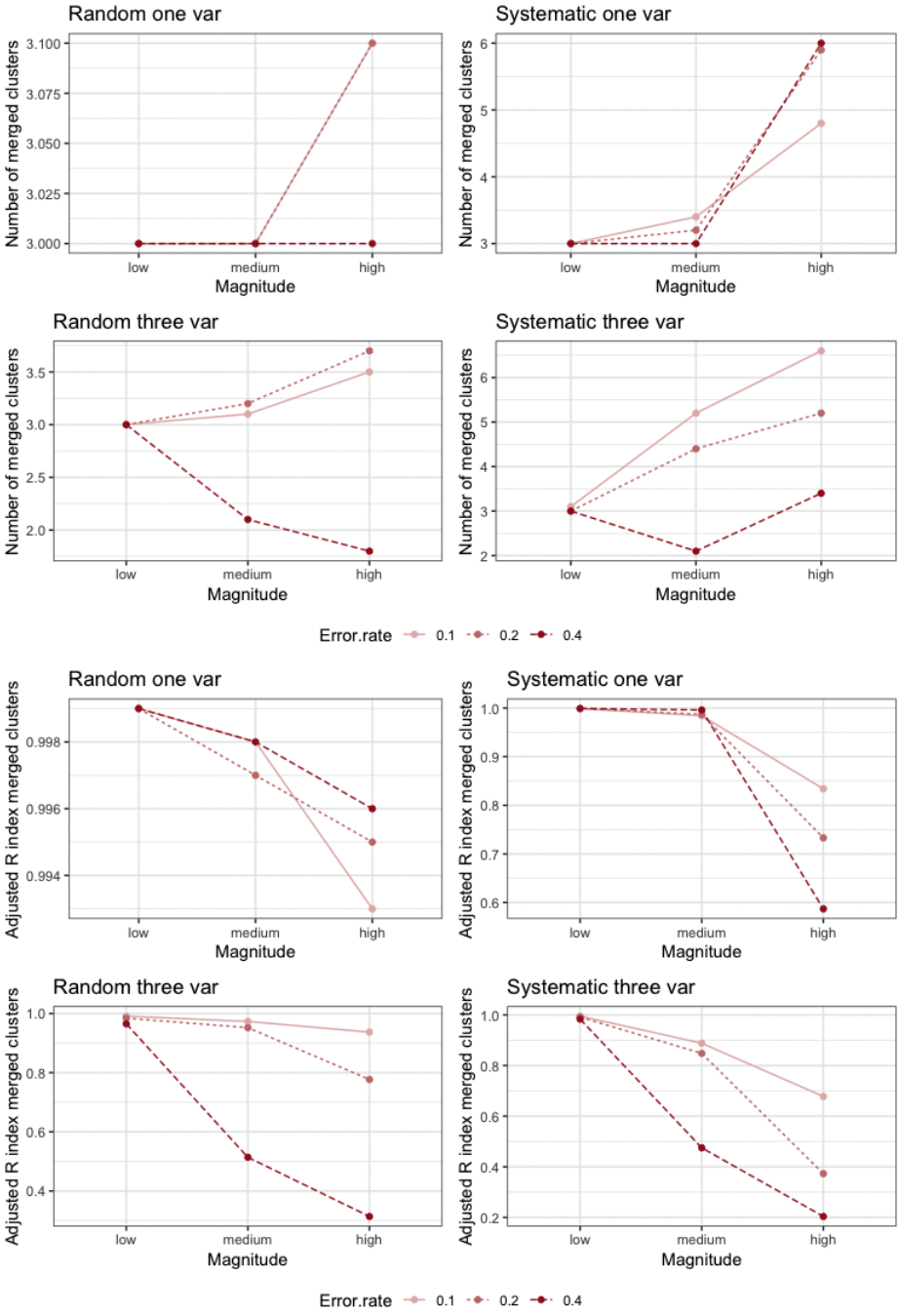


Figure 2.5- (top) Number of clusters and (bottom) cluster similarity for GMM merged clusters

DBSCAN estimates

The DBSCAN results presented in Table 2.2 and Figure 2.6 suggest that this clustering algorithm performs worse than GMM in the presence of measurement error. This is particularly striking when looking at cluster similarity. In more detail, when looking at the mean number of clusters obtained in each condition, it can be observed that for most conditions the presence of measurement error does not lead to (many) spurious clusters nor does it obscure clusters. The number of clusters is strongly inflated primarily when the error is systematic and high in magnitude (regardless of the number of variables affected and the error rate).

The degree of similarity of the clusters, however, appears more sensitive to measurement error. That is, DBSCAN returns substantially dissimilar clusters when it is applied to datasets containing random or systematic errors that affect all three variables (with the exception of the conditions in which the error rate is 0.1 and the magnitude low). In other words, for the aforementioned conditions the resultant clusters have very little in common with the ones obtained using the error-free data. Overall, the results do not substantially differ when considering only stable clusters.[3] Furthermore, as can be seen in Figure 2.7, contrary to expectations the noise cluster does not appear to capture observations that are subject to measurement error. This is the case even when the error is random and of very high magnitude, i.e. when the error is anticipated to lead to outliers, which should theoretically be assigned to the noise cluster. More specifically, Figure 2.7 provides an overview of the size cluster for each of the 36 simulation conditions. For most conditions, the number of observations included in the noise cluster is only slightly higher than the number of observations included in that cluster in the absence of measurement error. More specifically, in most cases the noise cluster size does not exceed 50, while for the error-free data this cluster consists of 36 observations. Therefore, it can be concluded that most observations that were subject to measurement error (i.e. a total of 200 or 400, depending on the condition) were not classified as noise.

3 It is worthwhile mentioning that for all 36 conditions the noise cluster was unstable for most bootstraps.

Table 2.2- DBSCAN clustering results by simulation condition

Error type	Variables incl. error	Magnitude	Error rate	Number of clusters	Adjusted R Index	Number of stable clusters	Adjusted R Index of stable clusters	Size of noise cluster
Random	one	low	0.1	4.3	0.965	3.0	0.992	44.0
			0.2	4.2	0.946	3.1	0.971	33.0
			0.4	4.3	0.934	3.1	0.966	35.7
		medium	0.1	4.3	0.946	3.2	0.975	41.6
			0.2	4.6	0.924	3.2	0.957	35.7
			0.4	4.7	0.889	3.0	0.896	34.0
		high	0.1	4.5	0.922	3.6	0.942	44.2
			0.2	5.2	0.879	3.2	0.909	34.3
			0.4	4.8	0.685	2.3	0.537	30.8
	three	low	0.1	4.4	0.909	3.5	0.938	47.4
			0.2	4.0	0.623	2.2	0.514	38.4
			0.4	3.6	0.412	2.0	0.381	40.8
		medium	0.1	3.7	0.574	2.8	0.560	26.8
			0.2	3.8	0.392	2.3	0.395	35.6
			0.4	2.9	0.054	1.3	0.049	36.9
		high	0.1	4.9	0.620	2.9	0.671	56.2
			0.2	3.7	0.148	1.8	0.149	43.5
			0.4	2.7	0.006	1.3	0.002	38.0
Systematic	one	low	0.1	4.7	0.951	3.1	0.991	52.6
			0.2	4.4	0.944	3.0	0.962	28.7
			0.4	4.2	0.934	3.0	0.943	29.6
		medium	0.1	5.8	0.871	3.0	0.879	39.7
			0.2	4.3	0.785	2.5	0.694	31.3
			0.4	3.7	0.700	2.1	0.522	30.3
		high	0.1	8.5	0.822	4.2	0.914	42.6
			0.2	7.5	0.699	5.4	0.736	31.8
			0.4	7.0	0.530	5.5	0.626	29.6
	three	low	0.1	4.8	0.893	3.2	0.934	54.0
			0.2	3.9	0.656	2.1	0.518	33.5
			0.4	3.2	0.409	1.6	0.330	30.6
		medium	0.1	4.4	0.527	1.8	0.365	40.2
			0.2	3.5	0.233	1.4	0.228	42.4
			0.4	2.6	0.015	1.0	0.000	35.1
		high	0.1	8.8	0.306	6.2	0.410	38.0
			0.2	9.7	0.075	5.2	0.096	34.2
			0.4	4.3	0.062	2.3	0.074	16.5

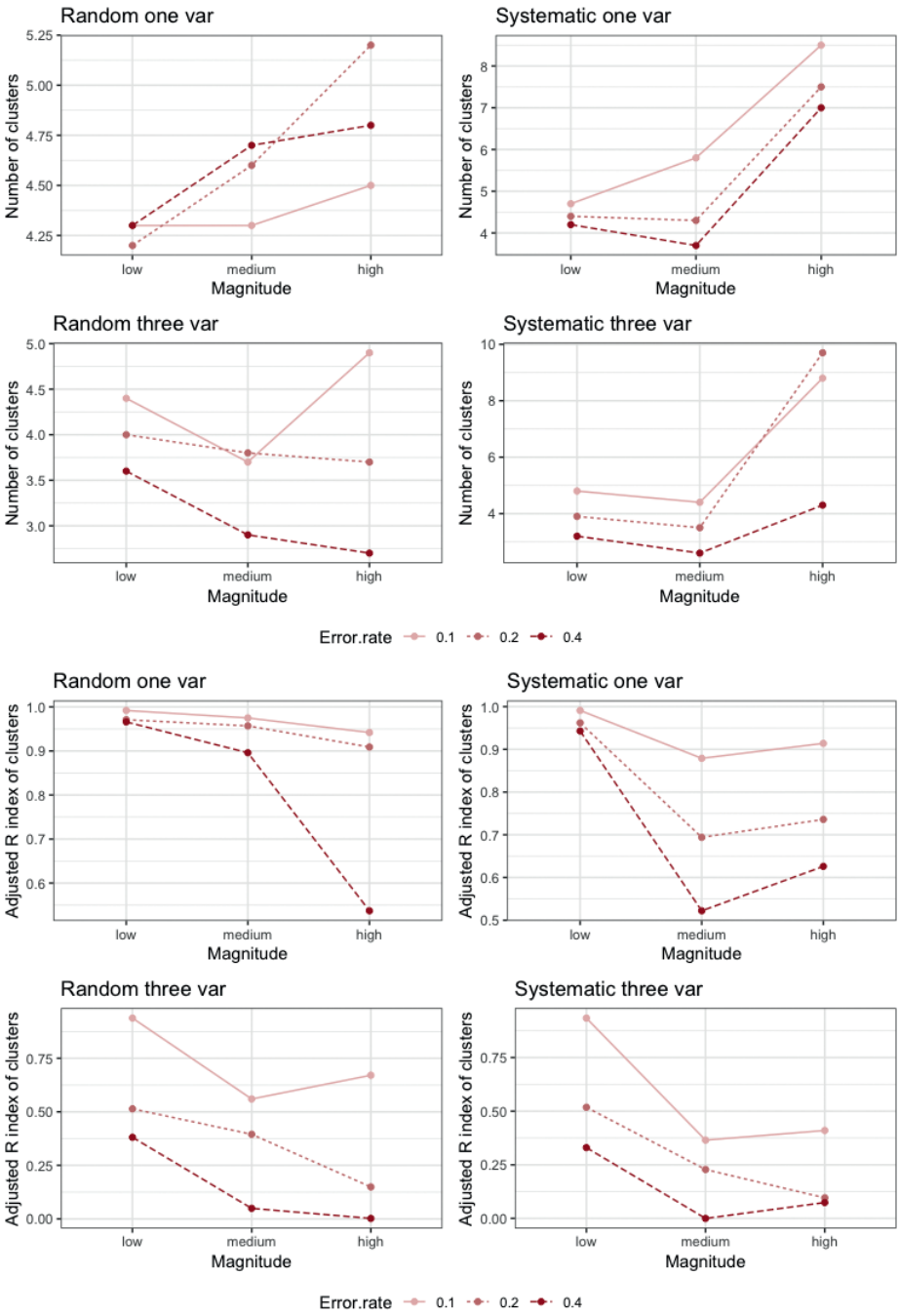


Figure 2.6- (top) Number of clusters and (bottom) cluster similarity for DBSCAN

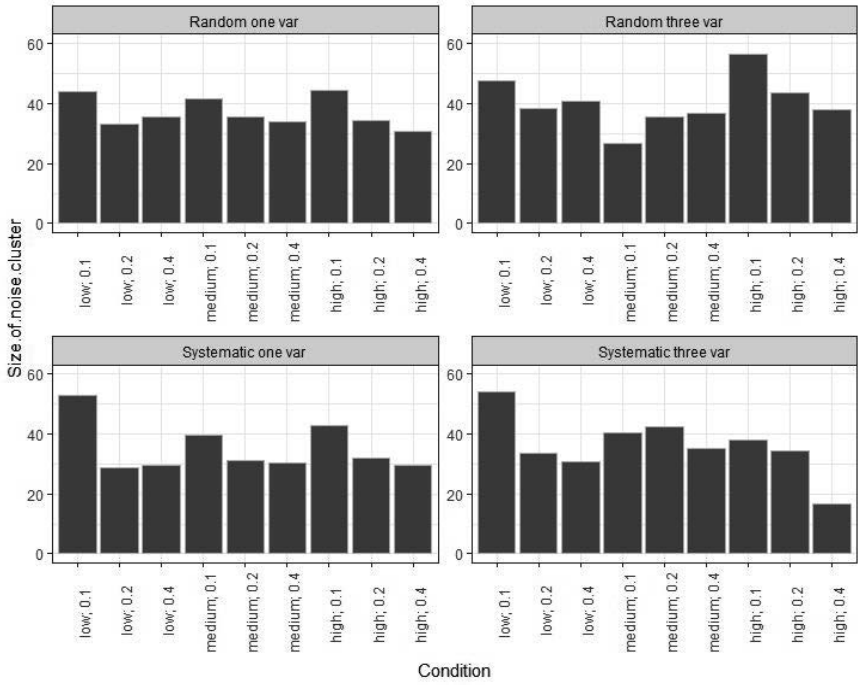


Figure 2.7- Noise cluster size by simulation condition (where the original noise cluster size = 36)

2.5 Discussion and conclusions

Clustering is a commonly applied method used in numerous disciplines that allows for the separation of observations into interesting groups, based on a predefined similarity measure. While an important and useful tool, this technique also suffers from an important shortcoming. Namely, in most cases, the clustering algorithms used disregard the problem of measurement error, which is both unrealistic and problematic. It is unrealistic as few, if any, data sources can be truly considered error-free, and it is problematic as measurement error is known to have the potential to severely bias estimates. In the context of clustering, measurement error can, for instance, produce spurious clusters or obscure clusters; it can also affect their shape, form, and stability.

Despite the threat that measurement error poses to the validity of clustering results, research available on the matter is scarce. Therefore, in this paper, we investigated the sensitivity of two commonly used model- and density-based clustering algorithms (i.e. GMMs and DBSCAN) to various types, severities, and levels of measurement error. In doing so, we examined how error affects the number of clusters found, the stability of the clusters, and their similarity to the clusters obtained in the absence of error.

Our results indicate that measurement error is particularly problematic and leads to unreliable clustering results when it is systematic as opposed to random, when it affects all (three) variables rather than only one, and, as expected, when its magnitude and/or rate is high. We also show that, overall, GMM is less sensitive to measurement error than DBSCAN, especially when looking at the merged clusters rather than the mixture components. DBSCAN appears highly sensitive to measurement error, in particular with regards to cluster (dis)similarity, regardless of whether all clusters or only stable clusters are considered. It also appears that, contrary to expectations, the noise cluster of the DBSCAN algorithm does not capture observations with measurement error.

The lower relative sensitivity of GMM estimates to measurement error is a rather surprising result. That is, while GMM can be viewed as the more restrictive clustering algorithm of the two (as, unlike DBSCAN, it makes an explicit assumption about the parametric form of the clusters), it seems to fare better in the presence of measurement error. These findings, however, should be treated with caution given the data structure of the simulated dataset. More specifically, in our analysis we simulated three almost perfectly spherical clusters and GMM algorithms are known to perform well when the clusters have such round shapes. DBSCAN, on the other hand, tends to be the preferred clustering method when the shapes of the clusters are arbitrary. Therefore, it is advisable to repeat the analysis using more complex data structures. This will also allow for the investigation of the impact of measurement error in a more realistic setup, as real-world data clusters tend to have various shapes and forms and are rarely perfectly separable.

It is also worthwhile mentioning that, while our analysis focuses on two important and popular types of clustering, it does not investigate the effect of measurement error on hierarchical clustering, a method which is widely used in particular in the social sciences. Therefore, future research should also examine how measurement error impacts such algorithms as Ward. We have also only focused on one type of systematic error, i.e. where the values of a variable are systematically over- or underestimated for some randomly selected subset of observations. It would be interesting to also look at how the two other types of systematic error, i.e. errors dependent on covariates or on the true value of the variable itself, affect clustering results.

Finally, given the strong potential implications of measurement error on clustering results, future research should also focus on investigating solutions that allow for the mitigation of its effects. Furthermore, new ways of performing error-aware clustering should consider the diverse nature of measurement error and account for both random and systematic type of errors.

Appendix 2.A Pseudocode illustrating the simulation design

Below we provide an example pseudocode illustrating the simulation design, which corresponds to the condition wherein all three variables contain systematic error that is small in magnitude and that affects 10 percent of the observations. The pseudocode includes the steps taken to simulate the "baseline" dataset and those taken to introduce measurement error according to the condition discussed above.

Step I: Simulate "baseline" dataset and perform clustering

1. Draw n_1 = 400 observations from the following MVN distribution:

$$G_1 \sim \boldsymbol{N}(\boldsymbol{\mu_1}, \boldsymbol{\Sigma_1}) \text{ where } \boldsymbol{\mu_1} = \begin{bmatrix} -2 \\ 9 \\ 12 \end{bmatrix} \text{ and } \boldsymbol{\Sigma_1} = \begin{bmatrix} 1.50 & 0.30 & 0.20 \\ 0.30 & 0.80 & 0.15 \\ 0.20 & 0.15 & 1.30 \end{bmatrix}$$

2. Draw n_2 = 400 observations from the following MVN distribution:

$$G_2 \sim \boldsymbol{N}(\boldsymbol{\mu_2}, \boldsymbol{\Sigma_2}) \text{ where } \boldsymbol{\mu_2} = \begin{bmatrix} 5 \\ 11 \\ 18 \end{bmatrix} \text{ and } \boldsymbol{\Sigma_2} = \begin{bmatrix} 2.00 & 0.40 & 0.15 \\ 0.40 & 1.60 & 0.25 \\ 0.15 & 0.25 & 1.00 \end{bmatrix}$$

3. Draw n_3 = 400 observations from the following MVN distribution:

$$G_3 \sim \boldsymbol{N}(\boldsymbol{\mu_3}, \boldsymbol{\Sigma_3}) \text{ where } \boldsymbol{\mu_3} = \begin{bmatrix} 4 \\ 4 \\ 5 \end{bmatrix} \text{ and } \boldsymbol{\Sigma_3} = \begin{bmatrix} 1.70 & 0.60 & 0.30 \\ 0.60 & 1.50 & 0.40 \\ 0.30 & 0.40 & 1.45 \end{bmatrix}$$

4. Perform clustering: fit GMM/ DBSCAN algorithms to the resultant dataset
   a. For GMM: fit models with number of clusters – k – varying from 1 to 10 and chose the model with lowest BIC
   b. For DBSCAN: set the minimum number of neighboring points to be four; choose the appropriate ε based on the k-nearest neighbor distance plot

Step II: Introduce measurement error into the "baseline" dataset and perform clustering (100 iterations)

5. Set the measurement error threshold[4] to 0.1 for all observations (that is t = 0.1)
6. For each observation in the dataset, draw a random number from a standard uniform distribution $-U_i \sim U(0,1)$
7. If $U_i \leq t$, add random draws to $X_{1,i}$,$X_{2,i}$ and $X_{3,i}$ from the following normal distributions $\mu_{x1} = 2.5$ *and* $\sigma_{x1} = 2$,$\mu_{x2} = -2.5$ *and* $\sigma_{x2} = 2$, *and* $\mu_{x3} = 1.25$ *and* $\sigma_{x3} = 2$
8. Perform clustering using GMM/DBSCAN (as described in (4))

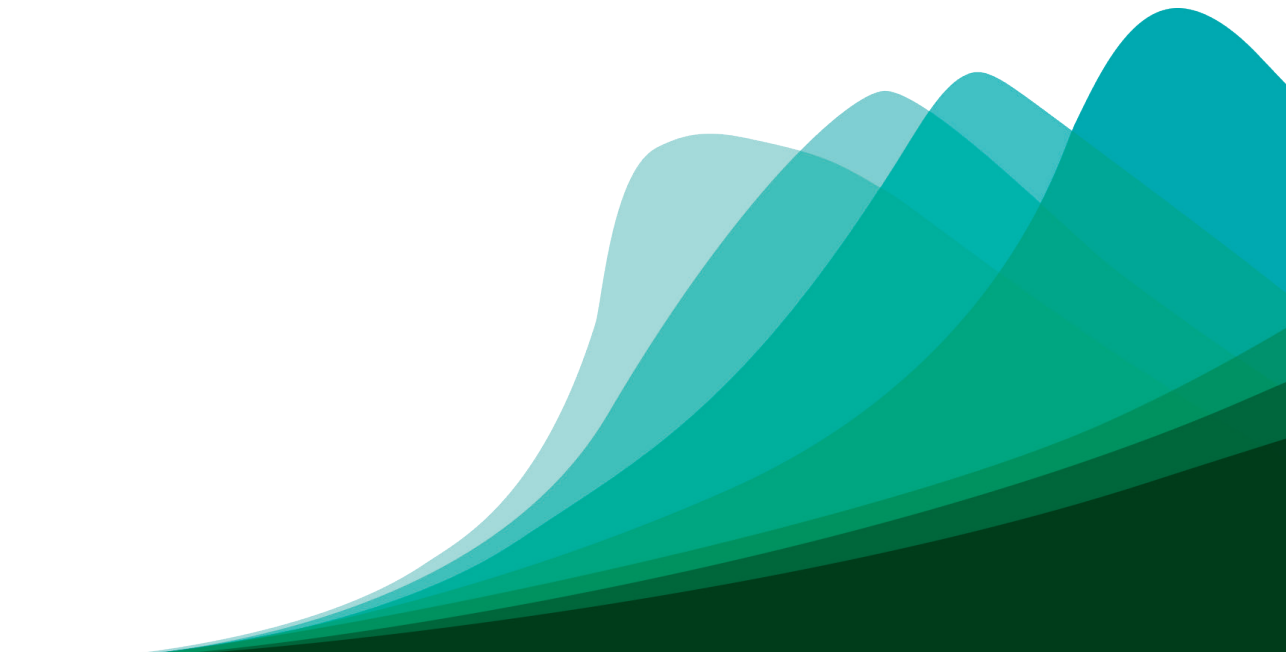
4 The threshold corresponds to the probability of being subject to measurement error.

a. For GMM: also merge mixture components into clusters using a threshold of 0.1 for the Bhattacharyya distance
b. FOR DBSCAN: also calculate cluster stability using a threshold of 0.7 for the Jaccard coefficient (50 iterations)

Step III: Compare clustering results in the absence and presence of measurement error

9. Compare clustering results obtained in (4) and (8)
   a. Compare number of clusters
   b. Compare cluster similarity using the Adjusted Rand Index

3



DEPENDENT INTERVIEWING: A REMEDY OR A CURSE FOR MEASUREMENT ERROR IN SURVEYS?

This chapter is based on: Pankowska, P., Bakker, B. F. M., Oberski, D. L.., & Pavlopoulos, D. (2019). Dependent interviewing: a remedy or a curse for measurement error in surveys. Manuscript submitted for publication (invited for resubmission after minor revisions).

Abstract

Longitudinal surveys often rely on dependent interviewing (DI) to decrease the level of random measurement error in survey data and reduce the incidence of spurious change. DI refers to a data collection technique that incorporates information from prior interview rounds into subsequent waves. While this method is considered an effective remedy for random measurement error, it can also introduce more systematic errors, in particular when respondents are first reminded of their previously provided answer and then asked about their current status. The aim of this paper is to assess the impact of DI on measurement error in employment mobility. We take advantage of a unique experimental situation that was created by the roll-out of dependent interviewing in the Dutch Labour Force Survey (LFS). We apply a hidden Markov model (HMM) to linked LFS and Employment Register (ER) data that cover a period before and after dependent interviewing was abolished, which in turn enables the modelling of systematic errors in the LFS data. Our results indicate that DI lowered the probability of obtaining random measurement error but had no significant effect on the systematic component of the error. The lack of a significant effect might be partially due to the fact that the baseline probability of repeating the same error was extremely high (i.e. when using standard, independent interviewing); therefore, the use of DI could not increase this probability any further.

3.1 Introduction

Measurement error in survey data is a well-known and well-documented phenomenon. A large volume of literature confirms that, if left unaccounted for, such error often biases estimates and can lead to inaccurate inferences and predictions (Alwin, 2007; Pankowska et al., 2018; Saris & Gallhofer, 2014; West & Blom, 2016). The magnitude of this problem is particularly high when using longitudinal survey data to estimate change or stability over time, as such second-order statistics have been shown to be severely affected by (random) measurement error (Bound et al., 2001; Fuller, 2009; Hagenaars, 1990, 1994; Van de Pol & De Leeuw, 1986). More specifically, when measurement error is random, observed over time changes are often inflated as they not only reflect true changes but also include changes in the error (Jäckle & Eckman, 2019). For this reason, survey methodologists have applied various tools to minimize the occurrence of measurement errors by improving data collection processes in longitudinal surveys (Groves et al., 2011).

One tool in particular that has been widely implemented is *dependent interviewing (DI)* — a method that uses prior information from responses provided in previous interview rounds to modify the phrasing and routing of questions in subsequent survey waves, as well as to facilitate within-interview edit checks (Jäckle, 2009; Jäckle et al., 2007; Mathiowetz & McGonagle, 2000). This interviewing technique has been widely implemented in various large-scale longitudinal surveys worldwide, such as the British Household Panel Survey (BHPS), the Dutch Labour Force Survey (LFS), and the US Current Population Survey (CPS) (Jäckle et al., 2007).

However, while DI is a promising tool to potentially reduce random measurement error, it also has some potential adverse implications for systematic measurement error, in particular when used proactively (Lynn et al., 2006). With *proactive dependent interviewing (PDI)*, the interviewer first reminds the respondents of the answer they provided in the previous round and then inquiries about their current state.[5] PDI has been shown to reduce spurious change (Hoogendoorn, 2004; Jäckle & Eckman, 2019; Lynn et al., 2006) as well as to lower the occurrence of the seam effect (Brüderl et al., 2017; Moore et al., 2009) — a phenomenon wherein between-wave change is overestimated while within-wave change is underestimated (Jäckle & Eckman, 2019). This interviewing setup, however, through various processes, might also lead to more autocorrelated errors (Eggs & Jäckle, 2015; Jäckle & Eckman, 2019). That is, when reminded of their previous answer, individuals might falsely confirm that this answer still holds. Such a false report of "no change" might lead to spurious stability if a true

5 DI can also be used reactively (RDI), whereby respondents are first asked the question independently and then, if an inconsistency is detected between the current and previous answer, a follow-up question is raised to verify whether a change occurred (Jäckle & Eckman, 2019; Uhrig & Sala, 2011). As RDI is primarily applied to numeric responses (Jäckle & Eckman, 2019) and is not expected to have strong implications for systematic error (Lynn et al., 2006), our paper focuses on the effect of PDI on measurement error.

change did occur. It can also lead to the copying over of an error across waves if no change occurred and the previously provided answer was wrong (Eggs & Jäckle, 2015; Hoogendoorn, 2004; Jäckle & Eckman, 2019).

Therefore, the overall effect of PDI on data quality appears uncertain and remains an issue for empirical investigation: on the one hand, this interviewing technique could reduce random error, but, on the other hand, it can increase the incidence of systematic error (as shown e.g. by Lugtig and Lensvelt-Mulders, 2014). From the perspective of substantive researchers, it appears that decreasing spurious change through the use of PDI might come at the expense of increasing spurious stability.

Given the two contradictory effects, and the lack of consensus in the literature regarding the overall utility of DI, this paper aims to disentangle the effect of PDI on random and systematic measurement error and in this way, to assess the overall effect of PDI on measurement error. For this purpose, rather than conducting our own experiment, we leverage the replacement of PDI with independent interviewing (INDI) which took place at the beginning of 2010 in the Dutch LFS. The questionnaire was changed as the routing in the former version was too complex, leading to mistakes in the interview. As no other major changes in the survey data collection process occurred in the time period under study, this change provides a natural experiment, which allows for an investigation of the impact of PDI on measurement error while treating independent interviewing (INDI) as the counterfactual.

To assess the magnitude of measurement error in the corresponding survey question, we use hidden Markov models (HMMs), a group of latent class models that allow for the estimation and correction of measurement error in categorical, longitudinal data, provided that the model is specified correctly (Biemer, 2004; Pankowska et al., 2018; Pavlopoulos & Vermunt, 2015). The main advantage of these models is that they do not require a “gold-standard”, error-free data source, which would serve as a benchmark for the survey data (Biemer, 2011; Vermunt & Magidson, 2002). To model systematic measurement error in the survey data without having to impose unwanted restrictions and risk poor identifiability, we use an extended, two-indicator version of HMMs (Bassi et al., 2000). These two indicators are obtained by linking data from the Dutch LFS and the Dutch Employment Register (ER).

The remainder of the paper is structured as follows: section 3.2 elaborates further on the use PDI and its effects on random and systematic measurement error; it then describes the roll-out of PDI in the Dutch LFS. Section 3.3 discusses the use of HMMs to assess and correct for measurement error as well as the model and data used in the analysis. Section 3.4 discusses the results obtained and, finally, section 3.5 offers concluding remarks.

3.2 Dependent interviewing (DI) and its effect on measurement error

3.2.1 Background

Dependent interviewing (DI) is an interviewing technique in which information provided by a respondent in prior interview rounds is used in subsequent waves; when using DI proactively, the wording of the question is tailored based on the previously provided response(s) (Jäckle, 2009). In this design, interviewees can be asked the question in three distinct manners: in *"remind, continue"* respondents are reminded of their previous answer and then asked the standard independent question; in *"remind, still"* they are asked whether the situation described still holds; in *"remind, confirm"* interviewees are asked to confirm whether their previous response is correct (Hoogendoorn, 2004; Jäckle, 2008, 2009; Jäckle et al., 2007; Jäckle & Eckman, 2019; Lugtig & Lensvelt-Mulders, 2014; Mathiowetz & McGonagle, 2000).

PDI is used in longitudinal surveys for two main reasons: (i) it has the potential to improve data quality by achieving higher longitudinal consistency and lower levels of random error (Jäckle, 2009; Mathiowetz & McGonagle, 2000) and (ii) it can increase survey efficiency and reduce respondent burden (Eggs & Jäckle, 2015; Jäckle, 2008). The importance of improving data quality is related to the fact that, as mentioned previously, longitudinal surveys in most cases suffer from random measurement error, which has the potential to severely inflate change estimates (Jäckle, 2009; Jäckle & Lynn, 2007; Lugtig & Lensvelt-Mulders, 2014; Lynn et al., 2006; Van de Pol & De Leeuw, 1986). Previous studies show that PDI has been effective in reducing spurious change and the seam effect in numerous different panel surveys (Jäckle & Eckman, 2019). The need to increase the efficiency of the interviewing process and to reduce respondent burden is tied to common complaints made by interviewees about having to answer the same question recurrently even when their circumstances have not changed. PDI reduces the need to repeatedly answer the same question and thus is thought to reduce respondent burden. Furthermore, tailoring the question to the respondents' specific situation and reminding them of their previously provided answers was shown to improve the flow of the interview and simplify the response task (Sala et al., 2011). These efficiency gains have also been linked to lower rates of (random) measurement error (Hoogendoorn, 2004; Jäckle, 2009; Lynn et al., 2006).

Overall, PDI is potentially an effective technique that can address several challenges faced by survey methodologists when dealing with repeated longitudinal surveys; however, it is not free of shortcomings, as there is some concern that PDI might lead to more systematic measurement through two main mechanisms. First, PDI might increase the incidence of error due to the phenomenon of *(cognitive) satisficing*, wherein respondents, rather than providing a well-thought-out, appropriate answer, tend to opt for the easy, credible response. In the context of PDI, this would imply falsely confirming that the previous answer still holds (Eggs & Jäckle, 2015; Hoogendoorn, 2004; Jäckle & Eckman, 2019; Lugtig & Lensvelt-Mulders, 2014). Second, PDI might also

have an adverse effect on the error due to the presence of *motivated misreporting*, a phenomenon whereby individuals, to shorten the duration of the interview, provide inaccurate answers that allow them to omit follow-up questions. This implies that when using PDI respondents will be inclined to report that the previous information still holds, as this will likely allow them to skip questions about their current state (Eggs & Jäckle, 2015).

Therefore, while there is some consensus that PDI improves survey efficiency and reduces respondent burden, its effect on data quality in general and measurement error in particular remains ambiguous. In short, lower levels of random error may come at the cost of higher probability of systematic error. Therefore, this paper investigates the nature of this relationship by examining the effects of the PDI, *"remind, still"* design on the measurement of the contract type question in the Dutch LFS.

3.2.2 Dependent interviewing (DI) in the Dutch Labour Force Survey (LFS)

The Dutch Labour Force Survey (LFS) is an address-based sample survey that provides information on the labor market characteristics of individuals residing in the Netherlands. It is carried out by Statistics Netherlands and, as of the end of 1999, it is a quarterly rotating panel survey consisting of five waves. For the contract related question (which is the focus of our analysis), dependent interviewing (DI), and more specifically the *"remind, still"* style of proactive DI (PDI), was in use in the LFS from the beginning until the end of 2009; at the beginning of 2010 it was replaced by independent interviewing (INDI). Survey respondents were asked about their employment contract using PDI if they met two conditions: (i) they reported in the previous wave that they had a temporary contract and (ii) they indicated that they did not change their job since the previous wave. Respondents who fulfilled both criteria were asked the following question regarding their contract type: *"Last time you were in temporary employment. Is this still the case?"*[6]. Individuals who changed jobs or those who did not experience a job change but had indicated previously that they had *"other"* position on the labor market (i.e. those that were not in paid employment) were asked the question in an independent fashion as follows: *"Are you currently in permanent employment?"*[7]. The contract question was skipped for respondents who in the previous wave reported having a permanent contract and who did not experience a job change; instead, these individuals' responses from the previous wave were copied forward.[8]

This setup, which is summarized in the flowchart of Figure 3.1, results in three possible scenarios: (i) an individual is subject to INDI if either (a) they indicated that a job change occurred, (b) they reported having "other" type of employment in the previous survey round *or*, (c) they first participated in the LFS after the end of 2009; (ii) an individual is asked the contract question using PDI if they (a) did not change their

6 In Dutch: "De vorige keer was u in tijdelijke dienst. Is dat nog zo?"

7 In Dutch: "Bent u op dit moment in vaste dienst?"

8 https://www.cbs.nl/en-gb/our-services/methods/surveys/korteonderzoeksbeschrijvingen/dutch-labour-force-survey–lfs–

job since the previous survey wave, (b) they reported being employed on a temporary basis in the previous round and, (c) they first took part in the survey before the end of 2009; (iii) the contract question is not asked altogether if (a) no job change occurred and (b) the individual previously reported being employed permanently.

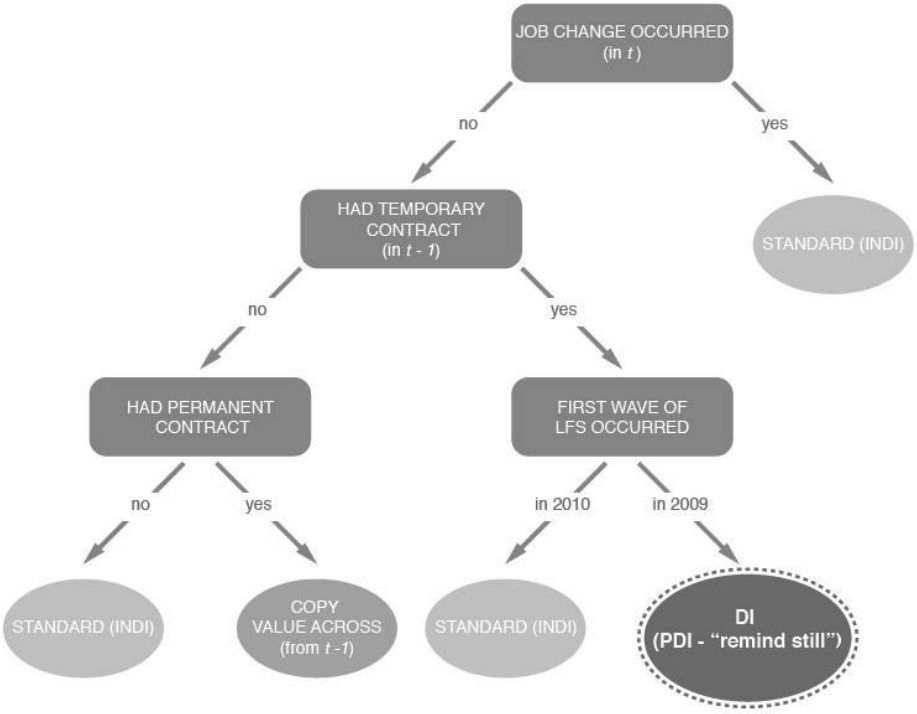


Figure 3.1- Summary of the interviewing setup in the LFS contract question

While all three scenarios occur in our dataset, our analysis focuses on comparing the levels of random and systematic errors when PDI was used with those when INDI was used, but PDI would had been applicable if it was not abolished at the end of 2009. This allows us to take advantage of the natural experiment setup caused by the replacement of PDI with INDI at the beginning of 2010. More specifically, we compare scenario ii (3.31 percent of cases in our sample), which we refer to as the treatment group, to a subset of scenario iii (2.28 percent of cases in our sample) - wherein no job change occurred, a temporary contract was reported in the previous wave, and the first round of the LFS was conducted after the end of 2009 - which we refer to as the *counterfactual* or the *control group*.

3.3 Methodology

3.3.1 Assessing and correcting for measurement error using hidden Markov models (HMMs)

Hidden Markov models (HMMs) are a latent variable modelling technique that can be applied to evaluate measurement error in categorical longitudinal survey data (Biemer, 2011; Pankowska et al., 2018; Pavlopoulos & Vermunt, 2015). Their rise in popularity can be attributed to the fact that, unlike other commonly used error assessment methods, they do not require the availability of error-free, "gold-standard" validation data that are most often unattainable in practice (Biemer & Wiesen, 2002; Pankowska et al., 2020). In this context, HMMs are used when the (dynamic) quantity of interest, e.g. over-time employment transitions, is measured in the panel survey with some degree of error. The models allow for the separation of true change from measurement error which, in turn, can produce error-corrected estimates of the quantity of interest as well as assessing the level of measurement error in the corresponding survey question (Biemer, 2011; Pankowska et al., 2018).

The standard HMM, which can be fit to surveys with at least three panel waves, consists of two components: (i) the structural component that models the true (latent) initial state probabilities X_0 and the true (latent) transition probabilities between X_{t-1} and X_t, where t = 1, ..., T ; and (ii) the measurement component that models the interactions of the survey observations (which contain error) A_t with the true values X_t at each wave t = 1, ..., T. The two components are estimated simultaneously. The model relies on two basic assumptions: first, the probability of a specific value of X occurring at time t only depends on its value in the previous time point, X_{t-1} – the so-called *Markov assumption*. This assumption can be stated formally as follows:

$$Pr(X_t = x_t | X_1 = x_1, \dots, X_{t-1} = x_{t-1}) = Pr(X_t = x_t | X_{t-1} = x_{t-1}) \quad (3.1)$$

where $Pr(X_t = x_t)$ denotes the probability of the latent state X_t taking on a specific value x_t out of k possible categories. Second, the probability of observing a specific value of A at time t only depends on the true value at the same time point – X_t – the so-called *local independence assumption* or — using a term that is more appropriate for longitudinal data — *independent classification error (ICE) assumption*. This assumption can be stated formally, as follows:

$$Pr(A_1 = a_1, \dots, A_T = a_T | X_1 = x_1, \dots, X_T = x_T) = \prod_{t=1}^{T} Pr(A_t = a_t | X_t = x_t) \quad (3.2)$$

where $Pr(A_1 = a_1, \dots, A_T = a_T)$ denotes the probability of observing a specific path or sequence of survey states, where each state – $A_1, \dots, A_T$ – takes on a specific value – $a_1, \dots, a_T$ – out of k possible categories. Combining the Markov and local independence

assumptions leads to the following probability of observing a certain path $A = (A_1, ..., A_T)$ in the survey data:

$$Pr(A = a) = \sum_{x_0=1}^{k} \dots \sum_{x_T=1}^{k} Pr(X_0 = x_0) \prod_{t=1}^{T} Pr(X_t = x_t | X_{t-1} = x_{t-1}) \prod_{t=1}^{T} Pr(A_t = a_t | X_t = x_t) \quad (3.3)$$

Where $Pr(X_0 = x_0)$ represents the initial state latent probabilities and $Pr(X_t = x_t | X_{t-1} = x_{t-1})$ represents the latent transition probabilities, which follow a first-order Markov process. $Pr(A_t = a_t | X_t = x_t)$ denotes the classification error (also referred to as emission) probabilities, which satisfy the local independence assumption and are used to estimate question reliability in surveys (Bassi et al., 2000; Biemer, 2011; Pankowska et al., 2018, 2020; Pavlopoulos & Vermunt, 2015).

If only three time-points are available, in addition to the two assumptions specified above, further restrictions in the form of time-invariant/constant misclassification (measurement error) rates and latent transitions rates are required to obtain model identification (Pankowska et al., 2018, 2020; Van de Pol & De Leeuw, 1986). Given these assumptions and restrictions, which are required to obtain identifiability, the standard, one-indicator HMM can be seen as rather limited in its capacity to model realistic error scenarios. While it is possible to relax some of the assumptions when using richer survey data (i.e. with more than three data points), the practical applicability of the model remains rather limited. To illustrate, even with multiple ($t > 3$) survey waves, one cannot simultaneously model both local dependence, which allows for the occurrence of systematic error, as well as time-varying measurement and/or structural parameters. It is worthwhile noting that, even models that only account for the occurrence of systematic error often suffer from identifiability issues (i.e. are "poorly identifiable"). As a result of these limitations, survey researchers have increasingly started using extended, multiple-indicators versions of the standard HMM, which are more flexible and allow for model specifications that are more reflective of reality (Pankowska et al., 2018, 2020; Pavlopoulos & Vermunt, 2015).

A basic two-indicator HMM, which can be obtained, for instance, by linking survey data to register/administrative records, has the following probability of observing certain paths $A = (A_1, ..., A_T)$ and B = $(B_1, ..., B_T)$:

$$\begin{aligned} Pr(A = a, B = b) = \sum_{x_0=1}^{k} \dots \sum_{x_T=1}^{k} Pr(X_0 = x_0) \prod_{t=1}^{T} Pr(X_t = x_t | X_{t-1} = x_{t-1}) \\ \prod_{t=1}^{T} Pr(A_t = a_t | X_t = x_t) \prod_{t=1}^{T} Pr(B_t = b_t | X_t = x_t) \end{aligned} \quad (3.4)$$

Where the latent initial state probabilities — $Pr(X_0 = x_0)$, the latent transition probabilities — $Pr(X_0 = x_0)$, and the survey emission probabilities — $Pr(A_t = a_t | X_t = x_t)$ — are

specified in the same way as in the univariate/one-indicator HMM described above. This extended specification also includes the register emission probabilities — $Pr(B_t = b_t | X_t = x_t)$ — that, in a similar way to the survey emission probabilities, also satisfy the local independence assumption. While this is the most basic two-indicator HMM specification, the model can be easily extended further by e.g. (i) accounting for (un)observed heterogeneity and time dependency in the transition and/ or emission probabilities and (ii) relaxing the local independence assumption for the survey and/ or register data.

3.3.2 Data and the empirical model

In our analysis, we make use of an extended HMM specification with two indicators that come from two independent data sources (i.e. the Dutch LFS and Employment Register). Such a specification allows us to model the possibility that PDI leads to more systematic error in the survey data and, at the same time, allowing the latent transition probabilities to depend on time and personal characteristics (following Pankowska et al., 2018 and Pavlopoulos and Vermunt, 2015).[9]

To obtain two indicators, we link the LFS data to records from the Dutch Employment Register (ER). The ER is an administrative dataset that combines information from various sources but predominantly consists of tax related data provided to the Dutch Tax Authorities by employers. It is managed by the Dutch Employee Insurance Agency (UWV) and contains monthly information for all insured employees in the Netherlands on such individual-level characteristics as wages, benefits, and labor relations.[10] The record linkage is performed at the individual level and the data from both sources are linked to the population register (PR) of the Netherlands. For the LFS, the linkage key is based on a combination of birth date, gender, postal code and house number. For the ER it is based on the social security number (BSN),[11] birth date, gender, postal code and house number. The linkage effectiveness of this procedure, i.e. the percentage of linked records in the LFS, is estimated by Statistics Netherlands to be around 98 percent. In the following we will assume perfect record linkage. Previous research has shown that even if there is linkage error, its effects on the estimates of HMMs is negligible unless this linkage error is large and strongly correlated with the process of interest (Pankowska et al., 2020). This is definitely not the case in our data.

9 While we also considered a model specification which allows for time-varying (systematic) error parameters (i.e. wave-heterogeneous question reliability), the fit of this model was significantly worse than of the one with time-invariant measurement parameters. Therefore, the findings we discuss in the results section are based on a model assuming constant reliability.

10 https://www.cbs.nl/nl-nl/achtergrond/2010/35/polisadministratie

11 A unique personal number allocated to everyone registered in the Netherland; https://www.government.nl/topics/identificationdocuments/contents/the-citizen-service-number

Table 3.1- Distribution of observations by DI eligibility (LFS year, job change in t and contract type in t-1) (N = 430,375; in %)

Job change	LFS contract at t-1					
	2009			2010		
	Permanent	Temporary	Other	Permanent	Temporary	Other
Yes	0.46	0.23	0.62	0.23	0.13	0.44
No	55.15	**3.31**	0.33	36.36	**2.28**	0.46

Note: The percentages correspond to the shares of specific groups in the overall sample and are calculated by dividing the number of individuals who fulfil the respective criteria by the overall sample size; the percentages of treatment and control groups are provided in bold

Our linked dataset consists of 86,075 LFS respondents of prime working age (i.e. 25 to 55 years old) who first participated in the survey either in 2009 (PDI in place) or 2010 (PDI abolished). It contains quarterly information on each individual for 5 time points, leading to a total sample size of 430,375 observations. Both the survey and register data are subject to item and unit nonresponse; we assume all missing values to be missing at random (MAR) given the model (Little & Rubin, 2019). Table 3.1 provides the distribution of observations by the conditions determining PDI eligibility. As can be seen from the table, overall PDI was used in a rather small fraction of the sample. That is, in approx. 3.3 percent of the cases, individuals were asked the question in a PDI fashion (i.e. 3.3 percent of the observations belong to the treatment group); in around 2.3 percent of the cases, PDI would have been used if it were not abolished (i.e. 2.3 percent belong to the control group/counterfactual).

In this linked survey and register dataset, the probability of observing particular employment contract paths — A and B —which depend on observed individual-level heterogeneity (Z) and the interviewing regime used (W), according to our two-indicator HMM, can be formalized as follows:

$$Pr(A=a,B=b|Z,W)=\sum_{x_0=1}^{k}\dots\sum_{x_T=1}^{k}Pr(X_0=x_0|Z)\prod_{t=1}^{T}Pr(X_t=x_t|X_{t-1}=x_{t-1},Z)$$
$$\prod_{t=1}^{T}Pr(A_t=a_t|X_t=x_t,X_{t-1}=x_{t-1},A_{t-1}=a_{t-1},W) \quad (3.5)$$
$$\prod_{t=1}^{T}Pr(B_t=b_t|X_t=x_tX_{t-1}=x_{t-1},B_{t-1}=b_{t-1})$$

where the (latent) initial state probabilities and transition rates — $Pr(X_0 = x_0 \mid Z)$ *and* $Pr(X_t = x_t \mid X_{t-1} = x_{t-1}, Z)$ — depend on observed individual level heterogeneity (i.e. the covariates education, gender and ethnicity) and the latent transitions also depend on time (i.e. are time-heterogeneous and depend on t and t^2). The inclusion of covariates in the structural part of the model implies that the Markov assumption holds conditional

on these covariates. The emission probabilities for both the survey and register data — $Pr(A_t = a_t | X_t = x_t, X_{t-1} = x_{t-1}, A_{t-1} = a_{t-1}, W)$ and $Pr(B_t = b_t | X_t = x_t, X_{t-1} = x_{t-1}, B_{t-1} = b_{t-1})$ — relax the local independence assumption allowing for systematic error in both data sources. In more detail, for both the LFS and the ER, we allow the error probabilities to also depend on the lagged true contract — X_{t-1} — and the lagged observed contract — A_{t-1} or B_{t-1}. Additionally, to compare the error levels under PDI and INDI, the emission probabilities also depend on the covariate W, which determines the interviewing regime used and can take 3 values:

- 0 (ref. category) INDI was used but PDI would have been used if it was not abolished;
- 1 INDI was used and would have been used regardless of whether DI had been abolished;
- 2 PDI was used.

In our analysis, we focus on comparing the error levels under PDI to those where PDI would have been used (i.e. category 2 vs. 0).

For the survey data, we use a restricted model that only allows for systematic error in situations where the errors are a consequence of the phenomena of *satisficing* and/or *motivated misreporting*. Specifically, the LFS log-linear error parameters take the following form $\alpha_{at,xt} + \beta_{at,at-1,xt,xt-1} + \alpha_{at,xt,w} + \beta_{at,at-1,xt,xt-1,w}$. In this specification, the term $\alpha_{at,xt} + \alpha_{at,xt,w}$ represents the random component of the error while the term $\beta_{at,at-1,xt,xt-1} + \beta_{at,at-1,xt,xt-1,w}$ represents the systematic component of the error. In both cases, the first parts of the expression (i.e. $\alpha_{at,xt}$ and $\beta_{at,at-1,xt,xt-1}$) represent the "baseline" random/systematic error probability while the second parts (i.e. $\alpha_{at,xt,w}$ and $\beta_{at,at-1,xt,xt-1,w}$) indicate how the use of different interviewing regimes affects the probabilities of obtaining random/systematic error. Put simply, to compare the effect of using PDI as opposed to standard INDI, we estimate additional random and systematic error parameters for when the contract question was asked in a PDI fashion. The parameters of the systematic error components are freed when the same error can be repeated due to the *"remind, still"* PDI — i.e. when $A_t = A_{t-1} = temp \neq X_t = X_{t-1} = \{perm, other\}$ or when it might cause spurious stability; that is, in a situation where an individual correctly reports having a temporary contract in t - 1, then experiences a true transition between t - 1 and t but erroneously confirms in t that she/he is still employed on a temporary basis — i.e. when $A_{t-1} = X_{t-1} = temp$ & $A_t = temp \neq X_t = \{perm, other\}$. In all other instances the systematic error parameters are set to 0.

For the register data, we only allow for the repetition of the same error over time, as previous research has shown the ER to suffer predominantly from this type of error (Pankowska et al., 2018; Pavlopoulos & Vermunt, 2015). That is, for the error parameters — $\alpha_{bt,xt} + \beta_{bt,bt-1,xt,xt-1}$ — we estimate the systematic component — $\beta_{bt,bt-1,xt,xt-1}$ — in situations where $B_t = B_{t-1} \neq X_t = X_{t-1}$; in all other cases we set this component to 0. Appendix 3.A lists all possible systematic error parameters (i.e. those estimated and those restricted to 0) and specifies which ones were freed and which ones were set to 0 in both the LFS and the ER data.

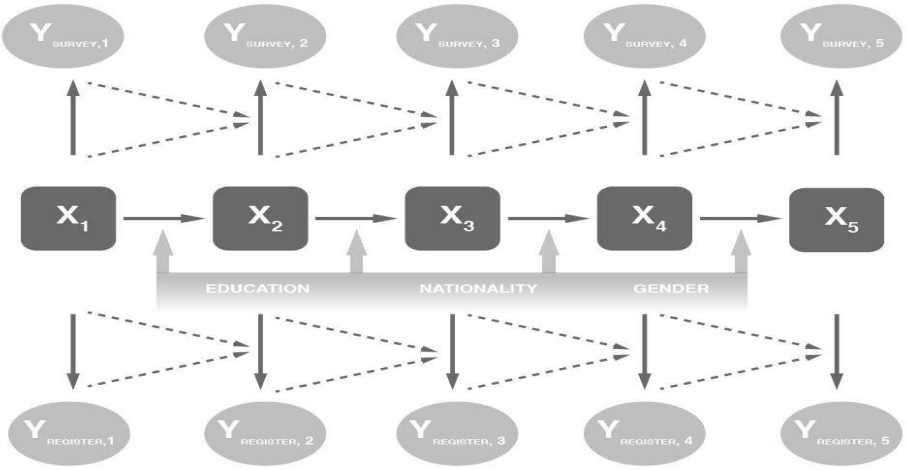


Figure 3.2- Path diagram of the two-indicator HMM with serially correlated error in the survey and register data and covariate dependent latent initial state and transition probabilities

In our model, k runs from 1 to 3 and represents the number of contract type categories {permanent, temporary, other}; T runs from 1 to 5 and corresponds to the months in which the (quarterly) survey took place. The model is estimated in the Latent GOLD software (version 4.5), using the Baum-Welch algorithm, which is an adapted expectation-maximization (EM) procedure (for further details about this process see McLachlan and Krishnan, 2008 and Pankowska et al., 2020). A path diagram of the model is provided in Figure 3.2.

3.4 Results

In this section, we first investigate whether the use of PDI, as shown by previous studies, indeed lowers the occurrence of random measurement error. We then look at whether, as hypothesized, PDI also leads to higher incidence of systematic error. In doing so, we compare the corresponding measurement error parameter estimates obtained when (i) PDI was used in 2009 to (ii) those obtained when INDI was applied in 2010 to cases that would have been eligible for PDI had it not been abolished. To reiterate, both scenarios include observations in which LFS respondents in t - 1 reported having a temporary contract and in t stated that they did not change their job. Therefore, all of these individuals fulfilled the criteria for PDI. However, only those who first participated in the survey in 2009 were actually asked the question in a PDI fashion; individuals who started the LFS in 2010 were subject to INDI. Table 3.2 presents the estimates of the random error parameters under PDI, where the reference category is *"INDI would have been PDI"*. When investigating the effect of PDI on random error, we estimated four additional error parameters when PDI is used (compared to INDI): reporting temporary in the LFS given that the true contract is permanent or other, and reporting permanent

or other given it is temporary. The remaining two parameters (permanent — other and other — permanent) were restricted to 0 as PDI was specifically applied when a temporary contract was reported and, therefore, should not have any effect in these two error scenarios.

Table 3.2- Random measurement error parameter estimates

LFS contract	True contract	Log-linear parameter	S.E.	Sig.
Temporary	Permanent	10.25	12.52	0.41
Permanent	**Temporary**	**-0.64**	**0.10**	**0.00**
Other	**Temporary**	**-0.47**	**0.20**	**0.02**
Temporary	Other	18.44	17.90	0.30

As can be seen from Table 3.2, the use of PDI in the LFS reduced the occurrence of random measurement error in instances where respondents erroneously reported to hold a permanent or "other" type of contract while in reality they were employed on a temporary basis (β = -0.64, p = 0.00 and β = -0.47, p = 0.02, respectively). More specifically, when asked the question in a PDI fashion compared to INDI, an LFS respondent, whose true contract at time *t* is temporary, is almost twice less likely to falsely report having a permanent contract (OR = 1.90) and slightly over 1.5 times less likely to report having "other" type of contract (OR = 1.60).

The probabilities of misreporting a contract type as temporary while in reality it is either permanent or other seem unaffected by PDI (β = 10.25, p = 0.41 and β = 18.44, p = 0.30, respectively).[12] The lack of significant effects when the true contract type is either permanent or other is to be expected given how this interviewing technique was set up in the LFS and given the eligibility criteria for PDI. More specifically, as individuals are only subject to PDI if they reported having a temporary contract in the previous wave, PDI will only decrease the probability of misreporting a true temporary contract as permanent or "other".

Table 3.3- Systematic measurement error parameter estimates

LFS contract (in t)	LFS contract (in t-1)	True contract (in t)	True contract (in t-1)	Log-linear parameter	S.E.	Sig.
Temporary	Temporary	Permanent	Permanent	-9.45	12.53	0.45
Temporary	Temporary	Other	Other	19.43	17.98	0.28
Temporary	Temporary	Permanent	Temporary	2.23	8.37	0.79
Temporary	Temporary	Other	Temporary	23.03	17.90	0.69

12 It is worthwhile mentioning that the very high coefficient estimates in this case are caused by the fact that the baseline probabilities (i.e. under INDI) of observing temporary given that the true contract type is either permanent or `other' are extremely low. Therefore, even a small increase in these probabilities in absolute terms can have a substantial relative effect.

To assess whether DI leads to higher rates of systematic error, we examine the parameter estimates that correspond to situations (i) where the erroneous reporting of a temporary contract can be repeated, and (ii) where the reporting of temporary contract is correct in t - 1 but then becomes incorrect in t due to a true transition that was not reported. As can be inferred from Table 3.3, which provides the corresponding parameter estimates, PDI does not seem to increase the probability of obtaining systematic error. It appears that PDI leads to neither error autocorrelation nor to spurious stability (i.e. falsely confirming the previously reported answer still holds while a true change occurred).

In more detail, the error parameter estimates corresponding to a situation whereby an individual falsely reports having a temporary contract in t - 1 and t while in both time points the true contract type is either permanent or other are insignificant (β = -9.45, p = 0.45 and β = -19.43, p = 0.28, respectively). Similarly, the probabilities of correctly reporting a temporary contract in t - 1, but failing to report a true transition to either permanent or temporary employment in t (and confirming to still hold a temporary contract instead) also seem unaffected by PDI (β = 2.23, p = 0.79 and β = 23.03, p = 0.69, respectively).[13] The lack of an effect on the systematic component of the error, in particular for the scenarios whereby the same error can be repeated, might be due to the fact that even at baseline (i.e. when using standard INDI), there is an extremely high probability of an LFS respondent repeating the same error if no true change occurred (i.e. β = 13.6, p = 0.03 when $LFS_t = LFS_{t-1} = temp \neq TRUE_t = TRUE_{t-1} = perm$ and β = 19.5, p = 0.03 when $LFS_t = LFS_{t-1} = temp \neq TRUE_t = TRUE_{t-1} = other$). These parameter estimates correspond to a probability of over 0.99; therefore, the use of PDI cannot increase the probabilities of repeating the error any further (i.e. there seems to be a ceiling effect). This result is not particularly surprising given the short gaps between the waves in the LFS. That is, any misreporting of a contract due to, for instance, confusion is likely to persist over a relatively short period such as three months, provided that no actual change occurred.

3.5 Conclusions and discussion

DI is an interviewing technique that is broadly applied in panel surveys to achieve higher longitudinal consistency and lower levels of random measurement error. The importance of minimizing random error in this context stems from the fact that longitudinal survey data are often used to study over time change or transitions; such second-order statistics are known to be highly sensitive to random measurement error. However, while DI helps to mitigate this problem, it potentially introduces a new one, in particular when used proactively, as it has also been hypothesized to increase

13 Again, the large coefficient estimates are caused by the fact that at the baseline (i.e. for INDI) these probabilities are either extremely high or extremely low.

the incidence of systematic error due to the phenomena of *cognitive satisficing* and *motivated misreporting*.

Given the potentially conflicting effects of PDI on survey data quality, in this paper we examined the effect of this interviewing technique on both the random and systematic components of the error. Our results confirm that PDI reduces the incidence of random error. On the other hand, we find no evidence for the claim that systematic measurement error is increased due to PDI. To restate, PDI in the LFS is associated with lower probabilities of misreporting a true temporary contract as permanent or other type of contract but it is not associated with higher probabilities of repeating the same error over time and it does not lead to spurious stability (i.e. not reporting a true change).

Thus, overall, in our case PDI appears to have a positive effect on data quality as it reduces random error while leaving the systematic component of the error unaffected. It can be seen, therefore, as a useful interviewing technique that helps to tackle the problem of spurious change. However, it is important to note that in our analysis the probability of repeating the same error was over 0.99, regardless of the interviewing regime (i.e. also in the absence of DI). These results indicate that the level of this systematic error was already so extreme in the Dutch LFS that the use of PDI could not have increased its magnitude any further (i.e. a ceiling effect had occurred). It is therefore possible that DI would have had a significant effect on the systematic component of the error had the baseline probability not been this high. Despite this limitation, the paper provides important findings for survey methodologists and designers of survey questionnaires. PDI is shown to be an attractive option for obtaining information on categorical characteristics in longitudinal surveys as it reduces random measurement error and, in cases where systematic error is high even with independent interviewing, PDI does not increase it any further. Therefore, in such cases, PDI reduces measurement error overall and can be a helpful tool in surveys.

It is important to note that, in our sample, the change from PDI to INDI affected a relatively small percentage of records (i.e. 5 percent). Therefore, future research should investigate the impact of changes in the interviewing regime on measurement error, when a greater proportion of the population is affected by these changes. When examining their impact, it is also worth going beyond the specific type of PDI used in our analysis to see whether the remaining two types of this interviewing method, i.e. *"remind, continue"* and *"remind, confirm"*, have similar effects on the quality of the survey data.

Appendix 3.A List of systematic error parameters in the LFS and ER

Observed contract in $t-1$	Latent contract in t	Latent contract in $t-1$	Observed contract in t		
			Permanent	Temporary	Other
Permanent	Permanent	Permanent			
Permanent	Permanent	Temporary			
Permanent	Permanent	Other			
Permanent	Temporary	Permanent			
Permanent	Temporary	Temporary	$\beta_{b_t,b_{t-1},x_t,x_{t-1}}$		
Permanent	Temporary	Other			
Permanent	Other	Permanent			
Permanent	Other	Temporary			
Permanent	Other	Other	$\beta_{b_t,b_{t-1},x_t,x_{t-1}}$		
Temporary	Permanent	Permanent		$\beta_{a_t,a_{t-1},x_t,x_{t-1}}/\beta_{b_t,b_{t-1},x_t,x_{t-1}}$	
Temporary	Permanent	Temporary		$\beta_{a_t,a_{t-1},x_t,x_{t-1}}$	
Temporary	Permanent	Other			
Temporary	Temporary	Permanent			
Temporary	Temporary	Temporary			
Temporary	Temporary	Other			
Temporary	Other	Permanent			
Temporary	Other	Temporary		$\beta_{a_t,a_{t-1},x_t,x_{t-1}}$	
Temporary	Other	Other		$\beta_{a_t,a_{t-1},x_t,x_{t-1}}/\beta_{b_t,b_{t-1},x_t,x_{t-1}}$	
Other	Permanent	Permanent			$\beta_{b_t,b_{t-1},x_t,x_{t-1}}$
Other	Permanent	Temporary			
Other	Permanent	Other			
Other	Temporary	Permanent			
Other	Temporary	Temporary			$\beta_{b_t,b_{t-1},x_t,x_{t-1}}$
Other	Temporary	Other			
Other	Other	Permanent			
Other	Other	Temporary			
Other	Other	Other			

4

HOW LINKAGE ERROR AFFECTS HIDDEN MARKOV MODEL ESTIMATES: A SENSITIVITY ANALYSIS

This chapter was published as: Pankowska, P., Bakker, B. F. M., Oberski, D. L.., & Pavlopoulos, D. (2020). How linkage error affects hidden Markov model estimates: A sensitivity analysis. *Journal of Survey Statistics and Methodology*, 8(3), 483–512. https://doi.org/10.1093/jssam/smz011

Abstract

Hidden Markov models (HMMs) are increasingly used to estimate and correct for classification error in categorical, longitudinal data without the need for a "gold standard," error-free data source. To accomplish this, HMMs require multiple observations over time on a single indicator 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 or incorrectly not linked. In this paper, we investigate the effects of linkage error on HMM estimates of transitions between 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 both false-negative and false-positive linkage error turn out to be problematic primarily if the error is large and highly correlated with the dependent variable. Moreover, under certain conditions, false-positive linkage error (mislinkage) in fact acts as another source of misclassification that the HMM can absorb into its error-rate estimates, leaving the latent transition estimates unbiased. In these cases, measurement error modelling already accounts for linkage error. Our results also indicate where these conditions break down and more complex methods would be needed.

4.1 Introduction

Despite numerous efforts to the contrary, survey and register data almost inevitably contain measurement error (Alwin, 2007; Biemer & Stokes, 2004; Kuha & Skinner, 1997). Such errors severely bias estimates of relationships between variables and, therefore, it is essential to account and correct for them (Carroll et al., 2006; Fuller, 2009; Kuha & Skinner, 1997; Saris & Gallhofer, 2007). For categorical variables, an attractive method of doing so — without requiring "gold standard" (error-free) validation data — is latent class models (LCMs) (Vermunt & Magidson, 2002).

The LCMs use repeated indicators of some categorical phenomenon of interest as input, and output estimates of the classification error rates of these indicators, otherwise known as "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 - which are used as inputs as part of a set of different survey questions or different administrative records – are 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 or administrative record at different time points, a particular variant of an LCM is used: the "hidden" (or "latent") Markov model (HMM) (Alwin, 2007; Alwin et al., 2018). In this paper, we focus on HMMs, which are regularly applied to categorical longitudinal data (Biemer, 2011; Biemer et al., 2017; Edwards et al., 2017).

The great advantage of LCMs is that all indicators are allowed to contain errors and, as such, LCMs can estimate the quality of a survey indicator without requiring perfect comparison data. However, this exciting feature of LCMs does not come cheap: a payment in untestable assumptions is required, in particular the "local independence" assumption, which requires that the errors in the repeated indicators occur independently (see e.g. Oberski et al., 2015).

This local independence assumption is unrealistic, harmful and, when only one indicator is available, also undetectable. It is *unrealistic*, because "common method variance" – that is, variance attributed to the measurement method as opposed to the constructs the measure represents - is typically found in studies able to detect it (Saris & Gallhofer, 2007) and because it is likely that, for instance, any personal "style" in answering a survey question carries over time (Billiet & Davidov, 2008). It is also highly probable that specific errors in registers are repeated for a certain period of time as shown by Pavlopoulos and Vermunt (2015). It is *harmful* because ignoring it leads to bias in the HMM parameter estimates (Georgiadis et al., 2003; Qu & Hagdu, 2012; Torrance-Rynard & Walter, 1997; Vacek, 1985); Appendix 4.A provides an illustration of the severity of the bias using employment mobility data from the Netherlands. 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, in general, possible to detect and model local dependence in LCMs (Hagenaars, 1988;

Oberski, 2016), in HMMs, the parameters that represent local dependence are only generally identifiable if a second indicator of the variable of interest 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.

Therefore, an attractive solution to the problem of local independence is to link different data sources, such as surveys and administrative registers. The attractiveness of this solution lies in the fact that neither of these two data sources is required to be error-free; it is only required that the survey errors are independent of the register errors, which 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 indeed found considerable local dependence (Bassi et al., 2000; Oberski et al., 2017; Pavlopoulos & Vermunt, 2015), confirming both the importance of relaxing this assumption and the attractiveness of data linkage.

Record linkage allows us to tackle the problems of measurement error modelling, but it introduces a new challenge: linkage error. Such errors, which occur when records of different individuals are wrongly linked or when records of the same individuals are wrongly not linked, are known to bias estimates of interest when left unaccounted for (Harron et al., 2017). Several estimators correcting for linkage errors have been suggested (e.g. Chambers, 2009; Goldstein et al., 2012; Lahiri & Larsen, 2005; Liseo & Tancredi, 2011); to illustrate, Di Consiglio and Tuoto (2018) show that these methods are effective in reducing linkage error bias in linear and logistic regression analyses. However, some of these estimators assume knowledge of the posterior probability of correct linkage for all pairs of cases. This knowledge is unavailable to most analysts in practice. The remaining solutions do not assume this knowledge but have only been developed for linear regression models (Chambers, 2009).

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 (transition rate) 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, however, non-negligible biases in the structural part of the model do occur. A novel geometric representation of the latent class estimation problem demonstrates why this is the case.

Section 4.2 first provides some background information on single- and multiple-indicator hidden Markov models and then discusses the topic of linkage error and its effects on HMMs. Section 4.3 presents the data and section 4.4 the methodology; in section 4.5 we discuss the results of our analysis. Section 4.6 provides conclusions.

4.2 Background

4.2.1 Hidden Markov models (HMMs) and measurement error

Hidden Markov models (HMMs) are a group of latent class models that are 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 is assumed to follow a multinomial distribution and 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, where $X = (X_1, \ldots, X_T)$, factorizes into the following product:

$$P(Y|X) = \prod_{t=1}^{T} P(Y_t|X_t) \tag{4.1}$$

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

$$P(X) = P(X_0) \prod_{t=1}^{T} P(X_t|X_{t-1}) \tag{4.2}$$

Finally, the observed data distribution $P(Y)$ is assumed to arise by combining the ICE and Markov assumptions that are mentioned above and then marginalizing over X. This yields the following marginal likelihood:

$$P_{\text{HMM}}(Y) = \sum_{X} P(Y|X)P(X) \tag{4.3}$$

with "structural" parameters $P(X_0)$ and $P(X_t | X_{t-1})$ – which correspond to the initial state and transition probabilities – and "measurement parameters" $P(Y_t | X_t)$ – which are 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 (4.3) over the structural and measurement parameters (Leroux, 1992). In practice, instead of the exponentially complex summation over all possible latent paths X in Equation (4.3), the more computationally efficient "forward-backward" (Baum-Welch) algorithm is used. This amounts to an adapted expectation-maximization (EM) procedure

(McLachlan & Krishnan, 2008, pp. 291-2). In the E-step of this procedure, the posterior probability $P(X|Y)$ is estimated by combining two computational steps: the forward and backward recursions. Specifically, in the forward step, the algorithm calculates the probability of arriving at a specific state at time t given the states that occurred up until that time point; in the backward step, this probability is calculated based on the states occurring at time points following t. Thus, each of the steps considers one time point at a time but in combination with the results of the respective previous computations. In the M-step the model's parameters are computed by summing over the states at each time point. This sum is weighted by the posterior probabilities. 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 (4.3). The E- and M- steps are iterated until convergence is reached.

The single-indicator HMM is attractive for two reasons. First, 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, as already discussed in the introduction, conditional independence may in practice be an unrealistic assumption. 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) suggest linking respondents' survey answers to administrative records. Such linked survey-administrative data then allow for the relaxation of the ICE assumption, replacing Equation (4.1) with

$$P(Y|X) = P(Y_{\text{survey}}|X)P(Y_{\text{admin}}|X) \tag{4.4}$$

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

$$P(Y|X) = \prod_{t=1}^{T} P(Y_{t,\text{survey}}|X_t) \prod_{t=1}^{T} P(Y_{t,\text{admin}}|X_t, X_{t-1}, Y_{t-1,\text{admin}}) \tag{4.5}$$

with $P(Y_{t,\text{admin}}|X_t, X_{t-1}, Y_{t-1,\text{admin}})$ modelled by logistic regression. This model allows for error dependence in the administrative data, while assuming survey and administrative answers to be conditionally independent. The advantages of record 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 linkage error may occur and cause bias in analyses of dependencies, such as linear and logistic regression Chambers and Kim (2015). Therefore, it 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 this specific group of models. This paper does not aim to examine these effects analytically 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 under certain circumstances. In the next section we provide an intuitive explanation of this phenomenon. In doing so, we first provide a formal definition of record linkage and the errors associated with it; we then present a theoretical consideration of how linkage errors (might) affect HMMs.

4.2.2 Linkage, its associated errors, and their effects on HMMs

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 data sources (Armstrong & Mayda, 1993; Winkler, 1999).

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 & Salmond, 2002). As data sources have been increasingly lacking high-quality unique identifiers, 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 & Mayda, 1993; Blakely & Salmond, 2002; Bohensky et al., 2010; Fellegi & Sunter, 1969; Winglee et al., 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: (1) 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 — and (2) when as a result of coding or measurement errors, unrelated records are wrongfully linked — a situation referred to as a *false-positive* linkage error (Bohensky et al., 2010; Winglee et al., 2005).

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 & Mayda, 1993; Sadinle et al., 2011).

The aim of record linkage is to divide set C into two separate sets – one that 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 definite 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 definite non-link, denoted by A_3) (Armstrong & Mayda, 1993; Fellegi & Sunter, 1969; Sadinle et al., 2011).

False-positive and false-negative types of error occur respectively when (1) a record pair that 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 $P(A_1 \mid U)$ and false-negative by $P(A_3 \mid M)$ (Armstrong & 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 assumption that the linkage/mislinkage probabilities of all pairs of records are known to the analyst. This assumption is often unrealistic in practice. Liseo and Tancredi (2011) propose a Bayesian approach to linkage problems, in which the analysis and linkage models are subsumed into a single latent variable model estimated via Markov Chain Monte Carlo (MCMC). A similar approach, implementing Bayesian imputation conditioned on the linkage probabilities, is suggested independently by Goldstein et al. (2012). Other studies that propose Bayesian approaches to correct for linkage error include those by Sadinle (2014, 2017), Steorts (2015) and Steorts et al. (2016). While the Bayesian approach is, in principle, comprehensive, it shares the drawback of the approach of Lahiri and Larsen (2005) that full knowledge of the linkage process is required by the analyst. 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 realistic approximation based on available aggregate linkage rates. As detailed in Chambers and Kim (2015), 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, the available methods to account for linkage error are difficult to implement for HMMs for practical or technical reasons. It is therefore important to investigate the sensitivity of such models to linkage error, which is the focus of this paper. While false-negative linkage error manifests itself as missing data, and a large

literature on the effects of various missingness mechanisms on maximum-likelihood (ML) estimates already exists (see e.g. Little and Rubin, 2019), false-positive linkage errors (mislinkages) have an entirely different, as yet unstudied, effect on HMMs. Therefore, our theoretical considerations elaborate on the effect of mislinkages; the simulation study, however, investigates the effects of both types of linkage errors.

Following Lahiri and Larsen (2005), mislinkage among declared links manifests itself 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 to 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:

$$P_{\text{linked}}(Y) = P(M)P_{\text{HMM}}(Y|\theta) + [1 - P(M)]\prod_{j=1}^{J} P(Y_j) \tag{4.6}$$

where the HMM likelihood has been expressed as $P_{\text{HMM}}(Y|\theta)$ to emphasize its dependence on the model parameters of interest, θ. 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 4.B 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 the 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.

4.3 Data

The data in our analysis are from the Dutch Labour Force Survey (LFS) and the Employment Register (ER), which have been linked using each citizen's unique identification number or the combination of birth date, sex, postal code, and house number as a linkage key. In this paper, we assume that this process does not involve linkage error and simulate the effect of linkage error by artificially introducing false-negative and false-positive linkages into the dataset.

Our sample consists of 15 months of 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, which can take one of the following values: "permanent contract," "temporary contract," or "other." For further details about the dataset, see Appendix 4.C.

4.4 Methodology

4.4.1 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. (2018), our simulations are based on a simplified model that retains the local independence assumption. The following equation estimates the probability of following a certain observed path according to our model:

$$P(C_i = c_{i,}\, E_i = e_i) = \sum_{x_0=1}^{3}\sum_{x_1=1}^{3}\dots\sum_{x_T=1}^{3} P(X_{i0} = x_0)\prod_{t=1}^{T} P\left(X_{it} = x_t \middle| X_{i(t-1)} = x_{t-1}\right) \prod_{t=1}^{T} P(C_{it} = c_t|X_{it} = x_t)\prod_{t=1}^{T} P(E_{it} = e_t|X_{it} = x_t)^{\delta_{it}} \tag{4.7}$$

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.[14] 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. This amounts to assuming an ignorable (MAR) missingness mechanism (Little & Rubin, 2019). 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.

14 In our analysis, we use data from January 2009 until March 2010 which corresponds to 17 months and, therefore, *t* runs from 1 to 17

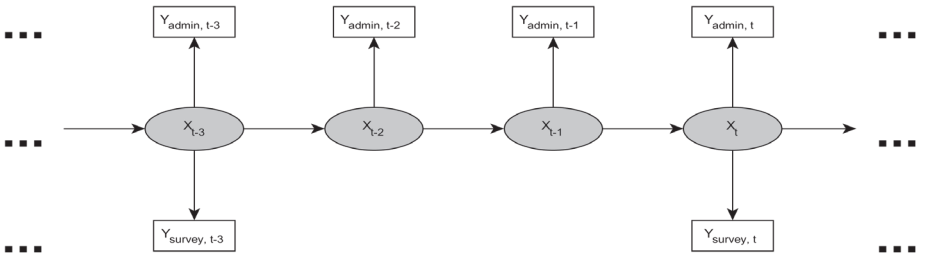


Figure 4.1- Hidden Markov model graph; rectangles are observed variables, while ovals are latent "true" variables. Absence of arrows indicates conditional independence

Figure 4.1 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 of the latent class model with missing data proceeds using maximum likelihood under the ignorability assumption (Little & Rubin, 2019; Vermunt & Magidson, 2013). Standard errors of the parameters can be obtained by inverting the expected or observed information matrix of the observed-data likelihood above.

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 4.2.

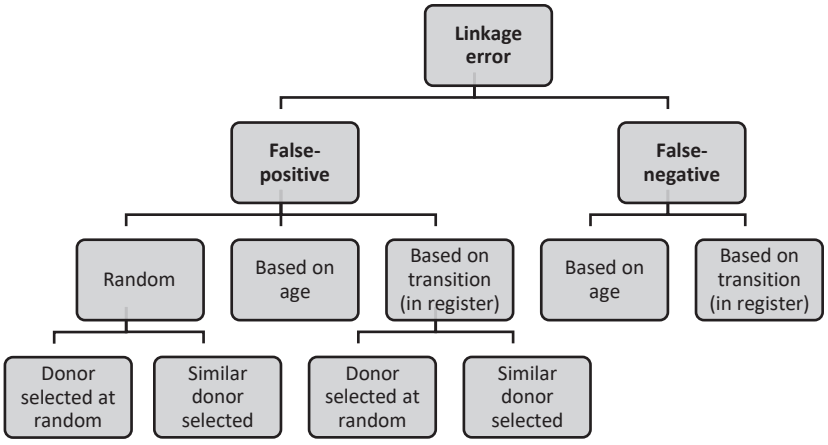


Figure 4.2- Conditions of the simulation study

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. For our code, please see the published paper's supplementary data online.

4.4.2 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 (1) age[15] and (2) the presence of a (three-monthly) transition from temporary to permanent employment in the register data[16]. 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 percent), medium (10 percent), and low (5 percent) overall exclusion error into our data; this error is equal to the proportion of correctly linked individuals in the data that are erroneously excluded.

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 probability of exclusion for those individuals who transitioned according to the register data is higher than that for 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 ensure 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 over- sampling of young individuals or 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 percent), medium (10 percent), and high (20 percent), respectively; the exclusion probability of older individuals (aged 35 to 54) remains at 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.

---

15 Pankowska et al. (2018) 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 the effect of age on the latent transition 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).

16 According to our model, over 99 percent of all contracts observed in ER are correctly classified and, therefore, the transition covariate we have created and the model estimates are highly correlated.

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 equal to that 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 4.D.1 provides pseudocode for generating one condition.

4.4.3 False-positive error simulations

The analysis of the false-positive linkage error, similarly to that for the false- negative, also follows three 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 percent, 10 percent, or 20 percent) 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 one of those individuals in turn, here referred to as individual A, is either (1) randomly matched to another person or (2) 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, wherein 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 is parallel to that of the exclusion error analysis. Our HMM is fitted to each of the simulated datasets, and the outcomes 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 4.D.2.

4.5 Results

4.5.1 The effect of false-negative error

The simulation results obtained for the various false-negative error conditions are shown in Table 4.1; the table provides the mean estimated three-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 4.E.1, which is included in Appendix 4.E, provides an illustration

of the relationship between the type (age or transition dependent), level (5 percent, 10 percent, 20 percent), 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 per- cent 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 percent).

A vastly 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 percent, for an overall linkage error of 5 percent, to 25 percent, when the linkage error amounts to 10 percent, and to as high as 84.3 percent when the error rate equals 20 percent. 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- negligible bias.

Table 4.1- Simulation Results: The Biasing Effects of All False-Negative Linkage Error Conditions (in %)

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	-	-	6.9	-	-
False-negative	Depends on age	5	15	1	6.6	0.3	4.6
		10	30	1	6.7	0.2	3.2
		20	70	1	6.6	0.3	3.8
	Depends on transition	5	15	5	6.2	0.7	10.6
		10	34	9	5.2	1.7	25.0
		20	90	17	1.1	5.8	84.3

Note: 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. 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 that takes the uncertainty of class memberships into account.

Overall, the results obtained suggest that the extended, two-indicator HHM is robust to false-negative linkage error when the exclusion probability depends on age, a covariate that is weakly or moderately correlated with the (structural) model estimates. In these situations, the bias introduced by linkage error is relatively small and thus the HMM estimates can be considered accurate. The model appears sensitive, though, to false-negative linkage error when the individual-level exclusion probabilities depend on whether a transition occurred, a covariate that is highly correlated with the latent variable and consequently the model outcomes. These scenarios lead to a substantial, non-ignorable bias.

Finally, it is worthwhile to note 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 that is moderately correlated with the model estimates; and second, where the probabilities are dependent on a variable exceptionally highly correlated with the model estimates. Our findings confirm this line of thought. More specifically, our results, similar to those reported by studies investigating missingness specifically, show that MNAR leads to substantial bias when the missingness is highly correlated with model estimates (Bakker & Daas, 2012; Galimard et al., 2016; Marshall et al., 2010).

4.5.2 The Effect of false-positive error

The results obtained when simulating various levels and types of false-positive linkage error are presented in Table 4.2 and in Figure 4.E.2, which is included in Appendix 4.E. 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 regard to age, gender, education, and ethnicity. While in the present case mislinking similar donors did not reduce the linkage error bias, this will not necessarily always be the case. If both the mislinkage probability and do- nor matching depend on a variable(s) that is (are) highly correlated with the transition estimates, it is likely that using similar rather than random donors would decrease the bias introduced by linkage error.

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 percent. 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 percent, 20–25 percent, and (well) over 60 percent when the mislinkage rate is low, medium, and high, respectively. Figure A.E.2 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 4.2- Simulation results- the biasing effects of all false-positive linkage error conditions (in %)

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	-	-	6.9	-	-
False-positive; mislinkage with random donor	Random	5	-	-	6.9	0	0.1
		10	-	-	6.9	0	0.3
		20	-	-	6.8	0.1	1.0
	Depends on age	5	15	1	6.9	0	0.3
		10	30	1	6.8	0.1	1.2
		20	70	1	6.7	0.2	2.6
	Depends on transition	5	15	5	6.4	0.5	7.8
		10	34	9	5.5	1.4	20.7
		20	90	17	2.4	4.5	64.6
False-positive; mislinkage with similar donor	Random	5	-	-	6.7	0.2	3.2
		10	-	-	6.7	0.2	3.2
		20	-	-	6.6	0.3	4.9
	Depends on transition	5	15	5	6.1	0.8	11.5
		10	34	9	5.1	1.8	26.6
		20	90	17	1.2	5.7	82.6

Note: 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. 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. 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. 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 that takes the uncertainty of class memberships into account.

Figure 4.3 demonstrates how mislinkage affects the measurement part of the model; that is, 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.[17] 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 correct for in the transition rate estimates.

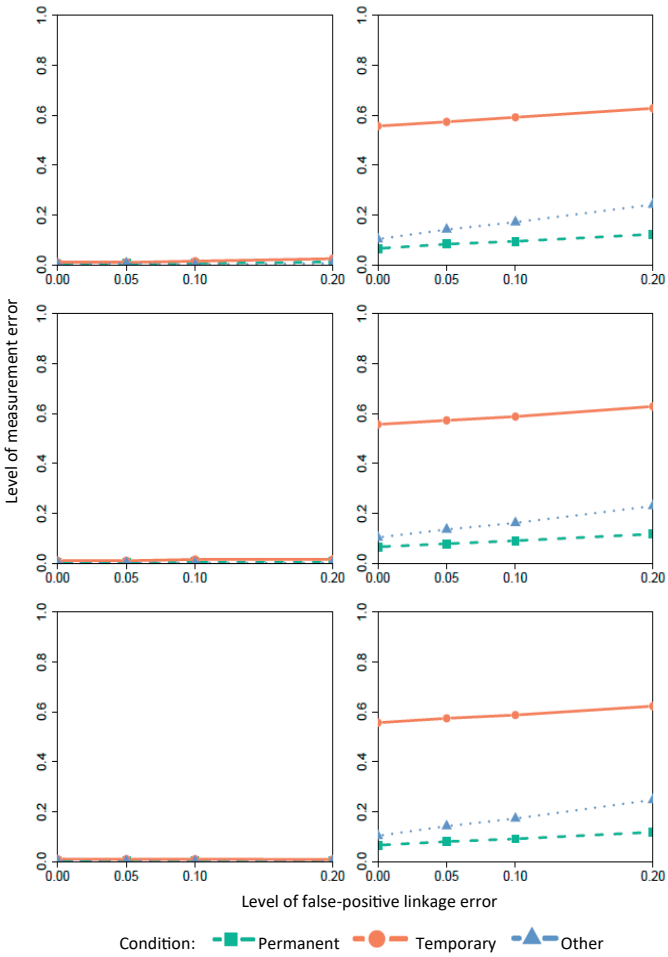


Figure 4.3- Level of measurement error by type and level of mislinkage

17 This pattern is not observed in the ER data, as the simplified HMM we use does not account for auto-correlation 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.

4.6 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 the study of transitions and change over time, which is often a quantity 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.

In this paper, we investigate the sensitivity of HMM estimates to linkage error. A geometric argument demonstrated that independent (false-positive) link- age 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 latent 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 HMMs for the purpose of estimating their structural parameters, such as transition rates. When individuals are randomly mislinked or not 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 significant bias only when the individual probability of being erroneously excluded or mislinked depended on the transition rate itself. The bias was particularly high for high rates of linkage error and when the aforementioned dependency was very strong; in the other instances investigated, the sensitivity of estimates of structural parameters to mislinkage appears relatively low.

Our results show that false-positive linkage error can often be absorbed by the model. In other words, mislinkage can manifest 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 using HMMs to correct for measurement error in structural parameter estimates, since, in particular cases, they allow for the use of linked data with relatively low sensitivity to linkage error. This is especially appealing, as the methods available to correct for link- age 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, and not the structural, parameters are of primary interest (e.g. Biemer, 2011), our results suggest that 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 link- age error, perhaps by extending the estimating equations approach discussed in (Chambers & Kim, 2015).

Furthermore, while our manuscript provided novel results on the effect of linkage error on point estimates, the effect on the variance of these estimates remains unknown. For false-negative linkage errors (i.e. missed links), the standard theory of missing data applies, and the observed information will equal the information without these errors minus the information that would have been obtained in the missed links (Little & Rubin, 2019). The effect of false-positive links (i.e. incorrectly linked records) on the variance, however, remains an open question for future work.

Appendix 4.A The Effect of local independence assumption violations on HMM estimates — an illustration using real data

As mentioned in the introduction, the local independence assumption, which is necessary for model identification for the standard, one-indicator HMM, is in many cases unrealistic for both survey and register data. If this assumption is violated, HMM estimates are likely to suffer from (considerable) bias and, as such, it is necessary to relax it, which is possible when using multiple indicators per time point.

We provide here an illustration of the biasing effects of local independence assumption violations using data on labor mobility in the Netherlands from the Employment Register (ER) and the Labour Force Survey (LFS) for the years 2009 and 2010. In doing so, we compare the temporary to permanent employment transition estimates obtained using a one-indicator HMM that only uses register data to those obtained using a two-indicator HMM applied to linked ER and LFS data. In the latter model specification, we relax the local independence assumption for the register data, as it is known from previous research (Pankowska et al., 2018; Pavlopoulos & Vermunt, 2015) that the measurement error in ER is autocorrelated and, thus, that the local independence assumption is violated. Table 4.A.1 compares the transition estimates obtained from both models; as can be seen, the one-indicator HMM, which erroneously retains the local independence assumption for the register data, significantly overestimates the transition rate from temporary to permanent employment. The relative bias resulting from ignoring the violation of the local independence assumption amounts to 310 percent.

Table 4.A.1- Transition estimates and bias for one- and two- indicator HMMs

Model specification	Transition estimate (temp → perm)	Absolute bias	Relative bias
One-indicator HMM • Only using ER • Retaining ICE	0.0689	0.0521	310%
Two- indicator HMM • Using ER and LFS • Relaxing ICE for ER	0.0168	-	-

Note: For computational reasons the simulation study uses a two-indicator HMM which does not relax the local independence assumption for the register data and does not model autocorrelated measurement error in the ER; therefore, the transition rate in the absence of linkage error resembles the one obtained from a single indicator HMM in Appendix 4.A.

Appendix 4.B 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 2X2 cross-tables are placed in a tetrahedron representing the simplex $\{x \in R^4 : \sum x_i = 1\}$ (Figure 4.B.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 2X2 cross-tables can be represented as a single point within the tetrahedron. An important subset of tables is the "independence surface" formed by all 2X2 independence tables, which is shown in Figure 4.B.1 as the shaded surface. Points along a line on this surface correspond to all independence tables with constant row or column margins.

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) and $Y_3 = 1$ (point P_1). The maximum-likelihood estimates of the conditional distributions given the latent class $P(Y_1,Y_2 \mid X = 0) = \eta_0$ and $P(Y_1,Y_2 \mid X = 1) = \eta_1$ are then found at the two intersections of the "solution line" $P_1 - P_0$ with the independence surface. This follows from the latent class model's assumption 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, X, must lie on the independence surface. The MLE of $P(X \mid Y_3 = 0)$ is then found as 1-length$(p_0 - \eta_0)$/length$(\eta_1 - \eta_0)$ and similarly, $\hat{P}(X \mid Y_3 = 1) = 1 - length((p_1 - \eta_1)/length((\eta_1 - \eta_0)$, 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 lie at greater distance from each other, accuracy estimates under the LCM increase, with the maximum attained at the corners of the tetrahedron (estimated sensitivity and specificity equal to one).

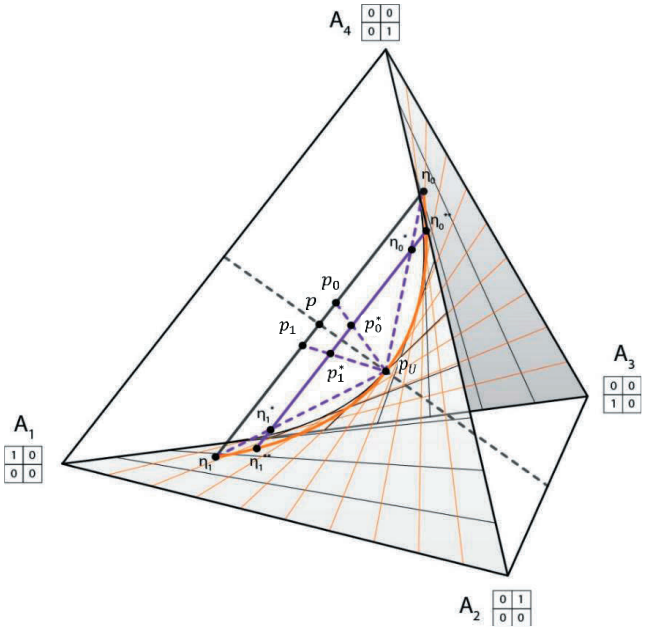


Figure 4.B.1- Geometrical view of fitting of a latent class model to data with independent linkage error

We now consider how the MLEs are affected by independent linkage error. Following Equation (4.6), we consider the distribution of linked records as a mixture over true matches and true non-matches, indicated by a random variable U. When false-positive linkage error is independent, $P(Y \mid U) = P(Y)$, the $P(Y)$ in the equation above reduce to the marginals under the model, $P(Y_j) = \sum_{Y_{k \neq j}} P_{HMM}(Y)$. This point, p_u in Figure 4.B.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 & Gilbert 1970, p. 699). The linkage error model in Equation (4.6) 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_1^* and p_0^* lie on a solution line parallel to the original solution line, with length $(p^* - p)$/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 4.B.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_1^{**} and , 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 4.B.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, $P(Y \mid U) \neq P(Y)$. 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, $p_{u,dep}$, 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 occurring. 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 4.C The combined LFS and ER dataset

4.C.1 Background information on the LFS and ER

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

The Employment Register (ER) is an administrative dataset managed by the Dutch Employee Insurance Agency (UWV). It contains monthly information on wages, benefits, and labor 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.

4.C.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 dropout, as all employers are obliged by law to submit their reports, 2,619 observations are missing, which amounts to just under 2 percent of the sample. Those observations are also assumed to be MCAR.

4.C.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 differs. 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, that is, the percentage of linked records, of 98.3 percent for those who had a first interview in 2009.

The ER is linked to the PR in three 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, ignoring any differences in the other variables. This procedure is repeated monthly. The overall linkage effectiveness is approximately 96–97 percent, depending on the chosen month; 99.8 percent 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 link- age effectiveness of the combined sources is approximately 97 percent. The unlinked records refer to cross-border workers from Belgium or 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 4.D Simulation design

The simulations are designed in the following way. First, we identify young individuals or individuals who had at least one three-monthly transition from temporary to permanent employment recorded in the register data (this step is skipped for random mislinkage conditions). Second, we assign one of two exclusion/mislinkage probabilities to each individual: a higher one for individuals identified in the first step (i.e. younger or who have "transitioned") and a lower one for all remaining ones (we assign the same probability to everyone in the random mislinkage conditions). Third, given the assigned probabilities, we select individuals for exclusion/mislinkage at random. Fourth, in the case of false-negative linkage conditions, we exclude the chosen individuals; in false-positive linkage conditions, we assign the selected individuals to a donor and replace their ER contract type with that of the donor. The assignment to the donor can be either completely random or based on similarity given the age, gender, nationality, and education of individuals. Finally, we run our HMM on the simulated datasets and compare the estimated transitions rates to those obtained when no linkage error is introduced into the dataset.

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

4.D.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 - $U_i \sim U(0,1)$
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

4.D.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 - $U_i \sim U(0,1)$
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

Appendix 4.E Illustration of simulation results

Figures 4.E.1 and 4.E.2 provide an illustration of the relationship between the type (random, age, or transition-dependent), level (5 percent, 10 percent, 20 percent), and bias introduced by false-negative and false-positive linkage error, respectively.

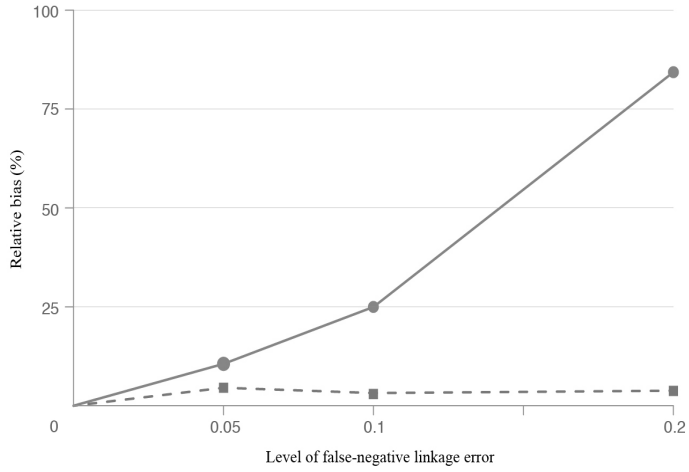
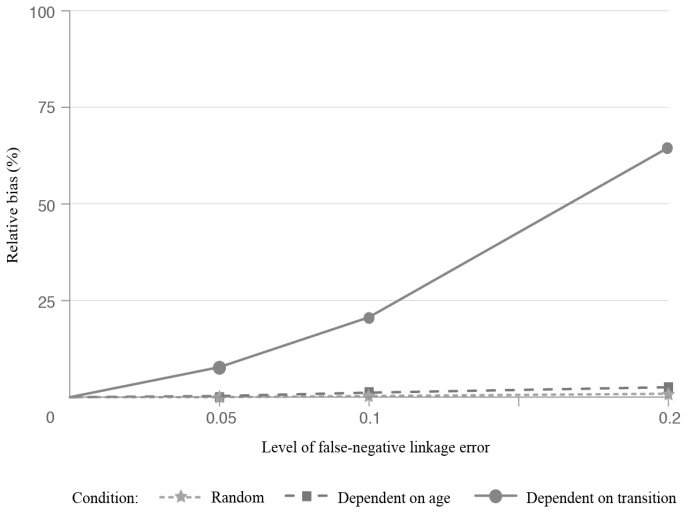


Figure 4.E.1- Relative bias by overall level of false-negative linkage error



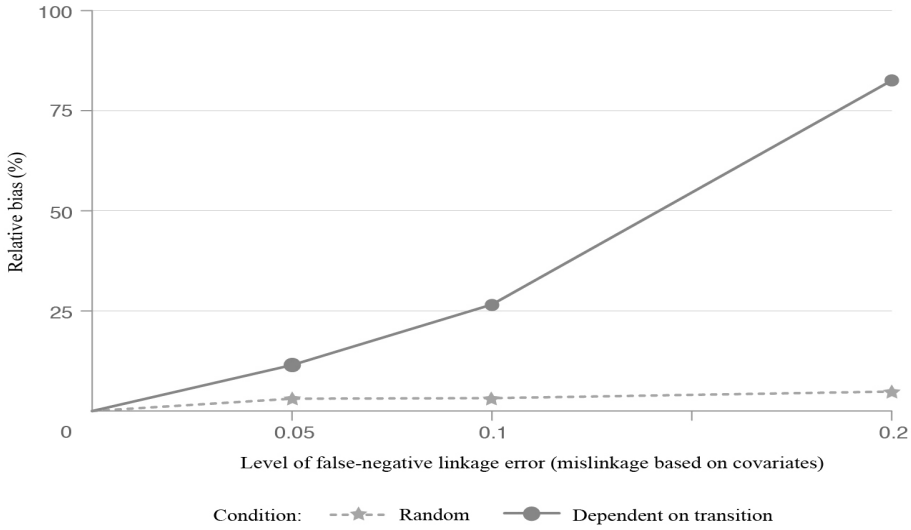


Figure 4.E.2- Relative bias by overall level of false-positive linkage errorReconciliation of inconsistent data sources by correction for measurement error: The feasibility of parameter re-use

5

RECONCILIATION OF INCONSISTENT DATA SOURCES BY CORRECTION FOR MEASUREMENT ERROR: THE FEASIBILITY OF PARAMETER RE-USE

This chapter was published as: Pankowska, P., Bakker, B., Oberski, D. L., & 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(3), 317–329. https://doi.org/10.3233/SJI-170368

Abstract

National Statistical Institutes (NSIs) often obtain information about a single variable from separate data sources. Administrative registers and surveys, in particular, often provide overlapping information on a range of phenomena of interest to official statistics. However, even though the two sources overlap, they both contain measurement error that prevents identical units from yielding identical values. Reconciling such separate data sources and providing accurate statistics, which is an important challenge for NSIs, is typically achieved through macro-integration. In this study we investigate the feasibility of an alternative method based on the application of previously obtained results from a recently introduced extension of the hidden Markov model (HMM) to newer data. The method allows a reconciliation of separate error-prone data sources without having to repeat the full HMM analysis, provided the estimated measurement error processes are stable over time. As we find that these processes are indeed stable over time, the proposed method can be used effectively for macro-integration, to reconciliate both first-order statistics — e.g. the size of temporary employment in the Netherlands — and second-order statistics — e.g. the amount of mobility from temporary to permanent employment.

5.1 Introduction

National Statistical Institutes (NSIs) often obtain information about the same phenomenon from different data sources (Van Delden et al., 2016). For example, the Dutch Labour Force Survey administered by Statistics Netherlands includes data that overlap to some extent with register data from the Dutch social security administration. The overlapping component of these datasets can be linked at the individual level. Such linked survey-register data often concern longitudinal measures of categorical variables such as employment, housing, and education, and are subject to editing procedures to improve data quality (De Waal, 2016; De Waal et al., 2011). However, even then, identical units do not always yield identical values (Guarnera & Varriale, 2016).

Two types of error may account for these discrepancies: measurement error and linkage error. However, linkage error, while an important error source in official statistics generally, is less of a concern for Statistics Netherlands due to the use of unique resident identifier numbers. We will therefore focus on the problem of measurement error.

Measurement error in surveys is a well-known and extensively studied phenomenon (Alwin, 2007; Saris & Gallhofer, 2014). Measurement error in administrative registers, by contrast, has only recently attracted attention (Bakker, 2012; Oberski et al., 2017; Oberski, 2015; Scholtus et al., 2015; De Waal et al., 2011). Such errors occur because registers result from data collection of public administration and are not originally intended for social-scientific research. When it occurs during data entry, measurement errors in administrative registers mirror familiar survey response errors; however, errors unique to registers also occur, including administrative delay, definition error, and errors caused by administrative incentives (Bakker & Daas, 2012; Huynh et al., 2002; Zhang, 2012).

Where measurement error is random or “classical”, the resulting data will not tend to bias “first-order” population estimates such as means, proportions, and totals (Bound et al., 2001). However, “second-order” estimates, such as domain mean differences, hazard ratios, and transition rates over time, are well-known to be severely biased by random measurement error (Bolck et al., 2004; Carroll et al., 2006; Fuller, 2009; Pavlopoulos et al., 2012). This bias may refer to an overestimation or an underestimation of these statistics.

For example, an important issue in labor market policy is the proportion of workers who change from employment with a flexible contract to employment with a permanent contract. If there is random measurement error in the type of employment contract, these transition rates are artificially (and severely) inflated as every misclassification in the contract type may lead to two errors in the measurement of transitions (Hagenaars, 1994).

On the other hand, if errors are carried over between time points, the observed transitions rates are artificially dampened, as some real changes are not observed. Considering that the source and the type of the measurement error differs between data sources, the problem faced by NSIs is not only that different data sources yield

different statistics, but also that measurement error may bias statistics in a different way in each of these sources.

There are several methods dealing with these differences in NSIs. Most commonly, the differences are ignored and only estimates from the source assumed to have the best quality are published. Another way is to assume that the quality of both sources is similar and take the mean of the estimates. However, a more advanced way of dealing with these differences is to apply macro-integration techniques. One of the usual integration strategies is that in the first step the stock data of two reference dates are integrated. In this step, the concepts, classifications and reference dates are harmonized, the data are completed by weighting or imputation if the data do not cover the entire target population, and, in order to minimize measurement error, the data are forced to meet identity relations defined beforehand. In the second step, data on the events between the two reference dates are made consistent by making use of the identity relations that the stock at reference date t plus all the changes add up to the stock at reference date $t + 1$. However, for the second step, only the source that is assumed to be of superior quality is used. In this second macro-integration step, one can try to preserve the original transitions in the (sub)populations as much as possible.

An alternative strategy to deal with this problem of inconsistency was recently introduced by Bakker (2012) for continuous cross-sectional data, by Oberski et al. (2017) for mixed type cross-sectional data, and by Pavlopoulos and Vermunt (2015) for categorical longitudinal data. In this *latent variable modelling* approach, the reconciliation and measurement error problems are solved simultaneously by modelling the two sources as conditionally independent measures of an underlying true value. In the cross-sectional models, this true value is related to other, similar, true values. Since repeated observations of a single linked survey-register variable may be more common in practice, we focus on the case of longitudinal data. In these models, the true value is related to itself over time in an autoregressive process that yields an extended — multiple data source — version of the hidden Markov model popularized by Biemer (2004, 2011), which in turn is a special case of the latent class model (Van de Pol & Langeheine, 1990; Vermunt, 2002). Previous work done at Statistics Netherlands used such models to integrate data from Labour Force Survey and social security administration (Pavlopoulos & Vermunt, 2015).

A problem with this procedure is that it is very time consuming and therefore expensive, since it requires the NSI to perform linkage between register and survey followed by re-estimation of the model for each new time period. This paper therefore considers the option of re-using existing parameter estimates from the above study in order to integrate data sources and correct statistics for measurement error. Re-use is potentially attractive because (1) it does not require re-estimation of the model, and (2) it can be applied not only to linked survey-register data, but also to each data source separately, forgoing the need for a time intensive linkage exercise.

However, parameter re-use can only be applied to regular production at NSIs if the parameters of the model remain the same over time. If the parameter estimates do

not exhibit stability over time, the corrections themselves will be biased. Therefore, an important question for the practical application of latent variable modelling at NSIs is whether there is indeed stability in the estimates when applying this procedure to real data. In this paper, we demonstrate how this question can be investigated using newly collected data on a topic studied previously and for earlier years by Pavlopoulos and Vermunt (2015). In other words, our analysis allows us to determine whether the aforementioned time- and cost-efficient methodology (based on using previously obtained parameters) can actually work in practice.

The next section describes the data used in the analysis; this is followed by a discussion of the empirical methodology, the results, and finally a brief conclusion.

5.2 Data

The dataset used for the analysis contains information from the Netherlands' Labour Force Survey that is conducted by Statistics Netherlands and the *"Polisadministratie"* (administrative data collected by the Employee Insurance Agency).

The Dutch Labour Force Survey (LFS) is a sample survey aimed at providing information about the relationship between individuals and the labor market. The target population consists of individuals aged 15 and older who reside in the Netherlands (excluding those in homes and institutions) and the information is collected at both the individual and household level.[18] Since the last quarter of 1999 the survey has been a rotating panel survey, consisting of five waves.

The Employment Register data (i.e. the *"Polisadministratie"* or ER) is an administrative dataset administered by the Dutch Employee Insurance Agency (EIA, or *UWV* in Dutch). The dataset contains monthly information on wages, benefits, and labor relations for all insured employees in the Netherlands. EIA uses the information collected to determine the level of benefits. The dataset combines information from various sources; the core of the information is delivered by the employers on their employees each month for tax purposes to the Dutch Tax Authorities, information from temporary work agencies and the Population Register (PR, in Dutch: *Basis Registratie Personen*-BRP)[19] is also used.

The data from both sources are linked at the individual level to the population register (PR) of the Netherlands. Therefore, the target population is restricted to the registered population in the Netherlands. For the linkage of the LFS with the PR, the linkage key is the combination of birth date, gender, postal code and house number. The ER is linked to the PR based on the social security number (BSN),[20] birth date, gender,

18 http://www.cbs.nl/en-GB/menu/methoden/dataverzameling/dutch-labour-force-survey-characteristics.htm

19 http://www.uwv.nl/overuwv/english/about-us-executive-boardorganization/detail/organization/data-services

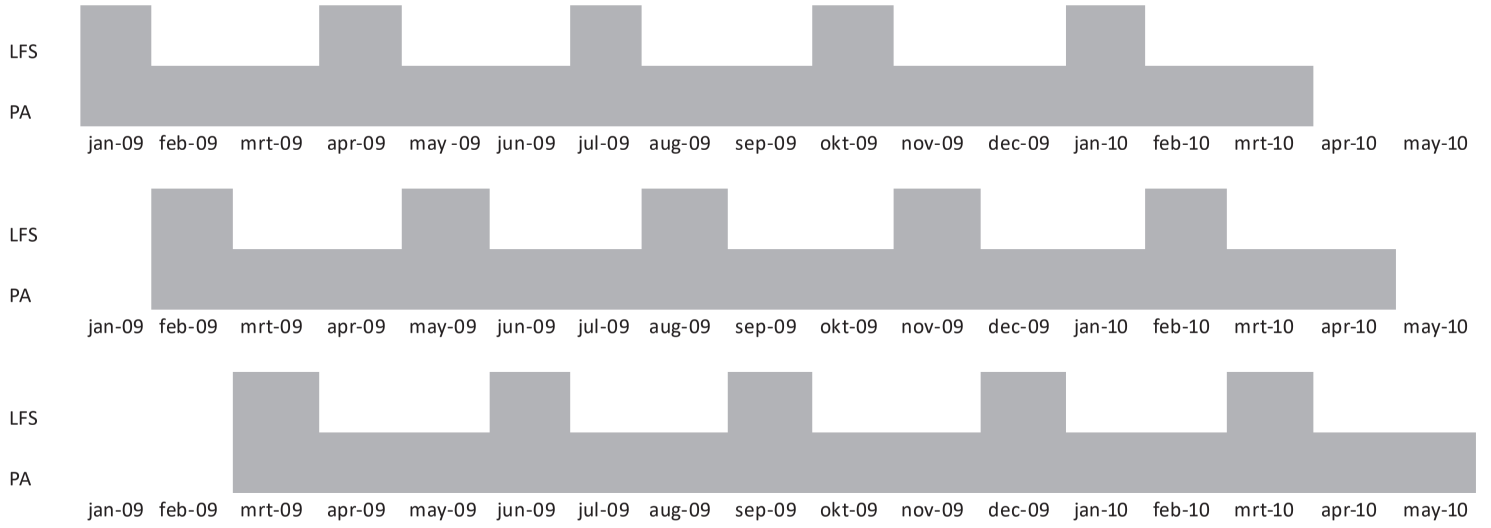
20 A unique personal number allocated to everyone registered in the Netherland; https://www.government.nl/topics/identificationdocuments/contents/the-citizen-service-number

postal code and house number. After selection of the individuals aged 25–55, the linkage effectiveness of the combined sources is approximately 97 percent.

The sample used for the analysis consists of 8,886 LFS respondents aged between 25 and 55 who participated in the LFS for the first time in the first trimester of 2009. For each individual included in the sample, the dataset contains information for a period of 15 months with the variables coming from the ER data observed on a monthly basis (i.e. 15 observations) and those from the LFS observed every 3 months (i.e. 5 observations). The time period the data correspond to, January 2009 to May 2010, is illustrated in Figure 5.1 (with the time period from January 2009 to March 2010 corresponding to those individuals first interviewed in January 2009; those from February 2009 to April 2010 to those firstly interviewed in February 2009; and those from March 2009 to May 2010 to those firstly interviewed in March 2009).

The panel dataset is unbalanced for the LFS as it suffers from attrition. More specifically, 8,708 individuals participated in the first round of the survey, 7,458 in the second, 6,856 in the third, 6,739 in the fourth and 6,560 in the fifth. For the non-survey months observations are assumed to be missing at random. While the ER officially cannot be subject to drop-out as submission of reports is obligatory for all employers, 2,619 observations (out of a total of 133,290) are missing. We assume that these missing observations are also missing at random.[21]

21 Those are primarily observations of workers who have passed away or emigrated from the Netherlands and, thus, there is no reason to believe their missingness is related to the variable of interest.



Note: 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 5.1- An illustration of the sample

The key variable of interest in the analysis is the contract type held by the individual for his or her main job (any secondary jobs are ignored in the analysis). The contract type can take on three distinctive and mutually exclusive values: "permanent contract" (i.e. a contract for an unlimited duration of time), "temporary contract" (i.e. a fixed contract for a limited duration of time) and "other" (which includes all other alternatives, i.e. self-employment, unemployment, unpaid employment and full-time education). While the last category is not of crucial importance for answering the research questions themselves (as the focus is on transition rates from temporary to permanent employment) it has to be included in the analysis to assure that the Markov assumption of mutual exclusivity and exhaustiveness of the latent classes is fulfilled.

The distributions of the contract types according to both data sources are displayed in Table 5.1. As can be observed, the aforementioned distributions as indicated by the survey and register data accordingly differ to a larger extent for permanent and temporary contracts than other types of contracts.

Table 5.1- Distribution of contract types according to the survey and the register, January, February and March 2009

	Survey	**Register**
January '09		
Permanent	0.637	0.586
Temporary	0.113	0.154
Other	0.250	0.261
Total	1.000	1.000
Cases	3,173	3,175
February '09		
Permanent	0.627	0.576
Temporary	0.120	0.163
Other	0.254	0.262
Total	1.000	1.000
Cases	2,857	2,849
March '09		
Permanent	0.642	0.596
Temporary	0.112	0.155
Other	0.247	0.249
Total	1.000	1.000
Cases	2,678	2,692

In order to gain more insight into the extent of the aforementioned inconsistencies, the contract type according to both datasets has been cross- tabulated for the entire sample. The results, presented in Table 5.2, show that while the discrepancies between the survey and register data concerning individuals who hold a permanent contract or occupy the state "other" are relatively small, those regarding individuals employed on a temporary contract are highly substantial.

Table 5.2- Cross- tabulation of contract type according to survey and the register

	Survey Data				
Register Data	Permanent	Temporary	Other	Total	Cases
Permanent	0.934	0.052	0.015	1.000	21,840
Temporary	0.517	0.441	0.043	1.000	5,347
Other	0.060	0.059	0.881	1.000	8,411
Total	0.665	0.112	0.224	1.000	35,598
Cases	23,654	3,983	7,961	35,598	-

Note: The frequency distributions are calculated for all observations in the sample which are non-missing for both the LFS and ER.

The disparities between the two datasets with regards to the contract type (and in particular to temporary contracts) outlined above have implications for the estimation of the transition rates between the different contract types. Namely, as depicted in Table 5.3, the transition rate from temporary employment in month t - 3 to permanent employment in month t equals 5.8 percent according to the survey data while it amounts to 7.3 percent according to the register data.

Table 5.3- Observed 3-month transitions in LFS and ER

Observed transitions from the survey data (LFS)			
	Contract in t		
Contract in t-3	Permanent	Temporary	Other
Permanent	0.983	0.006	0.011
Temporary	0.058	0.879	0.063
Other	0.016	0.037	0.947
Total	0.672	0.110	0.218
Observed transitions from the register data (ER)			
	Contract in t		
Contract in t-3	Permanent	Temporary	Other
Permanent	0.976	0.012	0.012
Temporary	0.073	0.869	0.058
Other	0.019	0.043	0.938
Total	0.623	0.148	0.229

Note: For both tables, these are the transition rates over a 3-month period and for 34,387 cases of the pooled sample. These cases come from the LFS- respondents that appear at least twice in the sample and have an observation for both LFS and ER.

5.3 Methods

5.3.1 Classification error model for survey and register

The methodology applied in this paper is based on the extended Hidden Markov Model used by Pavlopoulos and Vermunt (2015). The standard Hidden Markov model discussed by Biemer (2011) assumes that an observed categorical variable Y_t is generated in the following way:

- At $t = 0$
  - Sample a "true value" x_o from the unknown distribution $p(X_0)$
  - Sample the observed value y_0 from the unknown conditional distribution $p(Y_t | X_t)$. The off-diagonal entries in this unobserved cross-table are the misclassification rates and the diagonal entries the probability of a correct classification.
- At $t > 0$,
  - Sample a "true value" x_t from the unknown distribution $p(X_t | X_{t-1})$. The unobserved cross-table between X_t and X_{t-1} contains the unobserved transition rates of substantive interest. In our example, the parameter $p(X_t = permanent | X_{t-1} = temporary)$ is specifically of interest,
  - As before, sample the observed value y_t from the unknown conditional distribution $p(Y_t | X_t)$
- Advance one step in discrete time by setting $t \leftarrow t + 1$ until the maximum number of observed time points $t = T$ is reached.

The unknown parameters of this model are those describing the initial state distribution $p(X_0)$, the misclassification rates $p(Y_t | X_t)$, and the *AR(1)* autoregressive transition rates $p(X_t | X_{t-1})$. These parameters are identifiable by assuming equal misclassification and transition rates over time, i.e. $p(Y_t | X_t) = p(Y_{t'} | X_{t'})$ and $p(X_t | X_{t-1}) = p(X_{t'} | X_{t'-1})$ for all $t \neq t'$. Since only the joint distribution of the observed variables $p(Y_{t0}, Y_{t1}, \ldots, Y_t, Y_T)$ is observed and X_t is entirely missing, estimation of the unknown parameters often proceeds by marginal maximum likelihood, expectation-maximization, or Markov Chain Monte Carlo methods. In what follows we employ the Latent GOLD software, which uses a combination of expectation-maximization and marginal maximum likelihood estimation. It is also straightforward to implement covariates affecting the distribution of X; for the sake of clarity we have omitted these in the description but do include them in our extended model.

The standard hidden Markov model has the substantial disadvantage that it makes the assumption of conditional independence of errors, sometimes also referred to as the "independent classification errors" or ICE assumption. In other words, it assumes that when y_t was generated, its probability of occurring only depended on x_t and nothing else. This precludes, for example, the possibility that any errors that occurred at the previous time point were copied over to the current time point, since that would make the observed value dependent on both the true value and the observed value

at the previous time point, i.e. $p(Y_t \mid X_t) \neq p(Y_t \mid X_t, X_{t-1}, Y_{t-1})$. Since there are considerable indications that register errors are copied over time, the standard Hidden Markov model is inappropriate.

As mentioned before, in this paper, we follow Pavlopoulos and Vermunt (2015) in employing an extension to the standard HMM that allows for error-copying over time in the register. The parameters of this model are identified by linking the register to a survey measuring the same true value over time, in addition to assuming parameters are equal over time. A graphical illustration of the model for the first 4 months is given in Figure 5.2

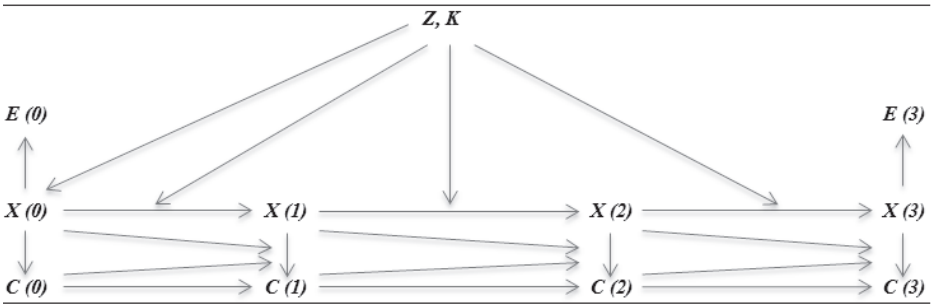


Figure 5.2- Path diagram for the hidden Markov model with two indicators, serially correlated and covariate dependent register errors and predictors for latent transitions and latent state probabilities

The extended HMM assumes that the observed register values $Y_t^{(r)}$ and survey values $Y_t^{(S)}$ were generated as follows:

- At $t = 0$,
  - Sample a "true value" x_0 from the unknown distribution $p(X_0)$,
  - Sample the observed register value $y_0^{(r)}$ from the unknown conditional distribution $p(Y^{(r)} \mid X)$,
  - Sample the observed survey value $y_0^{(s)}$ from the unknown conditional distribution $p(Y^{(s)} \mid X)$.

- At $t > 0$,
  - Sample a "true value" x_t from the unknown distribution $p(X_t \mid X_{t-1})$
  - Sample the observed survey value $y_t^{(s)}$ from the unknown conditional distribution $p(y^{(s)} \mid X)$
  - If the register at the previous time point had an error and no change in true value occurred, i.e. if $x_{t-1} \neq y_{t-1}^{(r)}$, and $x_{t-1} = x_t$ ("*previous error and no change*"),
    - Sample the observed register value $y_t^{(r)}$ from the unknown distribution $p(Y_t^{(r)} \mid$ previous error and no change). This distribution contains the probability of copying an error when no change occurred in the true value, $p(Y_t^{(r)} \mid$ previous error and no change),

- Else if $x_{t-1} = y_{t-1}$ or $x_{t-1} \neq x_t$ (there was no error, or true change occurred),
  - Sample the observed register value $y_t^{(r)}$ from the unknown conditional distribution $p(y^{(r)}|X)$
- Advance one step in discrete time by setting $t \leftarrow t + 1$ until the maximum number of observed time points $t = T$ is reached.

Again, covariates Z are easily included by extending $p(X|\ .\ )$ to $p(X|.,Z)$, where "." may indicate a set of random variables. In our model, this set of covariates always includes the timepoint to allow for variation over time in the transition probabilities. To control for unobserved heterogeneity in the transition probabilities, we further extend $p(X|.,Z)$ to $p(X|.,Z,k)$, where k denotes the latent class that the individual belongs to.

In addition to the output from the HMM, which also provides estimates of the transition rates and misclassification rates, the extended HMM also provides estimates of the error-copying rates. Moreover, the misclassification rates estimated for the register are conditional on no error having occurred previously. Since this cannot be known in practice, we will report both these estimates, and the overall error rates that average over previously occurring errors and correct reports.

The extended model allows for error-copying over time and therefore relaxes the ICE assumption. However, it does this by introducing the assumption that the survey and register values are conditionally independent, given the true value. In what follows we will evaluate the fit of these models before turning to interpretation.

5.4 Results

We first apply the extended HMM described above to the data from 2009; then, we repeat the analysis for the same cohort while fixing the measurement error specific parameters to those obtained by Pavlopoulos and Vermunt (2015) when analyzing data from 2007. The results of the two analyses are then compared to verify whether it is possible to correct for measurement error in data sources over the course of several years while only applying the full extended HMM analysis once at the initial stage.

5.4.2 Model fit

To assure that the model specification used by Pavlopoulos and Vermunt (2015) fits more recent data equally well, we estimated a total of nine different specifications of the hidden Markov model. Those specifications were also estimated by Pavlopoulos and Vermunt (2015) to reach the final version of the model. The goodness-of-fit measures of those models are summarized in Table 5.4. In more detail, the table includes the following information: the log-likelihood, the Bayesian Information Criterion (BIC), the Akaike Information Criterion (AIC) values as well as the number of model parameters.

The first three models used *(A', A" and A)* assumed, respectively, that only the survey data, only the register data and both datasets are subject to independent classification errors (ICE). The fact that the last of the three models fits the data best

provides support for the hypothesis that both data sources contain measurement error. As such, the subsequent six models are extensions of the model assuming the presence of classification errors in both the survey and register data.

The next three models estimated *(B', B" and B)* relax the ICE independence assumption of the measurement error for the survey, the register and both data sources, respectively. In the survey data this is related to the fact that the likelihood of making an error often varies according to age and proxy interview (Bergin, 2013; Bingley & Martinello, 2014; King et al., 2012).

The serial correlation of the measurement error in the register data is likely to result from the fact that companies submit information — including the contract type — to the Employment Office once or twice a year. This is likely to result in errors being carried over until an actual change in the contract type occurs or until some form of data quality control takes place (Bakker & Daas, 2012; Conrad et al., 2009; Groen, 2012; Zhang, 2012). Therefore, the probabilities of having an error in the register data are modelled in such a way that they depend on the lagged observed and lagged true contract.

As can be seen from Table 5.4, models B" and B, which relax the ICE assumption only for the register data and for both datasets respectively, perform somewhat better than model B', which assumes that only the survey errors do not satisfy the local independence assumption. This means that it is realistic to conclude that the error is indeed serially correlated in the register data but not in the survey data. Therefore, the final set of models (C', C" and C) extends those two models by including covariates for the latent transition and for the latent initial state probabilities and thus assumes that those transitions and probabilities are heterogeneous.

In more detail, model C' can be seen as a restricted extension of model B" as it assumes that the measurement errors are not locally independent for the register data and that the latent transitions depend on gender, age, education and country of origin. Model C" can be seen as a full extension of B" as it also assumes that ICE does not hold for the register data but, in addition to the latent transitions, it also assumes that the aforementioned covariates influence the initial state probabilities. Finally, model C can be seen as a full extension of model B as it assumes that ICE should be relaxed for both data sources and that the covariates influence both the latent transitions and initial state probabilities. The covariates are allowed to be time heterogeneous.

Table 5.4- Fit measures for nine models estimated with the linked LFS and ER data

	LL	BIC (LL)	AIC (LL)	Parameters	L^2	df	p-value
A': ICE Error LFS	-35,983	72,365	72,053	44	32,458460	8,842	8.9e-2635
A'': ICE Error ER	-58,742744	11,78857887	11,7574	44	77,979	8,842	4.6e-1084
A: ICE Error LFS and ER	-35,852	72,159	71,805	50	32,198	8,836	2.2e-2595
B': non-ICE Error LFS	-35,717719	71,926928	71,544	54	55,691	8,832	2.1e-6647
B'': non-ICE Error ER	-30,875	62,313	61,873	62	54,435	8,824	1.3e-6421
B: non-ICE Error LFS and ER	-31,048050	62,697699	62,230	66	60,361363	8,820	3.3e-7512
C': non-ICE Error ER with covariates	-31,025027	62,942944	62,248	98	61,588590	8,788	1.3e-7753
C'': non-ICE Error ER with covariates also initial state	-30,647	62,240	61,503	104	60,831	8,782	2.6 e-7615
C: non-ICE Error LFS and ER with covariates also initial state	-30,634636	62,287289	61,493	112	60,806	8,774	5.6e-7614

Note: A'- ICE for the survey; A"- ICE for the register, A- ICE for both datasets, respectively.
B' - survey error depends on age and proxy interview; B"- register errors serially correlated; B- combines B' and B".
C' extend B" by introducing gender, age, education and country of origin as predictors for the transitions.
C' extend B" by introducing gender, age, education and country of origin as predictors for both the initial state and the transitions, respectively.
C extends B by introducing gender, age, education and country of origin as predictors for both the initial state and the transitions.

As can be seen in Table 5.4 models C" and C appear to fit the data best. However, as the differences in the AIC and BIC between the two models are rather minimal and model C" is slightly less complex, it has been selected as our final model. The results from the comparison of the model fit statistics are similar to those of Pavlopoulos and Vermunt (2015) where model C" was also selected as the final model. This confirms that for a certain period of time the same model specification can be used to correct for measurement error.

5.4.3 The size of the measurement error

The size of measurement error in the survey and register data according to our analysis and that of Pavlopoulos and Vermunt (2015) is depicted in Tables 5.5 and 5.6 respectively. In order to estimate the error, the posterior probabilities of having a specific type of latent contract in each month have been used; those were estimated for all individuals included in the sample using the hidden Markov model.

In more detail, the tables report the classification error probabilities, which are represented by the probabilities $P(C_{i0} = c_0 \mid X_{i0} = x_0)$ and $P(C_{it} = c_t \mid X_{it} = x_t, X_{i(t-1)} = x_{t-1}, C_{i(t-1)} = c_{t-1})$ for the survey and register data, respectively.

Overall, both analyses produce very similar results and point to the same trends with regards to the level of measurement error indicating that the error is stable for this period of time. In other words, the analyses show that overall all three contract types are measured very accurately by the survey. The overall size of measurement error in the register data, on the other hand, appears very high especially for individuals holding a temporary contract.

Table 5.5- The size of the measurement error in the survey data according to Model C''

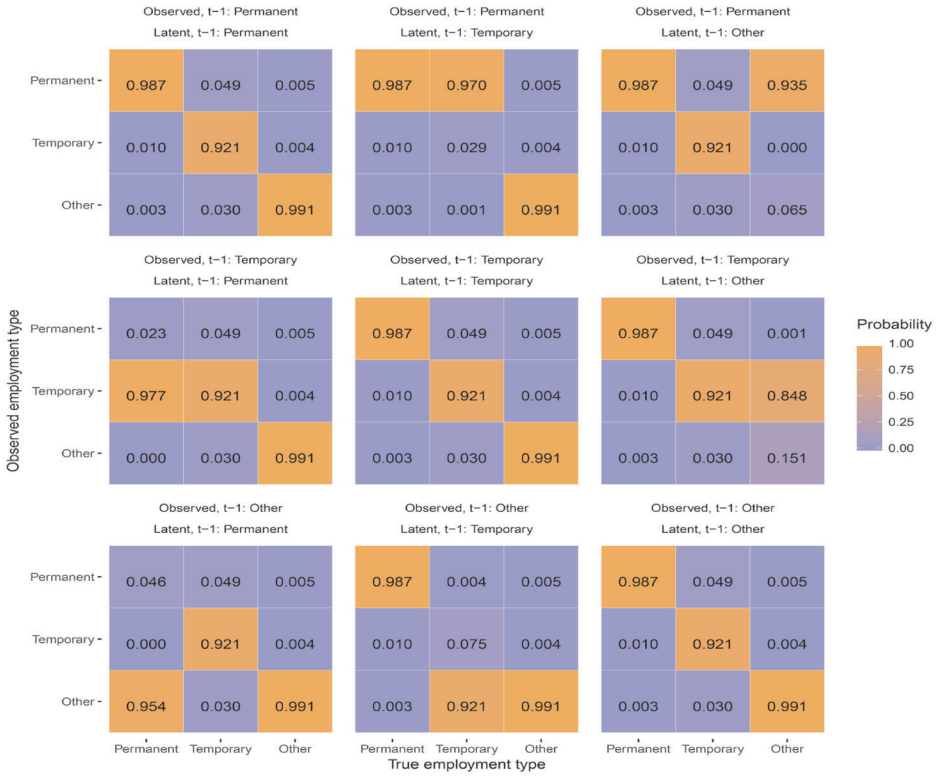
	Own analysis			**Pavlopoulos and Vermunt (2015)**		
	Observed contract in t			**Observed contract in t**		
Latent contract in t	Permanent	Temporary	Other	Permanent	Temporary	Other
Permanent	0.996	0.003	0.002	0.998	0.001	0.002
Temporary	0.090	0.878	0.033	0.125	0.832	0.042
Other	0.011	0.006	0.984	0.004	0.005	0.991

Given that the error probabilities in the register data are assumed to be serially correlated — by estimating an additional coefficient when a classification error was made in time point $t - 1$ and this error can be repeated in time t, we can extract more information on the structure of this error by studying more closely the conditional error probabilities. Figure 5.3 reports our estimates of the conditional probabilities of the error in the register data in time for all 9 combinations of latent and observed state in time $t - 1$. The 3 diagonal matrices represent the cases where no error in the register data was made in time $t - 1$, while off diagonal matrices represent the different cases of measurement error in $t - 1$. Figure 5.3 gives a completely different picture than Table 5.6.

The diagonal matrices — which are by construction identical — indicate that when no classification error is made in t - 1, the probability of an error in t is rather minimal.

Table 5.6- The size of the measurement error in the register data according to Model C''

	Own analysis			**Pavlopoulos and Vermunt (2015)**		
	Observed contract in t			**Observed contract in t**		
Latent contract in t	Permanent	Temporary	Other	Permanent	Temporary	Other
Permanent	0.877	0.106	0.017	0.888	0.081	0.031
Temporary	0.247	0.635	0.118	0.237	0.684	0.079
Other	0.033	0.013	0.954	0.032	0.017	0.951



Note: use of average posterior probabilities

Figure 5.3- Conditional probabilities of measurement error in register data according to Model C" (own analysis)

The conditional error probabilities obtained by Pavlopoulos and Vermunt (2015) are presented in Table 5.7. The results are almost the same as the ones reported by us and presented in Figure 5.3.

Table 5.7- Conditional probabilities of measurement error in the register data in time t when no error has been made in t - 1 according to the C'' model with fixed error parameters

	Latent contract in t		
Observed contract in t	Permanent	Temporary	Other
Permanent	0.986	0.045	0.005
Temporary	0.009	0.930	0.005
Other	0.004	0.025	0.990

The left-hand side of Table 5.8 extracts from Figure 5.3 the probabilities that an error repetition is possible. All the error probabilities when an error repetition is possible are extremely high. The relevant probabilities from Pavlopoulos and Vermunt (2015), as presented on the right-hand side of the table, are very similar to our results.

Table 5.8 - Conditional probabilities of repeating an error in time t that has been made in t - 1

	Own analysis			**Pavlopoulos and Vermunt (2015)**		
	Observed contract in t			**Observed contract in t**		
Latent contract in t	**Permanent**	**Temporary**	**Other**	**Permanent**	**Temporary**	**Other**
Permanent		0.977	0.954		0.973	0.961
Temporary	0.970		0.921	0.968		0.896
Other	0.935	0.848		0.913	0.842	

The last remaining situation to examine is to study the probability for a different classification error in time t when a classification error is made in t - 1. These probabilities are presented in Table 5.9 for the three latent states and for both our own data and those of Pavlopoulos and Vermunt (2015). All these probabilities are rather small and similar to those where no error is made in t - 1. The probabilities are also almost identical between the two studies.

Table 5.9- Conditional probabilities of making an error in time t that is different from the error made in t - 1

Own data			**Pavlopoulos and Vermunt (2015)**		
Latent contract in t			**Latent contract in t**		
Permanent	**Temporary**	**Other**	**Permanent**	**Temporary**	**Other**
0.013	0.079	0.009	0.014	0.070	0.001

Note: the probabilities of the last row come from table 4.3 of Pavlopoulos and Vermunt (2015).

Thus, the estimates of the conditional probabilities of the measurement error in the register data show a clear picture. The large size of the error that was illustrated in Table 5.6 is only due to the error in the initial registration of the contract type in the register. Once a mistaken value for the contract type is entered, then this will be carried

over almost for sure for many months. However, if a correct entry is made, then the probability of an error in the subsequent months is very small.

Overall, we can conclude that the nature and size of the measurement error in both the survey and register data appear very similar in 2007 (analyzed by Pavlopoulos and Vermunt, 2015) and in 2009 (as shown by us). The stability of the measurement error for this period of time enables us to apply the aforementioned error correction method in which we fix the error parameters according to the results obtained by Pavlopoulos and Vermunt (2015) in the analysis of our own data from 2009. This in turn allows us to correct for measurement error without having to undertake the full HMM analysis. The accuracy of this method when estimating first- and second- order statistics is explored below.

5.4.4 First-order statistics: The size of temporary employment

The latent distribution of the contract types, approximated according to our analysis and when substituting in the measurement error specific parameters from Pavlopoulos and Vermunt (2015), is presented in Table 5.10 and is contrasted with the observed distributions of the contract type according to the survey and register data, respectively. As in the case of the estimation of the average size of the measurement error, this has been carried out by using the average posterior probabilities of individuals holding a certain type of latent contract.

Table 5.10- The average size of temporary employment according to Model C''

	Survey	Register	Latent- own analysis	Fixing error parameters to those in Pavlopoulos and Vermunt (2015)
Permanent	0.653	0.585	0.611	0.613
Temporary	0.110	0.151	0.128	0.131
Other	0.237	0.264	0.261	0.257
Cases	36,321	130,671	133,290	133,290

As can be seen from the table, the results of our own analysis are almost identical to those using the fixed error parameters. Furthermore, the latent probability of belonging to a certain state always lies between the observed probabilities coming from the two data sources. Specifically, the latent probability of having a temporary contract equals approx. 13 percent for both analyses and is higher than is reported by the survey data while lower than reported by the register data (11.1 percent and 15.1 percent, respectively). The latent probability of being employed with a permanent contract (approx. 61 percent), is lower than suggested by the survey data (65.3 percent) while higher than suggested by the register data (58.5 percent). Finally, the latent probability of belonging to the "other" state equals approximately 26 percent and lies also in between the figures estimated using the survey and register data (23.7 percent and 26.4 percent, respectively).

This conveys good news for official statistics. In the presence of measurement error in our data, a macro integration of two data sources — even by using a crude measure such as the average of the two observed probabilities — can produce reliable results for the size of temporary employment.

5.4.5 Second-order statistics: The transition probabilities

Besides providing a reliable estimate of the size of temporary employment, the challenge for official statistics is to present a correct estimate of mobility from temporary employment. The dominant argument in the policy debate is that although temporary employment is inferior to permanent employment, it provides an effective stepping stone to permanent employment. For this argument to be true, mobility rates from temporary to permanent employment should be high. Table 5.11 presents the average latent transition probabilities between the various states associated with the three contract types. These transition probabilities have been calculated using model C'' in such a way that they refer to a 3-month period and are an average of the twelve 3-month periods that are included in the dataset.

Table 5.11- Latent 3-months transitions according to model C''

	Own analysis			**Fixing error parameters to those in Pavlopoulos and Vermunt (2015)**		
Contract in t-3	Permanent	Temporary	Other	Permanent	Temporary	Other
Permanent	0.987	0.004	0.009	0.989	0.004	0.008
Temporary	0.017	0.929	0.054	0.016	0.928	0.056
Other	0.006	0.030	0.963	0.006	0.029	0.965
Total	0.610	0.128	0.263	0.610	0.132	0.258

When looking at the estimates presented in Table 5.11, it can be once more noted that the two analyses provide almost identical results. Furthermore, when analyzing the transition rates in combination with those presented in Table 5.3 — i.e. the observed transition rates based on the survey and register data — it can be inferred that the latent transition rate from temporary to permanent employment is much lower than those estimated using both the survey and register data. That is, while according to the survey and register data out of all temporary employees in time $t - 3$, 5.8 percent and 7.3 percent respectively obtain a permanent contract in time t, our analysis suggests that this is only true for 1.6–1.7 percent of all temporary workers.

These findings suggest that a simple macro integration of the two data sources, which typically aims at the reconciliation of the distribution of the variable of interest in the two data sources at a given point in time, cannot provide reliable estimates of second-order statistics, namely mobility from temporary to permanent employment. These transition rates are overestimated by both the survey and the register data although these datasets contain a very different size and structure of measurement

error. Therefore, macro-integration would possibly not lead to transition rates lower than both sources.

5.5 Conclusions

National Statistical Institutes often have more than one indicator available for the same variable. The development of register data means that, increasingly, information on survey respondents can also be traced at the administrative level. This offers new opportunities to NSIs as they can corroborate findings from one data source using the other. However, these opportunities present new challenges that ought to be addressed.

As all data sources contain some measurement error, discrepancies emerge between the data sources in the measurement of a single variable even for the same individuals. Measurement error leads to bias in the estimates for first order statistics (estimates on one reference date) and second order statistics (estimates of transition rates between two reference dates).

Besides ignoring the problem or taking averages, these discrepancies are usually resolved by NSIs with the use of macro-integration. After separate integration of the stock data of two reference dates by harmonization, completion, and by forcing the data to meet certain identity relationships, on an aggregate level, a large part of the measurement error has been removed. However, the aggregate transitions rates are only corrected by the application of one identity relationship that the stock at reference date t plus all the transitions add up to the stock at $t + 1$. For this second step, most of the time only one source is used, the one that is assumed to be of superior quality. In practice this means that the first order statistics usually are close to the real values. However, the adjustment in the transitions is marginal when using only one identity relationship and only one source. Therefore, one can expect that the real transition rates, the second order statistics, differ more from the observed ones because not all measurement error could be removed.

In this paper, we study whether an alternative macro-integration method of the two datasets can produce more accurate results. In doing so we rely on the micro integration approach undertaken by Pavlopoulos and Vermunt (2015). This approach requires re-linkage of data and re-estimation of the model at every time interval. For this reason, it is considered extremely time-consuming and expensive by Statistics Netherlands. We therefore investigate whether estimates obtained by using this approach are invariable to time and therefore can be re-used in later time points without the need to re-link the datasets and re-estimate the statistical model.

Our results indicate that the size of the error in the measurement of the employment contract in the LFS and the ER is indeed stable over time. Therefore, HMMs provide a way to develop a powerful macro-integration method; as measurement error rates can be considered time constant, we can develop an error correction method, based on

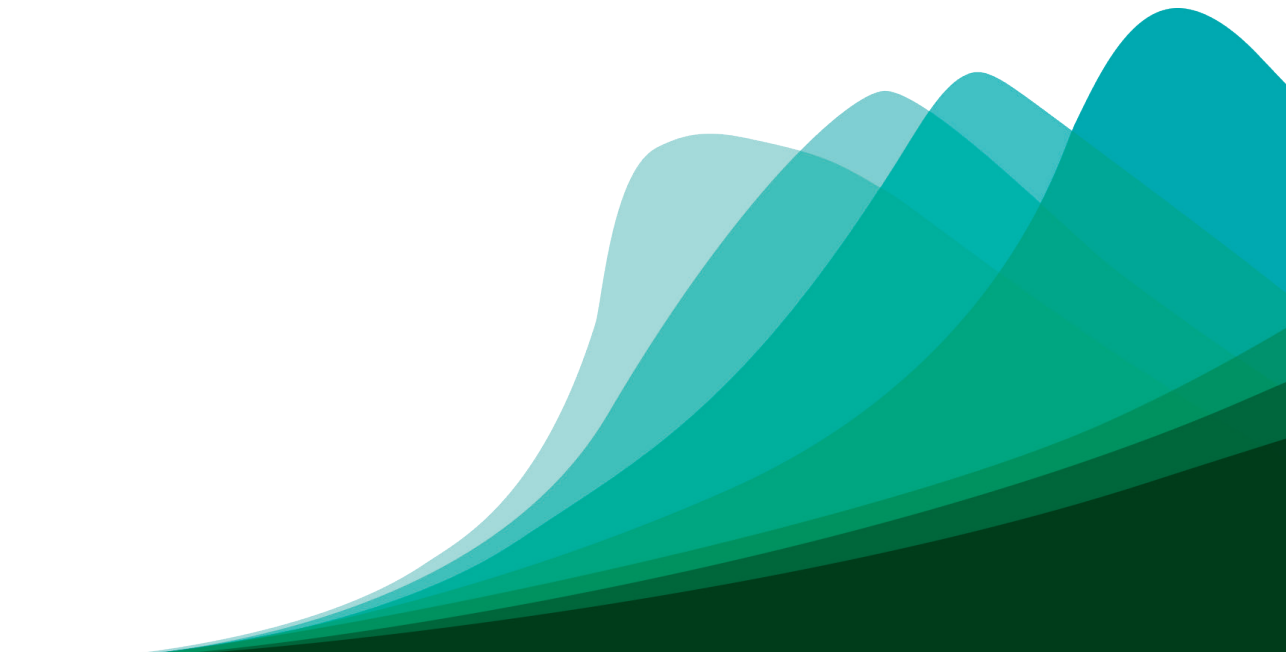
the use of fixed parameters from an initial HMM analysis, that can be easily included in the production of official statistics.

The method can be considered superior to traditional macro-integration approaches in particular for second order statistics. That is, while the findings suggest that first-order statistics can be approximated using traditional macro-integration, this is probably not the case when estimating second-order statistics. In more detail, the size of temporary employment in the Netherlands always lies between the estimates from the two data sources. Therefore, traditional macro-integration can easily provide rather accurate estimates of these statistics. In contrast, findings on second-order statistics indicate a different picture; according to the survey and register data, 5.8 percent and 7.3 percent of workers with temporary contracts are employed with permanent contracts 3 months later. This indicates a substantial amount of mobility in the labor market. However, our HMM model suggests that this mobility is only 1.7 percent. Therefore, a static reconciliation approach, which is what typically traditional macro-integration methods do, is unlikely to provide an effective error correction of second order statistics.

However, a formal comparison of the outcomes of traditional macro-integration and macro-integration based on HMMs should confirm that the last method is superior to the first. Moreover, it would be interesting to test also a more advanced way of traditional macro-integration. Instead of doing the integration process in two steps (first, the integration of the stock data on the two reference dates and second, make the transitions consistent with the stock data) doing it in one go in such a way that the identity relationships on the two reference dates and the transitions are met. Further research will provide a formal comparison of our approach with traditional macro-integration approaches.

Nevertheless, before the macro-integration approach based on HMMs enters the production of official statistics, further issues have to be addressed. Our analysis has ignored the effect of a possible linkage error between the two data sources. Although the probability of a linkage error is very small in our data, this error is particularly important when it is correlated with the outcome variable — here, the type of contract — and, thus, might bias the results. Moreover, the analysis did not fully take into account the overtime changes in the way LFS is conducted, such as the transition from dependent to independent interviewing or to a different mode of interviewing (i.e. face-to-face vs. phone or online surveys). Those aspects are likely to impact the error in the survey data and, therefore, further analysis which will investigate those aspects is required.

6



SUMMARY, CONCLUSIONS, AND DISCUSSION

P. Pankowska is the sole author of this chapter but elements of it are based on: Pankowska, P., Bakker, B. F. M., Pavlopoulos, D. & Oberski, D. L. (2020). Reconciliation of inconsistent data sources using hidden Markov models. Manuscript accepted for publication in the Statistical Journal of the IAOS.

6.1 Summary and conclusions

This thesis focused on the problem of measurement error and investigated the feasibility of using hidden Markov models (HMMs) to correct for such error in categorical, longitudinal data. In doing so, we have first illustrated how measurement error poses a substantial threat to the validity and accuracy of estimates. We then demonstrated the need to use multiple-indicator HMM specifications, which can account for the nonignorable presence of systematic/dependent errors. Finally, we showed that the use of such extended models is feasible. That is, even though such HMMs require record linkage, linkage error is largely not a problem. Furthermore, while their implementation process is complex and time-consuming, it can be simplified because error parameters can be re-used for a number of years. The thesis includes four chapters of original research work.

In **Chapter 2**, we examined the bias introduced by measurement error, using clustering as an illustrative example. In more detail, in this simulation study, we investigated the sensitivity of two commonly used model- and density-based clustering algorithms (i.e. GMMs and DBSCAN) to varying severities and magnitudes of random and systematic errors. In doing so, we looked at the similarity of the clusters obtained in the presence of measurement error to those obtained in its absence. We also considered the effects on the number of clusters found, to infer whether measurement error leads to the emergence of spurious clusters and/or obscures clusters. Our analysis shows that measurement error in many cases leads to non-negligible bias as the returned clusters are (highly) dissimilar to the ones obtained when the dataset is error-free. The number of clusters found in the data is also affected by the error. These effects are particularly strong when the error is systematic as opposed to random, and when it affects all variables in the dataset.

In **Chapter 3,** we looked at how different data collection processes might impact the nature and magnitude of measurement error. We did this by studying how the switch from dependent interviewing (DI) to independent interviewing (INDI) in the Dutch Labour Force Survey (LFS) affected the random and systematic components of the error in this data source. For this purpose, we applied an extended, two-indicator HMM to linked LFS and Employment Register (ER) data and allowed for error dependency in both data sources. Our results indicate that the use of DI lowered the probability of obtaining random errors but had no significant effect on systematic errors. What is more, our results also show that regardless of the interviewing regime used, the survey data contains autocorrelated error, wherein the probability of repeating the same error, provided that an error was made in the previous time point, is extremely high. The error in the register data is also shown to follow the same pattern. The findings of this paper indicate that both data sources examined are subject to non-negligible systematic error which needs to be considered when correcting for measurement error using HMMs. This in turn confirms the need for using extended, multiple-indicator HMMs which allow for the relaxing of the local independence assumption and modelling, among

other things, error autocorrelation without risking poor model identifiability. What is more, our analysis also demonstrates that, apart from correcting for measurement error, HMMs can also be used to assess how various data collection processes impact data quality.

In **Chapter 4**, we investigated whether and to what extent the use of multiple-indicator HMMs, which often requires record linkage, leads to biased estimates due to the presence of linkage error. In doing so, we tested the sensitivity of the structural parameter estimates of a two-indicator HMM — i.e. the transition rates between employment contract types — to varying degrees and types of false-negative (incorrectly not linked records) and false-positive (erroneously linked records) linkage error. The results of our simulation study show that the sensitivity of the method to both types of linkage error is low. It appears that only scenarios wherein the error rate is high and the probability of exclusion/mislinkage is highly correlated with the target variable lead to substantial bias. Moreover, our results also show that under certain conditions, false-positive linkage error acts as another source of measurement error that is absorbed into the error-rate parameters of the model, leaving the latent transition estimates unaffected. In these cases, HMMs also accounts for linkage error.

In **Chapter 5**, we focused on a more practical matter and explored the feasibility of using multiple-indicator HMMs, given their complex nature and, with that, the time and costs associated with their implementation. More specifically, we looked at whether it is possible to simplify the error correction procedure by running the full analysis once and then re-using the resultant error parameters as a correction factor for a number of years. The proposed solution is contingent on the assumption that the structure and size of the error remain constant; if this assumption is violated the estimates obtained using this procedure might be biased. Our analysis provides some evidence that in the absence of a major change in the data collection process, the size and structure of the error are time-invariant and, therefore, the error parameters can be carried forward a certain number of years. More specifically, we show that using the full error-correction procedure and applying a two-indicator HMM to linked LFS and ER data from 2009 yields virtually the same results as fitting this model to the data with parameters obtained from the 2007 version of the same dataset.

6.2 Discussion

HMMs have considerable potential to be used as a measurement-error-correction technique and produce more accurate statistical estimates. In this thesis, we demonstrated the importance of accounting for such errors and provided some evidence confirming that HMMs could be applied to this end. We will now summarize the main contributions and implications of this thesis. In doing so, we will also discuss the limitations of our research and provide suggestions for future research.

The presence and effects of measurement error

In the first part of the thesis, we focused on the biasing effects of measurement error and showed that it has the potential to strongly impact statistical estimates. The analysis conducted in Chapter 2 demonstrated that measurement error can strongly bias the results of cluster analysis, a phenomenon that has not been extensively studied within the clustering literature. This in turn confirms the need for more in-depth research focusing on the effects of measurement error on various statistical analyses.

A thus far understudied angle that is particularly worth pursuing going forward is the presence and effects of systematic error specifically. This suggestion is motivated by the results of our analysis which showed that this type of error tends to have a more severe effect than random error. This finding, combined with the results of Chapter 3, which confirm the presence of considerable error dependency in both the survey and register data (in addition to random error), implies that systematic error poses a substantial and realistic threat to the validity of estimates. Therefore, it is crucial to understand, test for and model such dependency, rather than assuming the error to be (predominantly) random.

The use of HMMs to correct for measurement error

Given the above, we then investigated whether it is plausible to use multiple-indicator HMMs to correct for both independent and dependent errors. The results of Chapter 4 paint a rather optimistic picture and show that although the use of multiple-indicator models requires performing record linkage, which itself might lead to linkage error, (the structural parameters of) HMMs are largely robust to this type of error. However, it is worthwhile noting that the integration of several sources might also lead to additional errors that can potentially introduce bias. Zhang (2012) and Bakker (2011) provide an overview of the errors associated with the use of such data. In doing so, the authors extend the Total Survey Error Framework proposed by Groves et al. (2011), which offers a systematic review of survey errors, to also account for errors that occur when using combined survey and register data. This extension assumes that the life cycle of integrated data sources consists of two phases. The first phase focuses on the errors occurring in each of the data sources separately (and includes measurement error) and the second phase concerns the errors associated with the integration process specifically. In more detail, the use of integrated data sources, apart from potentially resulting in linkage error, might also lead to the misalignment of variable definitions (as a result of variable harmonization) and to mapping errors. The latter are associated with the reclassification step and might occur when the primary, input-source measures do not clearly correspond to standard definitions. As a result, when harmonizing these measures their re-classification to standard categories may contain errors (Zhang, 2012). However, both these types of errors are forms of measurement error/misclassification that are already corrected for by the model. Therefore, it can be inferred that the application of multiple indicator HMMs to data sources should not introduce new potential sources of bias, as long as the model assumptions are not violated.

In addition, the findings of Chapter 5 provide some evidence that, while the proposed method is rather complex and time-consuming, it can be simplified and therefore applied as an error-correction procedure in practice (provided that the error size and structure are constant for the time period under consideration).

It is important to note, though, that the findings are based on an analysis that used only one sample and considered a relatively short time period, i.e. we looked at carrying forward estimates from the first quarter of 2007 to the first quarter of 2009 only. Therefore, to confirm the robustness of the findings, the analysis should be repeated using different linked datasets as well as varying time intervals. What is more, while the analysis of Chapter 3 confirms that a significant modification of the data collection process is likely to affect the size and/or structure of the error, the sensitivity of error parameters to various other changes, which might differ in terms of severity, should also be investigated. Examples of such changes include the transition from Computer-Assisted Personal or Telephone Interviewing (CAPI or CATI, respectively) to Computer-Assisted Web Interviewing (CAWI).

It is also worthwhile mentioning that, in order for the findings to be more generalizable, the performance and feasibility of the proposed method should be tested in a different context that goes beyond the topic of labor mobility and uses other variables than the individual's contract type. It would be also interesting to use data from different countries than the Netherlands and, if possible, additional sources apart from surveys and administrative registers.

Aspects that require further investigation

There are two main additional aspects that have not been examined directly in this thesis but need to be considered before the HMM-based approach could be applied in practice: model robustness and the use of error-corrected microdata in further analysis. In more detail, a thorough examination of the sensitivity of parameter estimates to varying model specifications containing different assumptions ought to be carried out. This is of particular importance as in practice it is extremely difficult, if not impossible, to use a model specification that accounts for all possible error sources and scenarios. Therefore, we need to understand whether and to what extent different violations of assumptions can bias the results obtained. To illustrate, in all of our analyses the measurement error probabilities in the survey and the register data are assumed to be time- invariant and homogenous. However, this might not be the case as the likelihood of making an error in both sources might change over time. For instance, this likelihood might decrease in the survey data as a result of interviewers getting accustomed to a questionnaire when using it for numerous consecutive waves; in the register data it might decline due to companies becoming more familiar with the software that is utilized to input, store and manage register data. What is more, the error probabilities might be heterogeneous and depend on data collection and/or individual-level characteristics. In these scenarios, a model specification that relies on the assumption that the error parameters are time-invariant and/or homogeneous is

incorrect and it is necessary to investigate the effect of such violations on the HMM estimates. Similarly, the sensitivity of the model estimates to the violation of the 1st order Markov assumption should also be examined.

Second, it is also important to consider how researchers can use error-corrected microdata in their analyses, while accounting for the uncertainty of the "true state" membership. One possible solution is to combine multiple imputation and latent Markov modelling, as proposed by Boeschoten et al. (2020). The use of multiply imputed datasets allows for the relatively straightforward estimation of various statistics of interest, while taking into account the uncertainty of the assignment of values to the latent variable that is caused by measurement error. It is also possible to generalize the method proposed by Boeschoten et al. (2020) and use reweighting. That is, each record in the dataset can be replaced by X records, where X corresponds to the number of latent categories. To illustrate, in our case each record in the linked LFS-ER dataset would be replaced with three records and the contract type variable for these records would be permanent, temporary and other, respectively. Then, the posterior probabilities of having permanent, temporary and other (latent) contract types for each record would be assigned as weights to the corresponding newly created records.

Alternatively, it might also be possible to run the measurement error procedure and the substantive analysis simultaneously. This can be done by extending the structural part of the HMM so that the resultant estimates answer the research question at hand. For instance, rather than first correcting for measurement error and then performing sequence analysis on the obtained data, one can make use of a mixture hidden Markov model (MHMM) that allows for both the correction of the error and the clustering of sequences (Helske et al., 2018).

6.3 Using HMMs to reconcile inconsistent data sources in official statistics

The results and findings presented in this thesis have strong implications for the feasibility of using extended HMMs to reconcile inconsistent data sources in official statistics specifically.

National Statistical Institutes (NSIs) often obtain information on the same phenomena from different data sources (Bakker, 2011; Van Delden et al., 2016). Even though these sources are in most cases subject to editing, which is used to detect and correct erroneous values (De Waal, 2016; Van Delden et al., 2016), identical units do not always yield identical values (Guarnera & Varriale, 2016). Such inconsistencies are mainly the result of measurement error in the data sources involved and are likely to lead to the unwanted publication of differing statistics.

The effect of measurement error on official statistics varies depending on the type of estimates published. To illustrate, random measurement error specifically does not tend to substantially bias "first-order" population estimates, such as means, proportions, and totals, but does, in most cases, severely overestimate (or less often, underestimate)

"second-order" statistics, such as (over-time) transition rates, hazard rations, or domain mean differences (Bolck et al., 2004; Bound et al., 2001; Pavlopoulos et al., 2012). Random error has also been shown to attenuate measures of associations between variables, such as correlations and linear regression coefficients (Liu, 1988).

NSIs apply several methods to account for the inconsistencies caused by measurement error. Most commonly, the differences are ignored and the estimates published are based on edited data coming only from the source that is assumed to have superior quality (De Waal et al., 2019). Alternatively, NSI's use weighting as well as micro- and macro-integration methods to obtain consistent estimates from different sources. These three methods differ with regards to the level of consistency achieved as well as the costs required for their implementation (De Waal et al., 2019).

When using weighting to achieve higher consistency, survey records are weighted using the totals of the register source (Särndal et al., 2003). For this solution, it is not necessary to link the sources on a micro-level, but rather it is sufficient to use the cross table of the weighting variables from the register source and weight the survey data using the marginals. An alternative approach is to use micro-integration, wherein the sources are first linked on the individual level and next the quality of the data is improved by identifying and correcting for errors at the unit level (Bakker, 2011; van Rooijen et al., 2016). Finally, the problem of inconsistencies can also be resolved using macro-integration, a process in which statistical outcomes are reconciled on the aggregate level. In macro-integration, the differences between the target and observed populations as well as the target variables and their measurements are first explained and then corrected for by using estimates from other sources or the knowledge of subject matter experts.

The methods discussed differ substantially with regards to the labor intensiveness and costs associated with their implementation. Weighting is a relatively inexpensive and easy to implement technique, which does not require data linkage; it is therefore often used by NSI's. Micro-integration, on the other hand, is significantly more labor- and cost- intensive. More specifically, determining the right edit rules and verifying the quality of the measured variables as well as performing record linkage requires a lot of time and effort. What is more, having developed the set of edit rules, its maintenance also requires substantial capacity, particularly when sources change. The costs of macro-integration are also relatively high, especially when subject matter experts play an important role. If the process is fully automated, though, it tends to be cheaper than micro-integration.

The results presented in this thesis suggest that HMMs can be potentially used as an alternative to the methods discussed when reconciling inconsistencies arising from measurement error in categorical, longitudinal data. To restate, we show that while this method, unlike weighting or macro-integration, requires record linkage, linkage error is not a major concern. We also show how this method can be simplified by carrying forward error parameter estimates. However, this can only be done in the absence of a major change in the data collection process, which implies that this method can still be

rather expensive, particularly if the data collection process of a survey or the laws and regulations impacting register data quality change frequently. The decision of whether this method should be applied in the production of official statistics depends then on the expected frequency of the aforementioned changes (i.e. the costs involved) and the importance of obtaining consistent and error-corrected variables for the users of official statistics (i.e. the revenues).

Another important aspect that should be taken into consideration is the fact that all the models that we refer to use linked data. This means that, if one of the sources used is much richer than the other (i.e. it contains more individuals and/or more time points per individual), such as is the case with register data compared to survey data, this method will lead to loss of information, as it only uses data available in both sources. Moreover, if the survey data suffer from selective non-response, the estimated measurement error can be biased too. For this reason, NSI's might prefer to use macro-integration or reweighting techniques as these methods use all the data available rather than just a linked subset. Further research should, therefore, look into the possibility of combining the aforementioned methods with hidden Markov modelling. In such a combined method, HMMs could be used to obtain estimates of measurement error from the linked data, while the final corrected (substantive) estimates could be based on all the data available and obtained using macro-integration or reweighting techniques.

To summarize, this thesis has laid down the groundwork for the use of HMMs as an error correction procedure. This methodology can be potentially used to reconcile inconsistent data sources by NSIs. However, before this method can be applied in practice, both specifically in official statistics and more broadly in other disciplines, further work is needed.

References

Aggarwal, C. C. (Ed.) (2010). *Managing and mining uncertain data* (Vol. 35). Springer. https://doi.org/10.1007/978-0-387-09690-2

Aggarwal, C., & Reddy, C. (Eds.) (2013). *Data clustering: Algorithms and applications* (1st ed.). CRC Press.

Aldenderfer, M., & Blashfield, R. (1984). *Cluster analysis*. Sage publications. https:// doi.org/10.4135/9781412983648

Alwin, D. F. (2007). *Margins of error: A study of reliability in survey measurement* (Vol. 547). John Wiley & Sons. https://doi.org/10.1002/9780470146316

Alwin, D. F., Baumgartner, E. M., & Beattie, B. A. (2018). Number of response categories and reliability in attitude measurement. *Journal of Survey Statistics and Methodology*, *6*(2), 212–239. https://doi.org/10.1093/jssam/smx025

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

Armstrong, J., & Mayda, J. (1993). Model-based estimation of record linkage error rates. *Survey Methodology*, *19*(2), 137–147. https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X199300214459

Bailey, K. D. (1975). Cluster analysis. *Sociological Methodology*, *6*, 59–128. https://doi.org/10.2307/270894

Bakker, B. F. M., & Daas, P. J. (2012). Methodological challenges of register-based research. *Statistica Neerlandica*, *66*(1), 2–7. https://doi.org/10.1111/j.1467-9574.2011.00505.x

Bakker, B. F. M. (2011). *Micro-integration.* (Methods). Statistics Netherlands. https://www.cbs.nl/en-gb/our-services/methods/statistical-methods/throughput/throughput/-/media/40d29d3f5dbe4cf58725b6835f33bf53.ashx

Bakker, B. F. M. (2012). Estimating the validity of administrative variables. *Statistica Neerlandica*, *66*(1), 8–17. https://doi.org/10.1111/j.1467-9574.2011.00504.x

Banfield, J. D., & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49(3), 803–821. https://doi.org/10.2307/2532201

Bassi, F., Hagenaars, J. A., Croon, M. A., & 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*(2), 230–268. https://doi.org/10.1177/0049124100029002003

Bergin, A. (2013). *Job changes and wage changes: Estimation with measurement error in a binary variable*. National University of Ireland Maynooth. http://mural.maynoothuniversity.ie/4434/

Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, *10*(2–3), 191–203. https://doi.org/10.1016/0098-3004(84)90020-7

Biemer, P. P. (2004). An analysis of classification error for the revised current population survey employment questions. *Survey Methodology*, *30*(2), 127–140. https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X20040027747

Biemer, P. P. (2011). *Latent class analysis of survey error* (Vol. 571). John Wiley & Sons. https://doi.org/10.1002/9780470891155

Biemer, P. P., & Bushery, J. M. (2000). On the validity of Markov latent class analysis for estimating classification error in labor force data. *Survey Methodology*, *26*(2), 139–152. https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X20000025534

Biemer, P. P., Groves, R. M., Lyberg, L. E., Mathiowetz, N. A., & Sudmanr, S. (Eds.) (1991). *Measurement errors in surveys.* John Wiley & Sons. https://doi.org/10.1002/9781118150382

Biemer, P. P., de Leeuw, E. D., Eckman, S., Edwards, B., Kreuter, F., Lyberg, L. E., Tucker, N. C., & West, B. T. (Eds.) (2017). *Total survey error in practice*. John Wiley & Sons. https://doi.org/10.1002/9781119041702

Biemer, P. P., & Wiesen, C. (2002). Measurement error evaluation of self-reported drug use: A latent class analysis of the US National Household Survey on Drug Abuse. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *165*(1), 97–119. https://doi.org/10.1111/1467-985X.00612

Biemer, P. P., & Stokes, S. L. (2004). Approaches to the modelling of measurement errors. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 485–516). John Wiley & Sons. https://doi.org/10.1002/9781118150382

Billiet, J. B., & Davidov, E. (2008). Testing the stability of an acquiescence style factor behind two interrelated substantive variables in a panel design. *Sociological Methods & Research*, *36*(4), 542–562. https://doi.org/10.1177/0049124107313901

Bingley, P., & Martinello, A. (2014). *Measurement error in the Survey of Health, Ageing and Retirement in Europe: A validation study with administrative data for education level, income and employment.* SHARE - Survey of Health, Ageing and Retirement in Europe. https://www.vive.dk/en/publications/measurement-error-in-the-survey-of-health-ageing-and-retirement-in-europe-5219/

Birant, D., & Kut, A. (2007). ST-DBSCAN: An algorithm for clustering spatial–temporal data. *Data & Knowledge Engineering*, *60*(1), 208–221. https://doi.org/10.1016/j.datak.2006.01.013

Blakely, T., & Salmond, C. (2002). Probabilistic record linkage and a method to calculate the positive predictive value. *International Journal of Epidemiology*, *31*(6), 1246–1252. https://doi.org/10.1093/ije/31.6.1246

Boeschoten, L. (2019). *Consistent estimates for categorical data based on a mix of administrative data sources and surveys* [Doctoral dissertation, Tilburg University]. Gildeprint. https://pure.uvt.nl/ws/portalfiles/portal/31415117/Boeschoten_Consistent_25_10_2019.pdf

Boeschoten, L., Filipponi, D., & Varriale, R. (2020). Combining multiple imputation and hidden Markov modelling to obtain consistent estimates of employment status. *Journal of Survey Statistics and Methodology,* Article smz05. https://doi.org/10.1093/jssam/smz052

Bohensky, M. A., Jolley, D., Sundararajan, V., Evans, S., Pilcher, D. V., Scott, I., & Brand, C. A. (2010). Data linkage: A powerful research tool with potential problems. *BMC Health Services Research*, *10*, Article 346. https://doi.org/10.1186/1472-6963-10-346

Bolck, A., Croon, M., & Hagenaars, J. (2004). Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political Analysis*, *12*(1), 3–27. https://doi.org/10.1093/pan/mph001

Bound, J., Brown, C., & Mathiowetz, N. (2001). Measurement error in survey data. In J. Heckman & E. Leamer (Eds.), *Handbook of econometrics* (Vol. 5, pp. 3705–3843). Elsevier. https://econpapers.repec.org/bookchap/eeeeconhb/5.htm

Bouveyron, C., Celeux, G., Murphy, T. B., & Raftery, A. E. (2019). *Model-based clustering and classification for data science: With applications in R* (Vol. 50). Cambridge University Press. https://doi.org/10.1017/9781108644181

Brüderl, J., Castiglioni, L., Ludwig, V., Pforr, K., & Schmiedeberg, C. (2017). Collecting event history data with a panel survey: Combining an electronic event history calendar and dependent interviewing. *Methods, Data, Analyses*, *11*(1), 45-66. https://doi.org/10.12758/mda.2016.013

Carroll, R. J., Midthune, D., Freedman, L. S., & Kipnis, V. (2006). Seemingly unrelated measurement error models, with application to nutritional epidemiology. *Biometrics*, *62*(1), 75–84. https://doi.org/10.1111/j.1541-0420.2005.00400.x

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

Chambers, R. (2009). *Regression analysis of probability-linked data*. Statistics New Zealand. https://pdfs.semanticscholar.org/f2d1/28fd449d62edf029b897b6015320a49757bf.pdf?_ga=2.27540540.1085603425.1585492829-1474965019.1585492829

Chambers, R., & Kim, G. (2015). Secondary analysis of linked data. In K. Harron, H. Goldstein & C. Dibben (Eds.), *Methodological Developments in Data Linkage* (pp. 83–108). John Wiley & Sons. https://doi.org/10.1002/9781119072454

Conrad, F. G., Rips, L. J., & Fricker, S. S. (2009). Seam effects in quantitative responses. *Journal of Official Statistics*, *25*(3), 339–361. https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/seam-effects-in-quantitative-responses.pdf

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* ERIC.

Dave, R. N. (1991). Characterization and detection of noise in clustering. *Pattern Recognition Letters*, *12*(11), 657–664. https://doi.org/10.1016/0167-8655(91)90002-4

Davé, R. N., & Krishnapuram, R. (1997). Robust clustering methods: A unified view. *IEEE Transactions on Fuzzy Systems*, *5*(2), 270–293. https://doi.org/10.1109/91.580801

De Waal, T. (2016). Obtaining numerically consistent estimates from a mix of administrative data and surveys. *Statistical Journal of the IAOS*, *32*(2), 231–243. https://doi.org/10.3233/SJI-150950

De Waal, T., Pannekoek, J., & Scholtus, S. (2011). *Handbook of statistical data editing and imputation*. John Wiley & Sons. https://doi.org/10.1002/9780470904848

De Waal, T., Van Delden, A., & Scholtus, S. (2019). Multi-source statistics: Basic situations and methods. *International Statistical Review, 0*(0),1–26. https://doi.org/10.1111/insr.12352

Di Consiglio, L., & Tuoto, T. (2018). When adjusting for the bias due to linkage errors: A sensitivity analysis. *Statistical Journal of the IAOS*, *34*(4), 589–597. https://doi.org/10.3233/SJI-170377

Edwards, S. L., Berzofsky, M. E., & Biemer, P. P. (2017). Effect of missing data on classification error in panel surveys. *Journal of Official Statistics*, *33*(2), 551–570. https://doi.org/10.1515/jos-2017-0026

Eggs, J., & Jäckle, A. (2015). Dependent interviewing and sub-optimal responding. *Survey Research Methods*, *9*(1), 15-29. https://doi.org/10.18148/srm/2015.v9i1.5860

Elnahrawy, E., & Nath, B. (2003). Online data cleaning in wireless sensor networks. SenSys '03: *Proceedings of the 1st International Conference on Embedded Networked Sensor Systems*, 294–295. https://doi.org/10.1145/958491.958527

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 226–231. https://dl.acm.org/doi/10.5555/3001460.3001507

Everitt, B., & Skrondal, A. (2002). *The Cambridge dictionary of statistics* (4th ed.). Cambridge University Press Cambridge.

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, *64*(328), 1183–1210. https://doi.org/10.1080/01621459.1969.10501049

Fienberg, S. E., & Gilbert, J. P. (1970). The geometry of a two by two contingency table. *Journal of the American Statistical Association*, *65*(330), 694–701. https://doi.org/10.2307/2284580

Figueiredo Filho, D. B., da Rocha, E. C., da Silva Júnior, J. A., Paranhos, R., da Silva, M. B., & Duarte, B. S. F. (2014). Cluster analysis for political scientists. *Applied Mathematics*, *5*(15), 2408. https://doi.org/10.4236/am.2014.515232

Flaherty, B. P. (2002). Assessing reliability of categorical substance use measures with latent class analysis. *Drug and Alcohol Dependence*, *68*(Suppl.), 7–20. https://doi.org/10.1016/S0376-8716(02)00210-7

Frigui, H., & Krishnapuram, R. (1996). A robust algorithm for automatic extraction of an unknown number of clusters from noisy data. *Pattern Recognition Letters*, *17*(12), 1223–1232. https://doi.org/10.1016/0167-8655(96)00080-3

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer. https://doi.org/10.1007/978-0-387-35768-3

Fuller, W. A. (2009). *Measurement error models* (Vol. 305). John Wiley & Sons.

Galimard, J.-E., Chevret, S., Protopopescu, C., & Resche-Rigon, M. (2016). A multiple imputation approach for MNAR mechanisms compatible with Heckman's model. *Statistics in Medicine*, *35*(17), 2907–2920. https://doi.org/10.1002/sim.6902

Gallegos, M. T., & Ritter, G. (2005). A robust method for cluster analysis. *The Annals of Statistics*, *33*(1), 347–380. https://doi.org/10.1214/009053604000000940

García-Escudero, L. A., Gordaliza, A., Matrán, C., & Mayo-Iscar, A. (2008). A general trimming approach to robust cluster analysis. *The Annals of Statistics*, *36*(3), 1324–1345. https://doi.org/10.1214/07-AOS515

Georgiadis, M. P., Johnson, W. O., Gardner, I. A., & 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*(1), 63–76. https://doi.org/10.1111/1467-9876.00389

Goldberger, J., & Ben-Reuven, E. (2016, April 24-26). *Training deep neural-networks using a noise adaptation layer* [Poster presentation]. 5th International Conference on Learning Representations, Toulon, France. https://openreview.net/forum?id=H12GRgcxg

Goldstein, H., Harron, K., & Wade, A. (2012). The analysis of record-linked data using multiple imputation with data value priors. *Statistics in Medicine*, *31*(28), 3481–3493. https://doi.org/10.1002/sim.5508

Gottschalk, P. (2005). Downward nominal-wage flexibility: Real or measurement error? *Review of Economics and Statistics*, *87*(3), 556–568. https://doi.org/10.1162/0034653054638328

Goyat, S. (2011). The basis of market segmentation: A critical review of literature. *European Journal of Business and Management*, *3*(9), 45–54. https://www.iiste.org/Journals/index.php/EJBM/article/view/647/540

Grace, Y. Y. (2017). *Statistical analysis with measurement error or misclassification*. Springer.

Groen, J. A. (2012). Sources of error in survey and administrative data: The importance of reporting procedures. *Journal of Official Statistics (JOS)*, *28*(2), 173–198. https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/sources-of-error-in-survey-and-administrative-data-the-importance-of-reporting-procedures.pdf

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2011). *Survey methodology* (2nd ed.). John Wiley & Sons.

Guarnera, U., & Varriale, R. (2016). Estimation from contaminated multi-source data based on latent class models. *Statistical Journal of the IAOS*, *32*(4), 537–544. https://doi.org/10.3233/SJI-150951

Gustafson, P. (2003). *Measurement error and misclassification in statistics and epidemiology: Impacts and Bayesian adjustments*. CRC Press.

Hadgu, A., Dendukuri, N., & Wang, L. (2012). Evaluation of screening tests for detecting Chlamydia trachomatis: Bias associated with the patient-infected-status algorithm. *Epidemiology*, *23*(1), 72–82. https://doi.org/10.1097/EDE.0b013e31823b506b

Hagenaars, J. A. (1988). Latent structure models with direct effects between indicators: Local dependence models. *Sociological Methods & Research*, *16*(3), 379–405. https://doi.org/10.1177/0049124188016003002

Hagenaars, J. A. (1990). *Categorical longitudinal data: Log-linear panel, trend, and cohort analysis*. Sage Publications.

Hagenaars, J. A. (1994). Latent variables in log-linear models of repeated observations. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (p. 329–352). Sage Publications.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). *Multivariate data analysis* (7th ed.). Pearson.

Hariri, J. G., & Lassen, D. D. (2017). Income and outcomes: Social desirability bias distorts measurements of the relationship between income and political behavior. *Public Opinion Quarterly*, *81*(2), 564–576. https://doi.org/10.1093/poq/nfw044

Harron, K. L., Doidge, J. C., Knight, H. E., Gilbert, R. E., Goldstein, H., Cromwell, D. A., & van der Meulen, J. H. (2017). A guide to evaluating linkage quality for the analysis of linked data. *International Journal of Epidemiology*, *46*(5), 1699–1710. https://doi.org/10.1093/ije/dyx177

Helske, S., Helske, J., & Eerola, M. (2018). Combining sequence analysis and hidden Markov models in the analysis of complex life sequence data. In G. Ritschard & M. Studer (Eds.), *Sequence Analysis and Related Approaches* (pp. 185–200). Springer. https://doi.org/10.1007/978-3-319-95420-2

Hennig, C. (2007). Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis*, *52*(1), 258–271. https://doi.org/10.1016/j.csda.2006.11.025

Hennig, C. (2010). Methods for merging Gaussian mixture components. *Advances in Data Analysis and Classification*, *4*, 3–34. https://doi.org/10.1007/s11634-010-0058-3

Hennig, C. (2015). What are the true clusters? *Pattern Recognition Letters*, *64*, 53–62. https://doi.org/10.1016/j.patrec.2015.04.009

Hoogendoorn, A. W. (2004). A questionnaire design for dependent interviewing that addresses the problem of cognitive satisficing. *Journal of Official Statistics*, *20*(2), 219. http://www.sverigeisiffror.scb.se/contentassets/ff271eeeca694f47ae99b942de61df83/a-questionnaire-design-for-dependent-interviewing-that-addresses-the-problem-of-cognitive-satisficing.pdf

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*, 193–218. https://doi.org/10.1007/BF01908075

Huynh, M., Rupp, K., & Sears, J. (2002). *The assessment of Survey of Income and Program Participation (SIPP) benefit data using longitudinal administrative records* (No. 238). U.S. Census Bureau. https://www.census.gov/sipp/workpapr/wp238.pdf

Jäckle, A. (2008). Dependent interviewing: Effects on respondent burden and efficiency of data collection. *Journal of Official Statistics*, *24*(3), 411–430. https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/dependent-interviewing-effects-on-respondent-burden-and-efficiency-of-data-collection.pdf

Jäckle, A. (2009). Dependent interviewing: A framework and application to current research. In P. Lynn (Ed.), *Methodology of longitudinal surveys* (pp. 93–111). John Wiley & Sons. https://doi.org/10.1002/9780470743874.ch6

Jäckle, A., & Eckman, S. (2019). Is that still the same? Has that changed? On the accuracy of measuring change with dependent interviewing. *Journal of Survey Statistics and Methodology,* Article smz021, https://doi.org/10.1093/jssam/smz021

Jäckle, A., Laurie, H., & Uhrig, S. (2007). *The introduction of dependent interviewing on the British Household Panel Survey* (ISER working paper series 2007-07). Institute for Social and Economic Research. https://www.iser.essex.ac.uk/research/publications/working-papers/iser/2007-07

Jäckle, A., & Lynn, P. (2007). Dependent interviewing and seam effects in work history data. *Journal of Official Statistics*, *23*(4), 529–551. https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/the-effects-of-dependent-interviewing-on-responses-to-questions-on-income-sources.pdf

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Jindal, I., Nokleby, M., Pressel, D., & Chen, X. (2019). A nonlinear, noise-aware, quasi-clustering approach to learning deep cnns from noisy labels. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 64–72. http://openaccess.thecvf.com/content_CVPRW_2019/papers/Deep%20Vision%20Workshop/Jindal_A_Nonlinear_Noise-aware_Quasi-clustering_Approach_to_Learning_Deep_CNNs_from_CVPRW_2019_paper.pdf

Jones, G., Johnson, W. O., Hanson, T. E., & Christensen, R. (2010). Identifiability of models for multiple diagnostic testing in the absence of a gold standard. *Biometrics*, *66*(3), 855–863. https://doi.org/10.1111/j.1541-0420.2009.01330.x

Kim, G., & Chambers, R. (2012a). Regression analysis under incomplete linkage. *Computational Statistics & Data Analysis*, *56*(9), 2756–2770. https://doi.org/10.1016/j.csda.2012.02.026

Kim, G., & Chambers, R. (2012b). Regression analysis under probabilistic multi-linkage. *Statistica Neerlandica*, *66*(1), 64–79. https://doi.org/10.1111/j.1467-9574.2011.00509.x

King, T., Cook, S., & Childs, J. H. (2012). Interviewing proxy versus self-reporting respondents to obtain information regarding living situations. *Proceedings for the Joint Statistical Meetings, Survey Research Methods Section*, 5667–5677. http://www.asasrms.org/Proceedings/y2012/Files/400243_500698.pdf

Kriegel, H. P., Kröger, P., Sander, J., & Zimek, A. (2011). Density-based clustering. *WIREs Data Mining and Knowledge Discovery*, *1*(3), 231–240. https://doi.org/10.1002/widm.30

Kuha, J., & Skinner, C. (1997). Categorical data analysis and misclassification. In L. Lyberg, P. P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz & D. Trewin (Eds.), *Survey measurement and process quality* (pp. 633–670). John Wiley & Sons. https://doi.org/10.1002/9781118490013.ch28

Kumar, M., & Patel, N. R. (2007). Clustering data with measurement errors. *Computational Statistics & Data Analysis*, *51*(12), 6084–6101. https://doi.org/10.1016/j.csda.2006.12.012

Lahiri, P., & Larsen, M. D. (2005). Regression analysis with linked data. *Journal of the American Statistical Association*, *100*(469), 222–230. https://doi.org/10.1198/016214504000001277

Leroux, B. G. (1992). Maximum-likelihood estimation for hidden Markov models. *Stochastic Processes and Their Applications*, *40*(1), 127–143. https://doi.org/10.1016/0304-4149(92)90141-C

Liseo, B., & Tancredi, A. (2011). Bayesian estimation of population size via linkage of multivariate normal data sets. *Journal of Official Statistics*, *27*(3), 491–505. https://www.scb.se/contentassets/ff271eeeca694f47ae99b942de61df83/bayesian-estimation-of-population-size-via-linkage-of-multivariate-normal-data-sets.pdf

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). John Wiley & Sons. https://doi.org/10.1002/9781119482260

Liu, K. (1988). Measurement error and its impact on partial correlation and multiple linear regression analyses. *American Journal of Epidemiology*, *127*(4), 864–874. https://doi.org/10.1093/oxfordjournals.aje.a114870

Lugtig, P., & Lensvelt-Mulders, G. J. (2014). Evaluating the effect of dependent interviewing on the quality of measures of change. *Field Methods*, *26*(2), 172–190. https://doi.org/10.1177/1525822X13491860

Lynn, P., Jäckle, A., Jenkins, S. P., & Sala, E. (2006). The effects of dependent interviewing on responses to questions on income sources. *Journal of Official Statistics*, *22*(3), 357–384. https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/the-effects-of-dependent-interviewing-on-responses-to-questions-on-income-sources.pdf

Maimon, O., & Rokach, L. (2005). *Data mining and knowledge discovery handbook*. Springer. https://doi.org/10.1007/978-0-387-09823-4

Malach, E., & Shalev-Shwartz, S. (2017). Decoupling "when to update" from "how to update". *NIPS'17: Proceedings of the 31st International Conference on Neural Information*, 961–971. https://dl.acm.org/doi/10.5555/3294771.3294863

Marshall, A., Altman, D. G., Royston, P., & Holder, R. L. (2010). Comparison of techniques for handling missing covariate data within prognostic modelling studies: A simulation study. *BMC Medical Research Methodology*, *10*(1), 7. https://doi.org/10.1186/1471-2288-10-7

Mathiowetz, N. A., & McGonagle, K. A. (2000). An assessment of the current state of dependent interviewing in household surveys. *Journal of Official Statistics*, *16*(4), 401–418. https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/an-assessment-of-the-current-state-of-dependent-interviewing-in-household-surveys.pdf

McLachlan, G. J., & Peel, D. (2004). *Finite mixture models*. John Wiley & Sons. https://doi.org/10.1002/0471721182

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

McVicar, D., & Anyadike-Danes, M. (2002). Predicting successful and unsuccessful transitions from school to work by using sequence methods. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *165*(2), 317–334. https://doi.org/10.1111/1467-985X.00641

Milligan, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, *45*(3), 325–342. https://doi.org/10.1007/BF02293907

Moore, J., Bates, N., Pascale, J., & Okon, A. (2009). Tackling seam bias through questionnaire design. In P. Lynn (Ed.), *Methodology of longitudinal surveys* (*pp.* 73–92). John Wiley & Sons. https://doi.org/10.1002/9780470743874.ch5

Oberski, D. L., Kirchner, A., Eckman, S., & Kreuter, F. (2017). Evaluating the quality of survey and administrative data with generalized multitrait-multimethod models. *Journal of the American Statistical Association*, *112*(520), 1477–1489. https://doi.org/10.1080/01621459.2017.1302338

Oberski, D. L. (2015). Estimating error rates in an administrative register and survey questions using a latent class model. In P. P. Biemer, B. West, S. Eckman, B. Edwards, & C. Tucker (Eds.), *Total Survey Error* (*pp.* 339–358). John Wiley & Sons. https://doi.org/10.1002/9781119041702.ch16

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*(2), 171–182. https://doi.org/10.1007/s11634-015-0211-0

Oberski, D. L., Hagenaars, J. A., & Saris, W. E. (2015). The latent class multitrait-multimethod model. *Psychological Methods*, *20*(4), 422–443. https://doi.org/10.1037/a0039783

O'Neill, D., & Sweetman, O. (2013). The consequences of measurement error when estimating the impact of obesity on income. *IZA Journal of Labour Economics*, *2*(1), 1–20. https://doi.org/10.1186/2193-8997-2-3

Onyancha, J., Plekhanova, V., & Nelson, D. (2017). Noise web data learning from a web user profile: Position paper. Proceedings of the World Congress on Engineering 2017, 2, 608–611. http://www.iaeng.org/publication/WCE2017/WCE2017_pp608-611.pdf

Pankowska, P., Bakker, B. F. M., Oberski, D. L., & Pavlopoulos, D. (2020). How linkage error affects hidden Markov model estimates: A sensitivity analysis. *Journal of Survey Statistics and Methodology,* 8(3), 483–512. https://doi.org/10.1093/jssam/smz011

Pankowska, P., Bakker, B. F. M., Oberski, D. L., & 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*(3), 317–329. https://doi.org/10.3233/SJI-170368

Pavlopoulos, D., Muffels, R., & Vermunt, J. K. (2012). How real is mobility between low pay, high pay and non-employment? *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *175*(3), 749–773. https://doi.org/10.1111/j.1467-985X.2011.01017.x

Pavlopoulos, D., & Vermunt, J. K. (2015). Measuring temporary employment. Do survey or register data tell the truth? *Survey Methodology*, *41*, 1. https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2015001/article/14151-eng.pdf?st=H1TgHaQP

Piccarreta, R., & Billari, F. C. (2007). Clustering work and family trajectories by using a divisive algorithm. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *170*(4), 1061–1078. https://doi.org/10.1111/j.1467-985X.2007.00495.x

Pickles, A., Bolton, P., Macdonald, H., Bailey, A., Le Couteur, A., Sim, C. H., & Rutter, M. (1995). Latent-class analysis of recurrence risks for complex phenotypes with selection and measurement error: A twin and family history study of autism. *American Journal of Human Genetics*, *57*(3), 717. https://www.ncbi.nlm.nih.gov/pubmed/7668301

Qu, Y., & Hagdu, A. (2012). Modelling correlations between diagnostic tests in efficacy studies with an imperfect reference test. In T. G. Gregoire, D. R. Brillinger, P. J. Diggle, E. Russek-Cohen, W. G. Warren & R. D. Wolfinger (Eds.), *Modelling Longitudinal and Spatially Correlated Data* (*pp.* 363–371). Springer.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66*(336), 846–850. https://doi.org/10.2307/2284239

Reynolds, D. (2009). Gaussian mixture models. In S. Z. Li & A. Jain (Eds.), *Encyclopedia of Biometrics* (pp. 827–832). *Springer*. https://doi.org/10.1007/978-0-387-73003-5_196

Romensburg, H. C. (2004). *Cluster analysis for researchers*. Lulu Press.

Sadinle, M. (2014). Detecting duplicates in a homicide registry using a Bayesian partitioning approach. *The Annals of Applied Statistics*, *8*(4), 2404–2434. https://doi.org/10.1214/14-AOAS779

Sadinle, M. (2017). Bayesian estimation of bipartite matchings for record linkage. *Journal of the American Statistical Association*, *112*(518), 600–612. https://doi.org/10.1080/01621459.2016.1148612

Sadinle, M., Hall, R., & Fienberg, S. E. (2011). Approaches to multiple record linkage. *Int. Statistical Inst.: Proc. 58th World Statistical Congress* (Session IPS059), 1064–1071. http://www.2011.isiproceedings.org/papers/450092.pdf

Sala, E., Uhrig, S. N., & Lynn, P. (2011). "It is time computers do clever things!": The impact of dependent interviewing on interviewer burden. *Field Methods*, *23*(1), 3–23. https://doi.org/10.1177/1525822X10384087

Sammut, C., & Webb, G. I. (2011). *Encyclopedia of machine learning* (2010th ed.). Springer.

Saris, W. E., & Gallhofer, I. (2007). Estimation of the effects of measurement characteristics on the quality of survey questions. *Survey Research Methods*, *1(1)*, 29–43. https://doi.org/10.18148/srm/2007.v1i1.49

Saris, W. E., & Gallhofer, I. N. (2014). *Design, evaluation, and analysis of questionnaires for survey research*. John Wiley & Sons. https://doi.org/10.1002/9780470165195

Särndal, C. E., Swensson, B., & Wretman, J. (2003). *Model assisted survey sampling*. Springer.

Scholtus, S. (2018). *Editing and estimation of measurement errors in administrative and survey data* [Doctoral dissertation, Vrije Universiteit, Amsterdam]. Faculteit der Sociale Wetenschappen, Vrije Universiteit Amsterdam. https://research.vu.nl/ws/portalfiles/portal/56329828/complete+dissertation.pdf

Scholtus, S., Bakker, B. F. M., & Van Delden, A. (2015). *Modelling measurement error to estimate bias in administrative and survey variables* (Discussion Paper No. 17). Statistics Netherlands. https://www.cbs.nl/-/media/imported/documents/2015/46/modelling_measurement_error.pdf

Steorts, R. C. (2015). Entity resolution with empirically motivated priors. *Bayesian Analysis*, *10*(4), 849–875. https://projecteuclid.org/euclid.ba/1441790411

Steorts, R. C., Hall, R., & Fienberg, S. E. (2016). A Bayesian approach to graphical record linkage and deduplication. *Journal of the American Statistical Association*, *111*(516), 1660–1672. https://doi.org/10.1080/01621459.2015.1105807

Sudman, S., Bradburn, N., & Schwarz, N. (1997). *Thinking about answers: The application of cognitive processes to survey methodology.* John Wiley & Sons.

Tan, P-N., Steinbach, M., Karpatne, A., & Kumar, V. (2019). *Introduction to data mining* (2nd ed.). Pearson.

Torrance-Rynard, V. L., & Walter, S. D. (1997). Effects of dependent errors in the assessment of diagnostic test performance. *Statistics in Medicine*, *16*(19), 2157–2175. https://doi.org/10.1002/(SICI)1097-0258(19971015)16:19<2157::AID-SIM653>3.0.CO;2-X

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press. https://doi.org/10.1017/CBO9780511819322

Uhrig, S. N., & Sala, E. (2011). When change matters: An analysis of survey interaction in dependent interviewing on the British Household Panel Study. *Sociological Methods & Research*, *40*(2), 333–366. https://doi.org/10.1177/0049124111404816

Vacek, P. M. (1985). The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics 41*(4), 959–968. https://doi.org/10.2307/2530967

Van de Pol, F., & De Leeuw, J. (1986). A latent Markov model to correct for measurement error. *Sociological Methods & Research*, *15*(1–2), 118–141. https://doi.org/10.1177/0049124186015001009

Van de Pol, F., & Langeheine, R. (1990). Mixed Markov latent class models. *Sociological Methodology*, *20*, 213–247. https://doi.org/10.2307/271087

Van Delden, A., Pannekoek, J., Banning, R., & Boer, A. de. (2016). Analyzing correspondence between administrative and survey data. *Statistical Journal of the IAOS*, *32*(4), 569–584. https://doi.org/10.3233/SJI-160972

Van Rooijen, J., Bloemendal, C., & Krol, N. (2016). The added value of micro-integration: Data on laid-off employees. *Statistical Journal of the IAOS*, *32*(4), 685–692. https://doi.org/10.3233/SJI-161013

Van Smeden, M., Oberski, D. L., Reitsma, J. B., Vermunt, J. K., Moons, K. G., & De Groot, J. A. (2016). Problems in detecting misfit of latent class models in diagnostic research without a gold standard were shown. *Journal of Clinical Epidemiology*, *74*, 158–166. https://doi.org/10.1016/j.jclinepi.2015.11.012

Vermunt, J. K. (2002). A general latent class approach to unobserved heterogeneity in the analysis of event history data. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class analysis* (pp. 383–407). Cambridge University Press. https://doi.org/10.1017/CBO9780511499531.015

Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class analysis* (pp. 89–106). Cambridge University Press. https://doi.org/10.1017/CBO9780511499531.004

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

West, B. T., & Blom, A. G. (2016). Explaining interviewer effects: A research synthesis. *Journal of Survey Statistics and Methodology*, *5*(2), 175–211. https://doi.org/10.1093/jssam/smw024

Winglee, M., Valliant, R., & Scheuren, F. (2005). A case study in record linkage. *Survey Methodology, 31*(1), 3–11. https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2005001/article/8085-eng.pdf?st=sVcJ1LbU

Winkler, W. E. (1999). *The state of record linkage and current research problems*. US Census Bureau. https://www.census.gov/srd/papers/pdf/rr99-04.pdf

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., & Ruzzo, W. L. (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics*, *17*(10), 977–987. https://doi.org/10.1093/bioinformatics/17.10.977

Zhang, L. C. (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, *66*(1), 41–63. https://doi.org/10.1111/j.1467-9574.2011.00508.x

Summary

Despite countless attempts to reduce it and address its causes, measurement error is a problem inherent to all data sources. (Alwin, 2007; Biemer et al., 1991; Kuha & Skinner, 1997). Its presence often leads to biased and inconsistent statistical estimates and, as a consequence, to erroneous findings and conclusions. It is therefore crucial to understand, account, and correct for measurement error to ensure research validity (Fuller, 2009; Grace, 2017; Kuha & Skinner, 1997).

Latent variable modelling is an increasingly popular solution to this problem, as it allows for the estimation of and correction for measurement error without the need for gold standard data. That is, the main advantage of latent variable models (LVMs) is the fact that, unlike alternative measurement-error-correction techniques, they do not make use of error-free validation data, which are rarely available in practice. Instead these models use repeated indicators of the same variable, either cross-sectionally from various sources or over time from the same source, to extract information about measurement error directly from the data (Biemer & Bushery, 2000).

A group of LVMs that are applied specifically to categorical, longitudinal data, and which are the main focus of this dissertation, are hidden Markov models (HMMs) (Biemer, 2004, 2011; Oberski et al., 2017; Pavlopoulos & Vermunt, 2015). HMMs are used when a (dynamic) quantity of interest is measured in a panel survey with some degree of error. The models allow for the separation of true change from measurement error which, in turn, can produce error-corrected estimates of the quantity of interest, and are also able to assess the level of measurement error in the corresponding variable (Biemer, 2011; Pankowska et al., 2018).

The standard HMM consists of two components: (i) the structural component that models the true (latent) initial state probabilities and the true (latent) transition probabilities; and (ii) the measurement component that models the interactions of the observed values (which contain error) with the true values at each time point. The two components are estimated simultaneously. The model relies on two basic assumptions: first, the probability of a specific value occurring at time t only depends on its value in the previous time point – the so-called *Markov assumption*. Second, the probability of observing a specific value at time t only depends on the true value at the same time point – the so-called *local independence assumption* or the *independent classification error (ICE) assumption*. While the standard, single-indicator HMM relies on the local independence assumption for identifiability, this assumption is often viewed as highly restrictive and unrealistic, as it does not allow for the modelling of the presence of systematic errors without risking poor model identifiability. To overcome this challenge, it is possible to use extended, multiple-indicator versions of HMMs. However, this solution introduces some new challenges. Most importantly, the use of multiple indicators usually requires performing record linkage, which might lead to linkage error – a new potential source of bias. Furthermore, the implementation of such extended models also tends to be complex and time-consuming.

Given the potentially strong, adverse effects of measurement error and the possibility of minimizing these using HMMs, the aim of this thesis is twofold: first to understand in more detail the problem of measurement error and second to investigate whether extended HMMs that are applied to linked data can be used for error, and to what extent this method can be feasibly implemented.

In more detail, **Chapter 2** examines the bias introduced by measurement error, using clustering as an illustrative example. More specifically, the simulation study investigates the sensitivity of two commonly used model- and density-based clustering algorithms (i.e. GMMs and DBSCAN) to varying severities and magnitudes of random and systematic errors. The results confirm that measurement error in many cases leads to non-negligible bias, as the returned clusters are (highly) dissimilar to the ones obtained when the dataset is error-free. The number of clusters found in the data is also affected by the error.

Chapter 3 looks at how different data collection processes might impact the nature and magnitude of measurement error, by studying how the switch from dependent interviewing (DI) to independent interviewing (INDI) in the Dutch Labour Force Survey (LFS) affects the random and systematic components of the error. The results indicate that the use of DI lowers the probability of obtaining random errors but has no significant effect on systematic errors. What is more, the results also show that regardless of the interviewing regime used, the survey data, similarly to the register data, also contains autocorrelated error. The findings of this paper indicate that both data sources examined are subject to non-negligible systematic error that needs to be considered when correcting for measurement error using HMMs. This in turn confirms the need for using extended, multiple-indicator HMMs, which allow for the relaxation of the local independence assumption and the modelling of error autocorrelation without risking poor model identifiability.

Chapter 4 investigates whether and to what extent the use of multiple-indicator HMMs, which often requires record linkage, leads to biased estimates due to the presence of linkage error. The results of the simulation study show that overall the sensitivity of the HMM (structural) parameter estimates to false-positive and false-negative linkage error is low. It appears that only rather extreme scenarios (i.e. high error rate and high correlation between the probability of error and model estimates) lead to substantial bias. Moreover, the results also show that under certain conditions, false-positive linkage error acts as another source of measurement error that is absorbed into the error-rate parameters of the model, leaving the latent transition estimates unaffected. In these cases, HMMs also accounts for linkage error.

Finally, **Chapter 5** focuses on a more practical matter. Given their complex nature and, with that, the time and costs associated with their implementation, the study explores the feasibility of using multiple-indicator HMMs in the first instance. More specifically, the study investigates whether it is possible to simplify the error correction procedure by running the full analysis once and then re-using the resultant error parameters as a correction factor for a number of years. The proposed solution is

contingent on the assumption that the structure and size of the error remain constant. The analysis provides some evidence that in the absence of a major change in the data collection process, the size and structure of the error are time-invariant and, therefore, the error parameters can be carried forward a certain number of years.

While the findings presented in this dissertation suggest that HMMs are a promising tool to correct for measurement error in categorical, longitudinal data, several additional aspects need to be considered before this approach can be applied in practice. Namely, the performance and feasibility of the method should be tested in a different context that goes beyond the topic of labor mobility and on data from different countries than the Netherlands. Also, if possible, additional sources apart from surveys and administrative registers should be considered. Furthermore, a thorough examination of model robustness and the sensitivity of parameter estimates to varying model specifications containing different assumptions ought to be carried out. Finally, it is also important to consider how researchers can use error-corrected microdata in their analyses, while accounting for the uncertainty of the "true state" membership.

Acknowledgments

First, I would like to thank the reading committee – Paul Biemer, Ton de Waal, Harry Ganzeboom, Wendy Smits, and Jeroen Vermunt – for taking the time to read my dissertation and providing valuable suggestions. Furthermore, I would also like to thank Paul Biemer and RTI International for giving me the opportunity to spend the summer of 2018 as an intern with the RTI Survey Statistics group in North Carolina. My time at RTI was a valuable and memorable experience both professionally and personally. I would also like to thank Jeroen Vermunt for his frequent and timely responses to all of my Latent Gold related queries.

Second, I would like to express my gratitude to my supervisors – Bart Bakker, Daniel Oberski, and Dimitris Pavlopoulos – for their continuous support and encouragement throughout my PhD and their ongoing assistance as I continue developing as a researcher. Bart, you introduced me to the world of Official Statistics and helped me to establish myself in the community by providing me with numerous opportunities to go relevant international conferences as well as present my research to CBS colleagues. Daniel, working with you taught me so much; you helped me to deepen my knowledge and become a much stronger methodologist. I highly appreciate the time you took to sit down with me and explain the inner workings of the models I was using. I value your ability to break down complex concepts into more manageable and understandable parts. Dimitris, you looked out for me from day one and I feel that your support went well beyond what is expected of a supervisor – I can't tell you how grateful I am for that.

I am also thankful to CBS colleagues – Laura Boeschoten, Peter-Paul de Wolf, John Michiels, Jeroen Pannekoek, Sander Scholtus, Barry Schouten, and Wendy Smits to name a few – who provided valuable input and feedback on my research and who were always willing to help and answer any questions regarding the data, methods, and beyond. I am equally grateful to all members of the SILC group and in particular to Harry Ganzeboom, Aat Liefbroer, and Ineke Maas, who were always happy to read and comment on my papers, which allowed me to substantially improve them and helped in getting them published.

To my fellow PhD colleagues who over these last five years have become close friends I want to say a huge thank you. You were an integral part of my experience! Maarten, from the moment we met at the beginning of my PhD you were incredibly helpful and made my whole experience better. You were always happy to help with Dutch translations and anything else that came up. You listened to my rants and you always made me feel better. I already miss our lunches and the hangouts by the E-wing coffee machine. Jolien, we met on the very first day of our PhDs and we immediately hit it off. I am still amazed at how similar our experiences were – for better and sometimes for worse. I feel that I can always talk to you about everything and that we truly understand each other. Sebastien, you always managed to make me laugh and put a big smile on my face. Especially on days when things were not going well you helped

me pick myself out of the slump. Silvia, you made my trips to the A-wing worthwhile! I always appreciated your dedication and hard work, both of which motivated me. It goes without saying that we also had our fair share of fun, like the memorable NYE and the time we committed a fashion faux pas at the department outing. Last but not least, Lucille you have been a great and very helpful office mate. I really enjoyed our vivid discussions of the Dutch economy, politics, and, most importantly, cats.

I also want to express my gratitude to my dear friends, who supported me throughout these years and who were always there for me. Alina, Anna, Anne, Jelena, Laury, Lisa, and Ola – you listened when I needed to complain and you were always happy to celebrate my achievements with me – thank you! Ola, I also want you to know how grateful I am for all of your help with the graphs and of course the beautiful cover page. Your contribution has brought a great deal of color into this rather statistical book. Yvonne, ik wil jou ook bedanken voor je steun en omdat je altijd in me geloofde en mij het beste toewenste.

Rich, there are so many things I want to thank you for I don't even know where to begin. You have helped me and supported me so much throughout my PhD both professionally and mentally. I honestly don't think I could have done this without you. I am so grateful for everything you have done for me!

Mamo, Tato, jestem Wam bardzo wdzięczna za wsparcie i wiarę we mnie, nie tylko podczas doktoratu, ale na przestrzeni całego mojego życia. Dało mi to odwagę i pewność siebie oraz pomogło ukończyć studia doktoranckie z powodzeniem. Dziękuję za to, że zawsze byliście gotowi mnie wysłuchać i okazać pomoc, gdy tylko o nią poprosiłam.

Daniel, tak bardzo mi przykro, że Ciebie tu już z nami nie ma.