# Handling missing data in multichannel life course analysis

Emery Kevin

UNIL | Université de Lausanne

FACULTÉ DES SCIENCES SOCIALES ET POLITIQUES

INSTITUT DES SCIENCES SOCIALES

# Handling missing data in multichannel life course analysis

THÈSE DE DOCTORAT

présentée à la

Faculté des sciences sociales et politiques
de l'Université de Lausanne

pour l'obtention du grade de

Docteur en mathématique appliquée aux sciences humaines et sociales

par

Kevin Emery

Directeur de Thèse
Prof. André Berchtold

Jury

Dr. Caroline Roberts
Prof. Matthias Studer
Dr. Brendan Halpin

LAUSANNE
2023

UNIL | Université de Lausanne

Faculté des sciences sociales et politiques

**IMPRIMATUR**

Le Décanat de la Faculté des sciences sociales et politiques de l'Université de Lausanne, au nom du Conseil et sur proposition d'un jury formé des professeurs

- M. André BERCHTOLD, Professeur, Directeur de thèse
- Mme Caroline ROBERTS, Professeure Assistante à l'Université de Lausanne
- M. Matthias STUDER, Professeur à l'Institut de Démographie et Socioéconomie (IDESO) de l'Université de Genève
- M. Brendan HALPIN, Maître de Conférences au Département de Sociologie de l'Université de Limerick, Irlande

autorise, sans se prononcer sur les opinions du candidat, l'impression de la thèse de Monsieur Kevin EMERY, intitulée :

**« Handling missing data in multichannel life course analysis. »**

Nicky LE FEUVRE
Doyenne

Lausanne, le 29 juin 2023

# Abstract

This thesis addresses the challenge of dealing with missing data, which is an inevitable issue in quantitative studies. The appropriate treatment of missing data is complex and can significantly impact statistical results and conclusions. A particular emphasis is placed on data for life course analysis. Life course data have characteristics that require specific treatment of missing data. First, the longitudinal structure is crucial since life courses are often considered as a whole. Then, due to their longitudinal nature, there are generally several missing observations in a row. This behaviour happens in a survey when individuals miss several waves of data collection or drop out of the survey. Next, they are generally coded as categorical data. Therefore, standard methods, that were generally developed for numerical variables, are difficult to apply. Finally, such data are subject to few transitions.

One of the major challenges encountered in life course methodology is the absence of a commonly accepted solution for handling missing data (Piccarreta and Studer, 2019; Liao et al., 2022). The main aim of this thesis is to fill this gap. Furthermore, this thesis proposes novel methods to enhance existing approaches and address gaps in the imputation of longitudinal categorical data.



# Résumé

Cette thèse aborde le défi de la gestion des données manquantes, qui constitue un problème inévitable dans les études quantitatives. Le traitement approprié des données manquantes est complexe et peut avoir un impact significatif sur les résultats statistiques et les conclusions. Un accent particulier est mis sur les données de parcours de vie. Les données de parcours de vie présentent des caractéristiques qui nécessitent un traitement spécifique des données manquantes. Premièrement, la structure longitudinale est cruciale, car les parcours de vie sont souvent considérés dans leur intégralité. De plus, en raison de leur nature longitudinale, il y a généralement plusieurs observations manquantes consécutives. Ce phénomène se produit lorsqu'il y a des vagues de collecte de données manquées par les individus ou lorsqu'ils abandonnent l'enquête. Ensuite, ces données sont généralement codées comme des données catégorielles. Par conséquent, les méthodes standard, généralement développées pour les variables numériques, sont difficiles à appliquer. Enfin, de telles données sont caractérisées par peu de transitions.

L'un des principaux défis de la méthodologie du parcours de vie réside dans l'absence d'une solution communément acceptée pour traiter les données manquantes (Piccarreta et Studer, 2019; Liao et al., 2022). L'objectif principal de cette thèse est de combler cette lacune. De plus, cette thèse propose des méthodes novatrices pour améliorer les approches existantes et combler les lacunes dans l'imputation des données catégorielles longitudinales.

# Acknowledgments

First, I would like to express my heartfelt gratitude to my supervisor, Professor André Berchtold. Without his unwavering guidance and support, this endeavour would not have been possible. Professor Berchtold's constant encouragement, his ability to gently steer me back on track when I veered off course, and, above all, his remarkable personal qualities make him an extraordinary supervisor.

I am also deeply appreciative of my esteemed jury members, Doctor Brendan Halpin, Professor Matthias Studer, and Doctor Caroline Roberts. Their expertise and invaluable insights on my thesis have significantly contributed to its improvement and have laid the foundation for my future research.

I extend my sincere thanks to my friends and colleagues at the University of Geneva and Lausanne. The valuable feedbacks from Rojin, Léonard, and Leonhard during the research seminars on sequence analysis have been instrumental in shaping my work. The enriching exchanges with my office mates, Zhivko, Serguei, Ibrahima, Anthony, and Julie, have been invaluable over the years. The unwavering support and companionship of Marc, Nathalie, Lucas, Géraldine, Robert, Fantine, Amelia, Dimitrios, Samuel, Iván, Jad, and Michael have been a constant source of strength throughout this journey. The quality time we have spent together, both within the university and outside, holds a special place in my heart.

I would like to express my sincere gratitude to the Swiss National Science Foundation for their generous funding of this thesis and to the NCCR-LIVES and its administrative team for their unwavering support. The doctoral programme they organised has provided me with invaluable knowledge in life course research and introduced me to exceptional individuals. Additionally, the annual Doctoriales and the valuable feedbacks from experts have been a fantastic learning experience.

I would like to express my gratitude to my friends, who have helped me maintain a necessary balance throughout this journey. Life is easy with such amazing people. I want to extend a special note of appreciation to my friends Johann and Sunny, who, on several occasions, served as test subjects for my presentations.

I am also thankful to my father and brother for their unwavering love and support. Finally, I would like to express my deepest love for my mother, whom I miss greatly.

# Contents


**1 Introduction** **1**

**2 State of the art** **5**

2.1 Definition . . . 5
2.2 Data collection and errors . . . 6
2.3 Patterns of missing data . . . 7
2.4 Mechanisms . . . 8
2.5 Issues related to missing data . . . 10
2.6 Preventing missing data . . . 10
2.7 Methods for dealing with missing data . . . 12
2.7.1 Deletion methods . . . 13
2.7.2 Weighting methods . . . 15
2.7.3 Likelihood and Bayesian methods . . . 16
2.7.4 Imputation . . . 19
2.8 Sequence analysis . . . 29

**3 Multiple imputation in longitudinal datasets** **36**

3.1 Introduction . . . 36
3.2 Methods . . . 38
3.2.1 Analysis of missing data . . . 38
3.2.2 Treatment of missing data . . . 39
3.3 Data . . . 44
3.4 Results . . . 46
3.4.1 Analysis of missing data . . . 46
3.4.2 Treatment of missing data . . . 50
3.5 Discussion . . . 60

**4 Comparison of imputation methods in the case of life course data** **65**

4.1 Introduction . . . 65
4.2 Imputation methods . . . 68

4.3 Simulation Framework . . . 74
4.4 Results . . . 83
4.5 Discussion . . . 87

**5 Comparison of two approaches in multichannel sequence analysis using the Swiss household panel 95**
5.1 Introduction . . . 95
5.2 Data . . . 97
5.3 Methods . . . 98
5.4 Results . . . 102
5.5 Discussion . . . 108

**6 Multichannel imputation 120**
6.1 Introduction . . . 120
6.2 Algorithms . . . 122
6.3 Simulation frameworks . . . 128
6.3.1 First simulation framework . . . 128
6.3.2 Second simulation framework . . . 135
6.4 Results . . . 137
6.4.1 First simulation framework . . . 138
6.4.2 Second simulation framework . . . 156
6.5 Discussion . . . 164

**7 Conclusion 168**
7.1 Multiple imputation in longitudinal datasets . . . 168
7.2 Imputation of life course data . . . 169
7.3 Comparison of MSA and EA in a real case . . . 170
7.4 Multichannel imputations . . . 171
7.5 Developments of seqimpute . . . 172
7.6 Further developments . . . 172

**Bibliography 175**

**A Technical elements 195**

**B Statistical models 198**
B.1 Multinomial model . . . 198
B.2 Random forest . . . 199
B.3 Variable-length Markov chains . . . 200

**C Men results** **201**

**D Detail of the variables** **206**

**E** **208**

E.1 States with an higher probability . . . . . 208

E.2 Details on the data . . . . . 209

E.3 Formulae for the criteria . . . . . 212

E.4 Normalisation of the criteria . . . . . 212

E.5 Total score by algorithm . . . . . 213

**F Additional results for the multichannel comparison** **218**

# Chapter 1

# Introduction

This thesis addresses the challenge of dealing with missing data, which is an inevitable issue in quantitative studies. The appropriate treatment of missing data is complex and can significantly impact statistical results and conclusions. For instance, a study conducted by Lall (2016) examined the impact of two commonly used methods for managing missing data on various statistical analyses. The results showed that in half of the cases, the choice of method significantly affected the essential statistical outcomes. Although Lall's research focused on the field of comparative and international political economy, these findings could extend to other fields as well. Despite the critical role of managing missing data, a review by Berchtold (2019) on articles published in top-tier social science journals found that over half of the cases did not acknowledge the presence of missing data, even when it was present. Moreover, when missing data were acknowledged, inadequate treatments were applied in most cases.

In this thesis, a particular emphasis is placed on data for life course analysis. As stated Bernardi et al. (2019):"The life course, therefore, can be defined as a multifaceted process of individual behaviour; that is, it evolves from the steady flow of individuals' actions and experiences, which modify their biographical states". Within the life course paradigm, life trajectories are studied as a whole rather than as a collection of events. One of the major challenges encountered in life course methodology is the absence of a commonly accepted solution for handling missing data (Piccarreta and Studer, 2019; Liao et al., 2022). The main aim of this thesis is to fill this gap. Furthermore, this thesis proposes novel methods to enhance existing approaches and address gaps in the imputation of longitudinal categorical data.

Life course data have characteristics that require specific treatment of missing data. First, the longitudinal structure is crucial since life courses are often considered as a whole. For example, events early in the trajectory may have an impact later on. Then, due to their longitudinal nature, there are generally several missing observations in a row, called gap of missing data. This behaviour happens in a survey when individuals miss several waves of data collection or drop out of the survey. Next, they are generally coded as categorical data.

Therefore, standard methods, that were generally developed for numerical variables, are difficult to apply. Finally, such data are subject to few transitions. For example, individuals do not change their civil status or with whom they live yearly.

We end this introduction by detailing the subject of the six remaining chapters, showing how they fit into the global objective of treating missing data in life courses and discussing their main contributions.

Chapter 2 provides a comprehensive overview of the state-of-the-art in dealing with missing data. The chapter introduces the concept of missing data and its impact on statistical analysis, discusses the different patterns and mechanisms of missingness, and reviews various methods for handling missing data in a longitudinal framework.

Furthermore, the chapter also presents ways to study life courses, since missing data can have different impacts on the statistical analysis, and the treatment of missing data may need to be tailored to the specific analysis. Overall, this chapter serves as a foundation for understanding the importance of addressing missing data and highlights the need for novel approaches to handle this issue in the context of life course studies.

Chapter 3 tackles the issue of missing data in longitudinal surveys, which shares similarities with missing data in life course studies due to the longitudinal nature and treatment of categorical data. Moreover, it allows us to have a first grasp on the treatment of missing data in a longitudinal setting.

This chapter introduces a multiple imputation procedure that takes into account the characteristics of longitudinal datasets. Multiple imputation is a popular method of treatment of missing data, whose idea is to replace missing data by several sensible values. We especially focus on the challenges induced by categorical data and logical missing data, which occur when questions are not applicable to certain individuals. For example, questions about voting habits are not asked to individuals that do not have the right to vote. Standard multiple imputation procedures such as *fully conditional specification* (Van Buuren, 2007) may create inconsistencies in the imputed values and distort the association between the variables. Therefore, we discuss a multiple imputation procedure that takes into account such issues. Moreover, we introduce a sequence of questions that allow determining which values are truly missing. Therefore, this chapter directly addresses the challenges related to categorical data and logical missing values that have not been fully addressed in previous studies that discussed the application of multilevel imputation models for panel data (Spiess et al., 2021), mostly focused on the impact of multiple imputation on the subsequent statistical analysis (Young and Johnson, 2015), or discussed logical missing but without considering its interaction with the challenges inherent to categorical data (Aßmann et al., 2017).

Chapter 4 focuses on the issue of missing data in univariate longitudinal data, which occurs when only one life domain is considered, such as professional, family, or health trajectories.

The aim of this chapter is twofold: to provide guidance on how to handle missing data and to introduce extensions to the *MICT* algorithm (Halpin, 2012, 2013, 2016b), a method designed for imputing categorical longitudinal data.

The typical methods used to treat missing data in longitudinal categorical data are compared with a framework built explicitly for life course data. Several comparisons of methods to handle missing data in a longitudinal context were done in the literature (e.g. De Silva et al., 2017; Kalaycioglu et al., 2016). However, they are often restricted to numerical variables. Anyway, at our knowledge, the longitudinal observations were never regarded as a whole, as it is the case with life course data, and the *MICT* algorithm was never included in these comparisons. Therefore, this chapter aims to fill this gap by reviewing the methods available for dealing with missing data in life courses, comparing them and by giving clear guidelines on their application.

Then, we introduce two extensions to the *MICT* algorithm. First, we explore the use of random forests models, rather than standard multinomial regression models, as imputation models. Random forest models have appealing properties for the imputation of life course data, such as a good handling of many predictors or non-linear effects. Secondly, we develop a method to handle missing data in trajectories that are not time-homogeneous. Such behaviour is common in life course trajectories, where transitions between ages 15-16 in terms of profession or family status may differ from those between ages 39-40. The standard *MICT* algorithm may not be optimal in such situations.

The following two chapters shift focus to multichannel trajectories. In the study of the life course, an individual's life domains often impact each other and are impacted by those of others (Bernardi et al., 2019). As a result, it is common to examine multiple associated trajectories simultaneously. For instance, in the case of women, family and professional trajectories are often intertwined (e.g. Piccarreta and Billari, 2007; Widmer and Ritschard, 2009; Aisenbrey and Fasang, 2017). As such, to gain a comprehensive understanding of women's life courses, these two domains cannot typically be examined in isolation.

In Chapter 5, we examine two methods for combining information from different trajectories, namely multichannel sequence analysis and extended alphabet. While this chapter does not directly address the issue of missing data, it is still relevant to the handling of missing data in life course data. Indeed, the usefulness of a method for handling missing data is often evaluated by its impact on statistical analyses. Clustering is the most commonly used statistical analysis with life course data. Therefore, it is necessary to have a clear understanding of the tools available for building clustering of multichannel trajectories before introducing the complexities related to missing data.

This chapter aims to fill two gaps. Firstly, there has been no comparison of these two methods on real data. Previous studies, such as Gauthier et al. (2010), have compared these methods using simulation frameworks, and informal guidelines can be found in the literature

(e.g. Piccarreta, 2017). However, an understanding of how these methods behave on real data and a screening of the situations where one or the other method is most appropriate is lacking. Then, we develop a framework based on the latest methodological advances to compare and validate clusterings of multichannel trajectories. The validation of these clusterings differs from the standard clustering validation methods due to a range of essential characteristics that need to be considered.

The *MICT* multiple imputation algorithm appears as a promising solution to handle missing data in life course, but can only impute one trajectory at a time in its standard formulation. Therefore, the main objective of chapter 6 is to develop an extension of the *MICT* algorithm to the imputation of several trajectories. This contribution is one of the main innovations of this thesis. This algorithm's parameters and global performance are discussed in different contexts and compared with existing solutions. In particular, we make use of the findings of chapter 5 to compare the impact of different multiple imputation methods on clustering results.

A conclusion lists the main contribution of this thesis, summarises each chapter and provides ideas for further developments.

# Chapter 2

# State of the art

This chapter provides an overview of the current state of research on missing data and its treatment, with a focus on life course data. We introduce the different concepts related to missing data that are used throughout this thesis. The goal is to lay a foundation for this thesis and identify methodological gaps that we aim to address.

We start by defining missing data and explore the various types of errors that can arise during data collection. Specifically, we highlight the errors that are associated with missing data. Then, we explore the various patterns of missing data and the three mechanisms underlying missing data. Next, we discuss the challenges that missing data pose and the different methods available to handle them, including strategies for preventing missing data during data collection. Finally, we explore the life course paradigm, focusing specifically on sequence analysis.

To enhance readability, statistical details related to the different concepts are included in Appendix A.

## 2.1 Definition

The first question when discussing missing data is: what is missing data? The broadest definition is simply *all information that is lacking*. To go further, Rubin (1976) differentiates between missing data that was anticipated beforehand and missing data that was unintended. An example of expected missing data, which we call logical missing in this thesis, is that of an experimental design, where individuals that did not receive the treatment have no data on outcomes related to the treatment. In the context of surveys, individuals that are not working are not expected to answer questions related to work. In this thesis, we consider as missing

*the information that was planned to be collected but that was not obtained.*

Therefore, the values that were expected to be lacking beforehand are not considered as really missing.

Even a simple definition of missing data, as we have chosen, does not remove all ambiguity. Indeed, it is not always apparent whether a value is missing. For example, we could imagine a survey where an individual is asked to fill in her/his second nationality (without the choice to say s/he has no second nationality). A lack of response to this question encompasses two situations: either the individual does not have any second nationality, or s/he has one but does not want to provide it. Another example appears in panel surveys when a question is dependent on another. For example, if an individual does not answer the question about having a child, there will be no information on further questions about the number of children or their age; hence, we cannot determine with certainty if these pieces of information are missing or not.

## 2.2 Data collection and errors

To understand why it is important to handle missing data in longitudinal studies, we first need to consider the various errors that can occur during data collection and the errors that we aim to address in this thesis. Even though this thesis does not focus directly on survey methodology, it is still useful to discuss how data collection errors can affect the accuracy and quality of longitudinal data, and how this impacts our ability to deal with missing data.

In this section, we draw on the concept of "total survey error" developed by Groves (2005) to explore the different sources of error that can arise during data collection. The total survey error paradigm considers four main types of error: coverage, sampling, nonresponse, and measurement errors (Lyberg and Weisberg, 2016).

Coverage errors stem from the discrepancy between the population from which the sample is drawn and the target population. In longitudinal surveys, these errors can manifest themselves when individuals relocate in or out of the country or when a part of the population under study cannot be contacted, leading to a mismatch between the population used to draw the sample and the actual population under study.

Sampling error arises because the sample is not a perfect representation of the population. Probability sampling methods, such as simple random sampling, stratified random sampling, and cluster sampling, are typically used to create a sample that has similar characteristics to the population (see e.g. Tillé and Wilhelm (2017) for details on probability sampling). However, probability sampling can be expensive and time-consuming. Non-probability sampling methods are sometimes used as a less expensive and less time-consuming alternative, but they often result in larger sampling errors (see e.g. Vehovar et al. (2016) for a discussion on non-probability sampling).

Non-response error occurs in longitudinal surveys when selected individuals either refuse to respond, provide incomplete responses, or stop responding altogether in a longitudinal survey. This issue is widely developed in the following sections.

Measurement errors refer to discrepancies between the value obtained for an individual and the "true" value. In longitudinal studies, panel conditioning (see e.g. Sturgis et al., 2009) and changes in the data collection process (see e.g. Dillman, 2009) can introduce measurement errors. Furthermore, the way in which data is collected can affect measurement error differently. Retrospective data collection generally produces fewer changes in variables than prospective data collection due to recall error (Lynn, 2009b). In this regard, life-history calendars are a popular tool for minimising recall error when gathering retrospective information. Typically, they take the form of a two-way grid, where one axis represents time and the other comprises the aspects being investigated (Freedman et al., 1988). In the context of life course research, collect information across various life domains enables individuals to connect events in different areas, effectively reducing the likelihood of recall errors (Morselli et al.).

Although measurement errors can sometimes result in missing data, such as when post-survey checks reveal inconsistencies or unreliable responses that have been discarded, missing data is mostly linked to nonresponse error in Groves (2005) "total survey error" framework. In particular, this thesis is not meant to correct for coverage, sampling and measurement errors. More generally, we assume that data have already been collected, so we consider what is called secondary data, and we aim to make the best use of it.

## 2.3 Patterns of missing data

In the two upcoming sections, we define missing data more precisely by examining their distributions, referred to as patterns, and mechanisms. This section focus on missing data patterns, while the next section delves into mechanisms. We first introduce general patterns of missing data and we then focus on patterns specific to longitudinal data.

Little and Rubin (2019) identified several distinctions concerning these patterns. The first distinction is between item-level and unit-level missing data. Unit-level missing data occur when all the information about a subject is missing. This situation happens, for example, when a person selected for the study simply refuses to participate. Item-level missing data appear when only a part of the information is missing for a given subject. This pattern occurs, for instance, when a subject refuses to answer sensible questions, such as questions about income or health. Another distinction is between univariate and multivariate missing data, which refers to missing data affecting one or more variables, respectively. Additionally, the concept of a monotone pattern was introduced, where missing data occur in an order such that when a value is missing for one variable, it is also missing for the following variables.

In longitudinal data, missing data have the tendency to appear under the form of gaps, meaning several observations in a row that are missing. In prospective surveys, this is induced by individuals missing several waves of data collection in a row. Attrition is a special type of

such behaviour. Some individuals simply stop taking part in a survey, inducing missing data to all subsequent waves of data collection. In retrospective surveys, gaps are typically induced by missing spells, where a spell refers to consecutive time points in the same state.

## 2.4 Mechanisms

In his seminal work, Rubin (1976) introduced a classification of missing data into three categories. Understanding these categories is essential in assessing the influence of missing data on the data and statistical outcomes. Consequently, it enables the identification of the most suitable approach for managing missing data, considering that the validity of a method often relies on the mechanism responsible for the missingness. In this section, we present Rubin's classification and provide an illustrative example that will also be used to explain the imputation methods in the following subsection. Finally, we discuss potential tests that can be used to identify the missing mechanism.

According to Rubin (1976), missing data can be classified into three categories: missing completely at random (MCAR), when missing data are a random sample of the entire dataset; Missing At Random (MAR), when the probability of missing depends only on other variables and Missing Not At Random (MNAR) when the probability of missing depends on the missing values themselves. The mechanisms are more formally defined in Appendix A.

Another distinction that is critical for the application of some methods is between ignorable and non-ignorable missing data. Missing data is considered ignorable if it satisfies either the MCAR condition, or MAR along with an additional property that the parameter governing the generation of missing data is independent of the parameter being estimated (Allison, 2001). Nonetheless, in most applications, MAR is considered ignorable.

To illustrate these mechanisms with longitudinal categorical data, let's imagine we measured a specific variable at fifteen occurrences for ten individuals. This variable takes two values ("state A" and "state B"). In addition to these 15 measurements, we know the sex of each person. The dataset is shown in Figure 2.1. Examples of the three mechanisms could be:

- *MCAR* if missing data is just randomly arising.
- *MAR* if women are more likely to have missing data or if the probability of missing at time $t$ depends on the state observed at time $t-1$ (e.g. every time a state A appears, a missing value is more likely afterwards). Therefore, missing data depends on something that is still observed (the state at time $t-1$ in this case)
- *MNAR* if one specific state is missing more often or if a missing value is more likely when a transition occurs. Therefore, the probability of missing depends on something that is not observed any more.

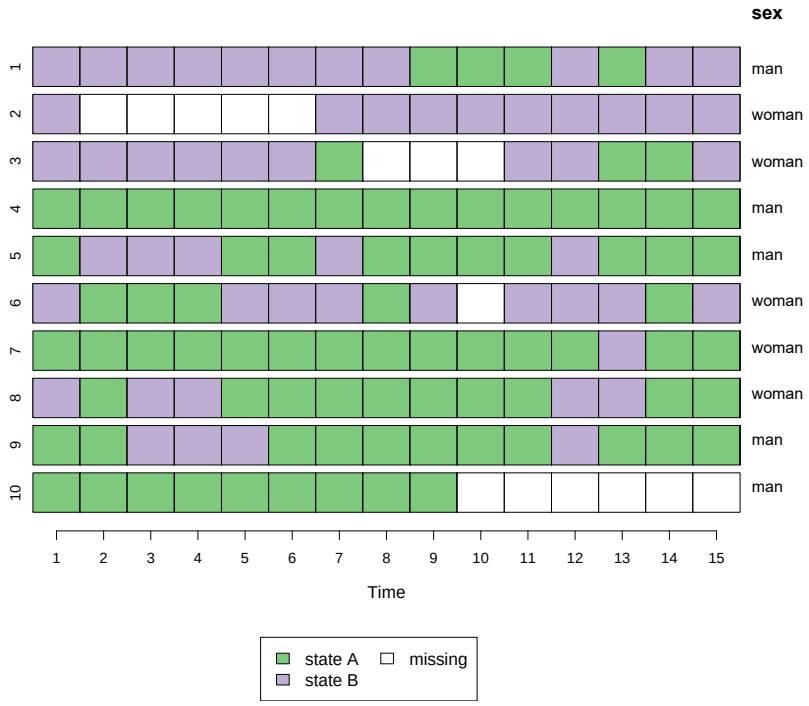



Figure 2.1: Trajectories of ten individuals measured at fifteen time points.

In this example, women are potentially more likely to have missing data than men. Therefore, the missing mechanism may be MAR. However, it is not possible to eliminate the possibility that the mechanism is MNAR.

Since these definitions depend on values that are not observed, it is generally impossible to determine which is the true mechanism in a dataset. Only MCAR is distinguishable from MAR, and statistical tests have been developed for this purpose. For a multivariate numerical outcome, Dixon and Brown (1983) compared the univariate means with t-tests. Little (1988) introduced a likelihood ratio test to assess the difference in means between missing patterns, assuming that the data are multivariate normal. Park and Davis (1993) extended Little's test to longitudinal data. Jamshidian and Jalal (2010) tested the normality and homoscedasticity assumptions simultaneously. When this joint hypothesis is rejected, they propose a nonparametric test to assess the equality of the covariances. Rouzinov and Berchtold (2022) proposed a regression-based approach applicable to numerical and categorical data. However, these tests have low statistical power (Chassan and Concordet, 2023) and are of limited use since in most cases MCAR is an unrealistic assumption (Van Buuren, 2018).

## 2.5 Issues related to missing data

Missing data is a pervasive problem that can have a significant impact on statistical analysis and the validity of research results. There are three main issues related to missing data: reduced statistical power, potential for biased results, and difficulties in applying standard statistical methods (Little and Schenker, 1995).

Firstly, missing data can reduce the amount of information available for analysis, leading to decreased statistical power and inability to detect effects accurately. To illustrate, consider a scenario where we have the salary information for a thousand men and women. In such a case, it would be easier to detect statistical differences between men and women than if we had only one data point for each sex. Therefore, each additional missing data limits the possibility of establishing a possible difference without being wrong.

The second issue that arises due to missing data is the potential for biased results. Statistical bias refers to the tendency of an estimator to systematically produce results that deviate from the actual values of the population parameter being estimated (see Appendix A for a formal definition). To illustrate, let's consider a hypothetical scenario where we want to determine the average income of the Swiss working population, which serves as the population parameter. In this case, the estimator would involve computing the average income based on a sample of this population. When missing data are MCAR, it may not be a significant issue. However, when there are specific reasons why some individuals do not respond, as is often the case, it may be problematic. For example, individuals in vulnerable situations such as unemployment, migration background, or poor health are more likely to leave a longitudinal study (Rothenbühler and Voorpostel, 2016). Therefore, vulnerable situations will be under-represented in the subsample compared to the target population and, hence, probably induce bias in most statistical measures related to vulnerability. In particular, the mean income computed on the sample will likely be underestimated, because individuals in vulnerable situations are more likely to have low incomes.

Finally, numerous statistical methods, such as regression, are primarily designed for complete data, posing challenges when attempting to apply them to datasets with missing information. In cases where these methods are still employed, all incomplete data is typically excluded from the analyses.

## 2.6 Preventing missing data

It may be commonplace but the best way to handle missing data is not to have missing data in the first place, or at least limit missing data. While there are various techniques available to treat missing data, these methods rely on assumptions and may introduce bias or inaccuracies. Prevention can be seen as a form of treatment in itself. Given the critical role of prevention in

minimising the burden of missing data, we outline several prevention strategies before delving into the methods used to handle missing data.

McKnight et al. (2007) identified five aspects of a study that can be linked to missing data:

- Many choices are made when it comes to the *overall study design*, which may impact missing data. In a survey, the design of the questionnaire itself is crucial. For example, shorter questionnaires induce a higher participation rate (Guo et al., 2016) and forced responses cause more interruptions during the completion of the questionnaire (Décieux et al., 2015) and, hence, more missing data. A large body of literature does exist regarding guidelines for researchers for the creation of good questionnaires and, hence, limit missing data (see e.g. Lynn, 2009a).

  Strategies could be developed to minimise the rate of non-response and attrition. For example, in the Swiss household panel (SHP), which is a longitudinal panel survey whose principal aim is to study social changes in Switzerland (Voorpostel et al., 2016), different measures have been put in place: incentives both for the interviewers and the households, a refusal conversion procedure and staying in contact with the respondents. Individual and collective incentives for interviewers may increase their productivity (Thurkow et al., 2000). For participants, incentives, which take the form of money sent with the announcement of the new wave in the SHP, have a positive impact on survey participation (see e.g. Lipps et al. (2019) for a thorough discussion of incentives in surveys). In most surveys, a procedure, called *refusal conversion*, is applied to convince individuals that either refuses to participate in the current wave or have not participated in previous waves to participate. Dangubic and Voorpostel (2017) detail the procedure applied to the Swiss household panel and its positive impact on participation. For example, in 2015, the number of participating households increased by approximately 6%.

- The *characteristics of the target population and the targeted sample* must be considered. For example, proposing questionnaires in the mother tongue may help reach a specific foreign population, such as what is done with the Parchemin study, which focuses on undocumented migrants, where questionnaires are provided on the most common languages spoken by them (Jackson et al., 2019). Moreover, a study conducted on elderly populations differs in methodology from those conducted on adults (e.g. Oris et al., 2016).

- The *mode of data collection and measurement* is of importance, and for each of them, strategies are applicable to reduce missing data. Web surveys allow reaching the largest sample at the lowest cost but suffer from a greater nonresponse rate than other modes of data collection (Daikeler et al., 2020). Along this line, Voorpostel et al. (2021) realised a pilot study on the Swiss household panel to compare standard phone data collection, full

web design, and a mixed design combining phone-based collection for household questions with web-based individual questionnaires. The study revealed that phone-based data collection achieved the highest response rate. However, the lowest response rate of the mixed design might be compensated by the widest population reached. Anyway, as technology evolves, the data collection field is also evolving. Lately, the use of smartphones to collect data has been under study (e.g. Link et al., 2014; Roberts et al., 2022). With smartphones, additional types of data may be collected, such as, for example, the time spent on each application or GPS locations.

- *Treatment implementation* is specific to experimental design, where the sample is split between a control and a treatment group. The burden on the individuals of the treatment group should be minimum, for example, in terms of the number and duration of sessions. Something similar applies to surveys, where the burden on the interviewee should be minimised while still gathering the desired information. Some surveys, such as the European Union Statistics on Income and Living Conditions (see e.g. Arora et al. (2015) for a description), apply rotating samples, which means a sample stays in the survey for a definite number of data collection waves before being replaced by another one.

  Another possibility is to induce a rotation between the topics addressed by a survey, such as in the European Social Survey (see e.g. Jowell et al. (2006)), where a part of the questionnaire is fixed every year and another part change every year, based on the suggestion from researchers. With this rotative design, more subjects can be investigated without increasing the questionnaire length.

- *Data entry* is all that is linked to the transformation of the data collected to make them usable, such as, for example, recopying answers to paper questionnaires on a computer. Since it is generally a monotone and cumbersome process, it is worth simplifying it to the maximum. In this area, methods that do not require data to be re-entered, such as web-based questionnaires, have a clear advantage.

Nevertheless, even with a meticulous planned study, missing data cannot be entirely avoided (see e.g. Eekhout et al., 2012; Berchtold, 2019).

## 2.7 Methods for dealing with missing data

As this thesis focuses on comparing imputation methods for life course data and proposing novel approaches, a thorough understanding of the currently available methods for handling missing data is essential. In this section, we provide a detailed review of the main families of methods, starting, for each of them, with a general overview and then focusing specifically on

life course data. We will use the example introduced in the previous section to illustrate them. In particular, we discuss their strengths and weaknesses and how the missing data mechanisms impact their use.

Methods to deal with missing data are classifiable into four categories: deletion, weighting, likelihood and Bayesian, and imputation methods (Molenberghs et al., 2014).

### 2.7.1 Deletion methods

Deletion methods are a common approach that involves removing observations with missing data. There are two main types of deletion methods: complete case analysis and available case analysis. When a strategy to deal with missing data is used, deletion procedures are still the standard procedure in social science research (Berchtold, 2019), even if they are unsuitable in most cases.

With complete case analysis, also known as listwise deletion, all observations with at least one missing value are deleted. Therefore, in longitudinal data, a single value in any variable and time point, lead to the deletion of an entire case. In the illustrative example, the second, third, sixth, and tenth trajectories have at least one missing value and are, hence, deleted, giving the completed dataset shown in Figure 2.2. It is composed of six trajectories, where men are overrepresented.

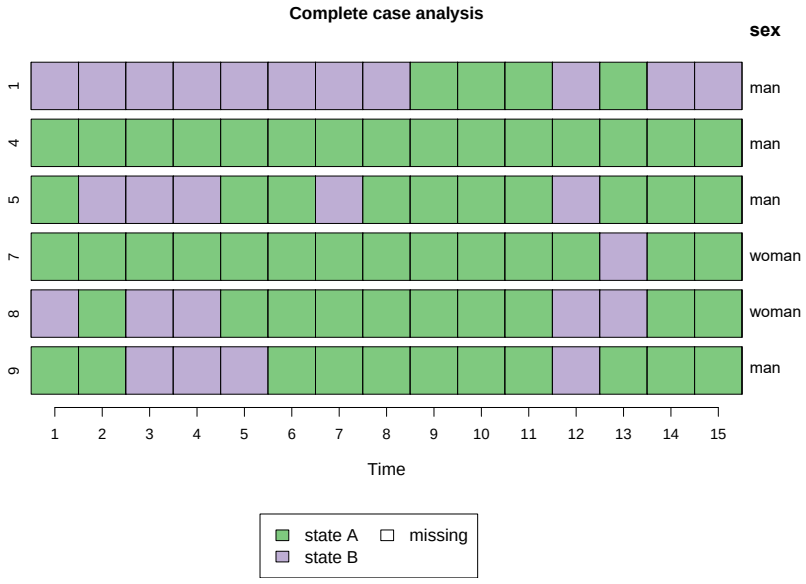



Figure 2.2: Complete dataset resulting from the application of complete case analysis.

Complete case analysis produces unbiased results when missing data are MCAR. However, this approach may lead to the loss of a substantial number of observations. This can cause

inflated standard errors and reduced statistical power since both measures depend on the sample size (King et al., 2001). For missing data that are MAR or MNAR, complete case analysis generally leads to bias, and in some cases, it can even result in spurious conclusions (see e.g. Perkins et al. (2018)). However, some specific MAR and MNAR scenarios can yield unbiased results, such as in a linear regression analysis when the probability of missingness does not depend on the outcome (Glynn and Laird, 1986; Little, 1992), or in a logistic regression when missingness is dependent on the outcome or the exposure but not both (see e.g. Bartlett et al. (2015) for a thorough exploration of the scenarios that lead to unbiased results in logistic regression).

With available case analysis, also known as pairwise deletion, only the observations with missing values in the variable(s) under study, are deleted. Regarding our illustrative example, the application of available case analysis depends on the analysis. For example, if we are explicitly interested in the transition between times 3 and 4, pairwise deletion leads to a dataset of nine trajectories (Figure 2.3), where only the second trajectory is deleted. However, only seven trajectories remain when the focus is on the transition between times 9 and 10.

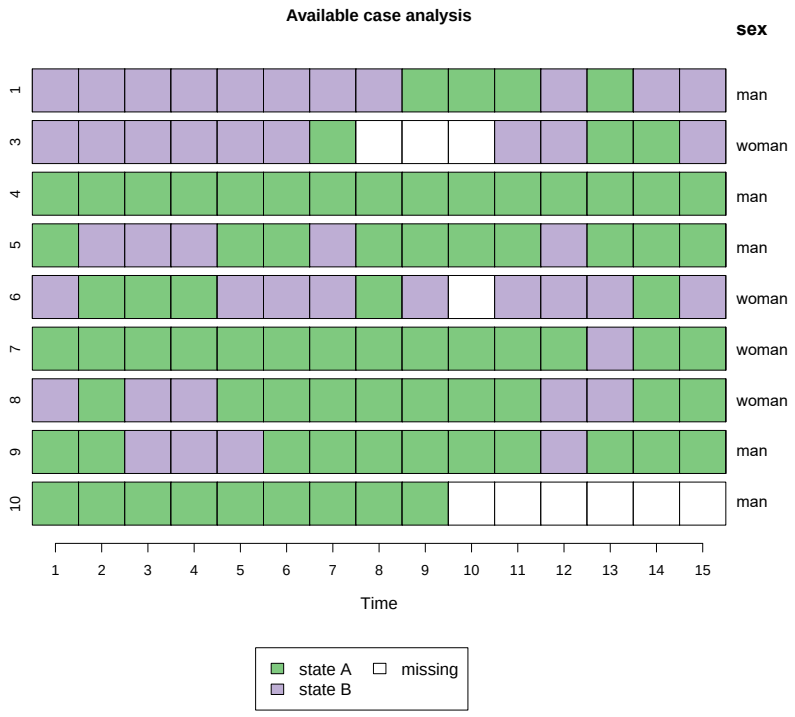



Figure 2.3: Dataset resulting from the application of available case analysis.

Available case analysis shares the same conclusion as complete case analysis regarding the missing data mechanisms, i.e. it is unbiased when the missing mechanism is MCAR but biased

in most MAR and MNAR scenarios. Compared to complete case analysis, available case analysis has the advantage of utilising every piece of information available, resulting in a generally smaller reduction of the sample size. However, the method has several drawbacks. First, analyses are difficult to compare since they are realised on different datasets. Then, it could lead to implausible results, such as correlations outside the range -1 to 1 or variance-covariance matrices not positive definite (Pigott, 2001). Finally, there is no clear way to compute the standard errors, since their computation is dependent on the sample size that may vary between each analysis (Graham, 2012).

### 2.7.2 Weighting methods

The idea behind weighting methods is to partly correct the issues related to complete case analysis by assigning different weights to the remaining observations. The application of weighting methods do not differ depending on whether the data are longitudinal or not.

To compute the weights, a first body of methods rely on adjusting some characteristics of the sample obtained from complete case analysis to the original. Adjustments can be made for each combination of covariates, known as "cell weighting", or for the marginal distribution of each covariate using methods like "raking" or "linear weighting". Alternatively, adjustments can be made for a specific population parameter, known as "GREG weighting" (for more details on these methods, refer to Kalton and Flores-Cervantes (2003)). For example, in the complete dataset obtained through complete case analysis (Figure 2.2), the representation of men and women is different from that of the original dataset. Conducting an analysis on this dataset alone would result in a larger impact of men on the results compared to women. Therefore, a solution is to assign a higher weight to women's trajectories than to men's in order to compensate for their under-representation in the complete dataset. In this scenario, weights of 3/2 and 3/4 are assigned to women and men, respectively.

However, with a high number of covariates, these methods are not suitable since very few, or even no observations, could appear in some combinations of covariates. For example, with 10 binary covariates, there will be 1024 possible combinations. Therefore, the most common approach relies on logistic regression weighting (which is also often called propensity score weighting). The probability to be missing is predicted through a logistic regression model. The weight of each observation is then set as the inverse of the predicted probability to be observed for this observation. The pool of variables that can be included in the model can be quite large, going from characteristics of the individuals, such as the age or the sex, to paradata, such as notes of the interviewer on contact records (Buskirk and Kolenikov, 2015).

As a method of dealing with missing data, the use of weights has both advantages and weaknesses. When missing data are MCAR, complete case analysis provides unbiased results, and weighting does not provide any improvement. As for deletion methods, it suffers from a

reduction of the statistical power and does not make use of partially observed cases, which often contains substantial information (Little et al., 2022). Weighting methods can correct for bias when the MAR hypothesis depends on variables that are used to compute the weights. However, it is limited to variables available both for respondents and non-respondents. These methods will generally lead to bias when the MAR mechanism depends on other variables. Finally, when missing data are MNAR, weighting methods, like most methods, lead to bias.

### 2.7.3 Likelihood and Bayesian methods

The third family of methods to handle missing data are the likelihood and Bayesian approaches, which assume that a statistical model has generated the data. We first provide an overview of these methods and highlight the key differences between likelihood and Bayesian methods. Then, we detail how these methods apply with longitudinal categorical data. Finally, we discuss how the application of these methods is influenced by the missing data mechanism.

#### General framework

Likelihood and Bayesian methods assume that a statistical model generated the data and seek to estimate this model. Likelihood methods adopt a frequentist perspective, treating the parameters of the model as fixed. In contrast, Bayesian methods involve making prior assumptions about the distribution of parameters, enabling the representation of uncertainty regarding the missing data generation process through the specification of priors (Daniels and Hogan, 2008).

These methods apply when the target of the analysis is either directly the parameter of the statistical model or a parameter that is derivable from it (e.g. regression coefficients, correlations, or global transitions between the states in a trajectory). When it is not the case, for example, for clustering, missing data have to be imputed from the conditional distribution. This is discussed in the next section, which concerns imputation methods.

A key point of the Bayesian framework is the specification of the prior distribution. On the one hand, this could capture information from previous research or hypothesis about the missing mechanism. In this case, the prior is called “informative”. On the other hand, the impact of the prior could be minimum, called “non-informative” prior. Different strategies exist to construct “non-informative” priors (e.g. Berger and Bernardo (1992); Jeffreys (1998)). However, in the context of missing data, we mostly rely on informative priors that capture assumptions about missing data (Daniels and Hogan, 2014). Using a “conjugate prior” simplifies computations by ensuring that the prior and posterior distributions are identical. For example, with a binomial likelihood, a prior following a beta distribution leads to a posterior distribution that also follows a beta distribution (see e.g. Gelman et al., 1995, for a thorough introduction

on prior distributions and more generally on Bayesian analysis).

### Likelihood methods with longitudinal data

The application of a Gaussian model, which is, in most cases, the standard strategy, is generally not suitable with longitudinal categorical data. Indeed, the goal of the statistical analysis of a categorical longitudinal dataset is rarely a parameter that can be derived from a Gaussian model. For example, the focus is sometimes on the probabilities of moving from one state to another, which cannot be extracted from a Gaussian model.

Markov chains are typically used (see e.g. Brémaud (2013) for a comprehensive description of Markov chains) with longitudinal categorical data. The most naive model is the independence model, where each time point is generated independently of the others. A $\ell$-order Markov chain supposes that the probability of experiencing a given state at time t depends only on the value at time $t-\ell, \ldots, t-1$. Hidden Markov models (Baum and Petrie, 1966) have a more complex structure. They suppose a latent Markov model which determines at each time point the probability distribution that generates the observed value. Double Chain Markov Models (Berchtold, 1999) combine an observed and a latent Markov chain. With high-order Markov chains and, even more, with Double Chains Markov Models, the number of parameters to estimate increases quickly. To overcome this issue, one can use both these models in combination with a mixture distribution transition model (Berchtold and Raftery, 2002), where a single transition matrix and a vector quantifying the impact of each lag are estimated. Variable-length Markov chains (Bühlmann and Wyner, 1999) are another type of Markovian model having the specificity that the number of past states sufficient to summarise the whole past varies with each situation. These models assume that the process is homogeneous.

Estimating these models when missing data are involved is generally more complex. Yeh et al. (2010) identified three methods to estimate a Markov chain when the trajectories are subject to ignorable missing data. The first strategy, called the One-Step Method, computes the transitions observed on the trajectories. The main drawback of this method is that it does not consider the longitudinal structure of the trajectories. The authors have shown that it lags behind the two other ones, which present similar performance in the setup they considered. For the sake of illustration, we suppose we are interested in the global transition probability between state A and state B in our example. Missing data gaps break the trajectories into sub-trajectories (Figure 2.4). For example, the second trajectory with states 2 to 6 missing is broken into two sub-trajectories: one consisting of one state (2.1) and one of nine (2.2).

The second strategy considers the missing values in the computation of the likelihood. For example, let's consider the sixth trajectory. The contribution of the transition from state B at time 9 to state B at time 11 is added to the computation by considering the possibility of the state at time 10 being either A or B.

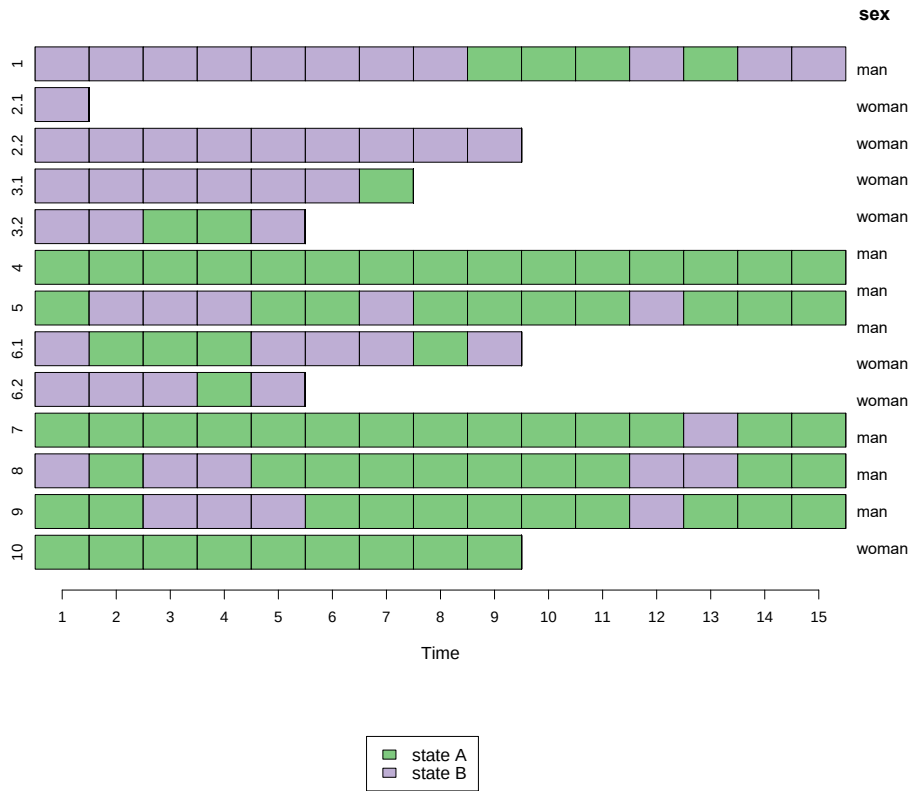



Figure 2.4: Sub-trajectories used to fit the Markov model.

Finally, the estimation could be done through the expectation-maximisation (EM) algorithm (M'Kendrick, 1925; Hartley, 1958; Dempster et al., 1977). It alternates between two steps: one that finds the expectation of a function of the missing data and another that estimates the parameters of the complete data. Several extensions of this algorithm have been proposed, mainly to cope with situations where the maximisation step has no easy computational form or to speed up the process (Meng and Rubin, 1993; Liu and Rubin, 1994; Liu et al., 1998; Meng and Van Dyk, 1997).

Moreover, Yeh et al. (2012) discuss the computation of the hidden Markov model with ignorable missing data. To our knowledge, there was no adaptation of their estimation with missing data concerning the mixture transition model, the Double-Chain Markov Model, or Variable-length Markov Chains.

### Impact of the missing mechanism

The notion of ignorability plays a central role for Likelihood and Bayesian methods. When the missing data generation process is ignorable, only the model of the data needs to be specified,

while when it is non-ignorable, one also needs to specify the missing data generating model (see Appendix A for the statistical details).

There are two ways of modelling non-ignorable data (Glynn et al., 1986). Selection models (Heckman, 1976) specify the distribution of the complete data and the conditional distribution of the missing mechanism on the complete data, while pattern-mixture models (Rubin, 1977; Little, 1993) specify the marginal distribution of the missing data and the conditional distribution of the complete data on the missing values (see Appendix A for more details). However, the distributional assumptions made for these models cannot be verified. Indeed, for every MNAR model, there exists a corresponding MAR model that yields the same fit on the data (Molenberghs et al., 2008). Therefore, non-ignorable models are mainly used in a sensitivity analysis, which is applied to determine the impact of a departure from the MAR assumption.

### 2.7.4 Imputation

The process of imputation involves replacing missing data with either one plausible value (single imputation) or several plausible values (multiple imputation). We first introduce single imputation methods and the pitfalls linked to it. Then, we focus on multiple imputation. Specifically, we begin by introducing its general functioning before delving into the main methods to produce multiple imputations with longitudinal categorical data, namely joint modelling, fully conditional specification, and the MICT algorithm.

#### Single imputation

Numerous single imputation methods are available. Strictly speaking, every strategy that fills the missing data with some values is a single imputation method. In the case of longitudinal data, we can cite for example:

- Last observation carried forward, where the last observed value is imputed to the following missing data. For example, for the second trajectory, the state B observed at time 1 is copied to times 2 to 6. Figure 2.5 shows the resulting completed dataset. However, this method cannot impute gaps of missing data starting the sequences and it is not suitable for trajectories showing high transition rates. Moreover, it can lead to bias even when the data are MCAR (Molenberghs and Kenward, 2007).

- The imputation of the mode, meaning the most observed state at a given time. For example, at time 2, state A is observed six times, while state B appears only three times. Therefore, the missing value arising at time 2 for the second trajectory is imputed with a state A. This process leads to the completed dataset shown in Figure 2.6.

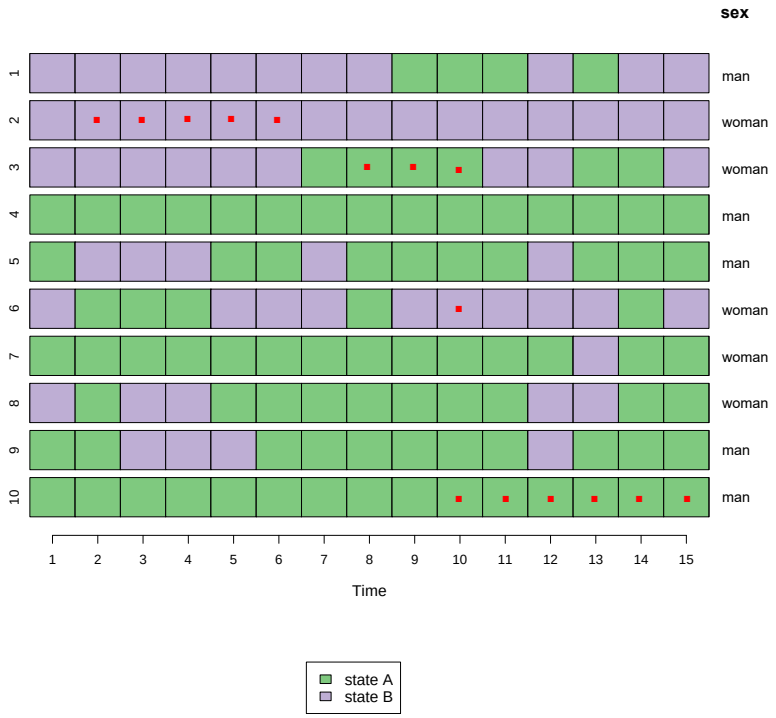



Figure 2.5: Completed dataset obtained with the last observation carried forward. Red dots localise the imputed values.

- A random draw from the distribution at a given time point. For example, the first missing value of the second trajectory has a probability of 2/3 to be imputed with state A and 1/3 with state B. Figure 2.7 shows a completed dataset induced with this method. Contrary to mode imputation and last observation carried forward, as the process is random, doing it again will most likely lead to a different imputed dataset.

In addition, multiple imputation methods, that are detailed below, are generally applicable to build a single completed dataset.

The main drawback of single imputation methods is that they do not account for the variability of the missing data, yielding an underestimation of the true variance of the parameters (see e.g. Schafer and Olsen, 1998; Donders et al., 2006).

### Multiple imputation

Multiple imputation was designed to solve the lack of variability issue arising from single imputation. A number $M$ of possible imputation values is determined for each missing data, giving $M$ completed datasets. The statistical analysis is then realised independently on each imputed dataset before aggregating the results. The decisive advantage of multiple imputation

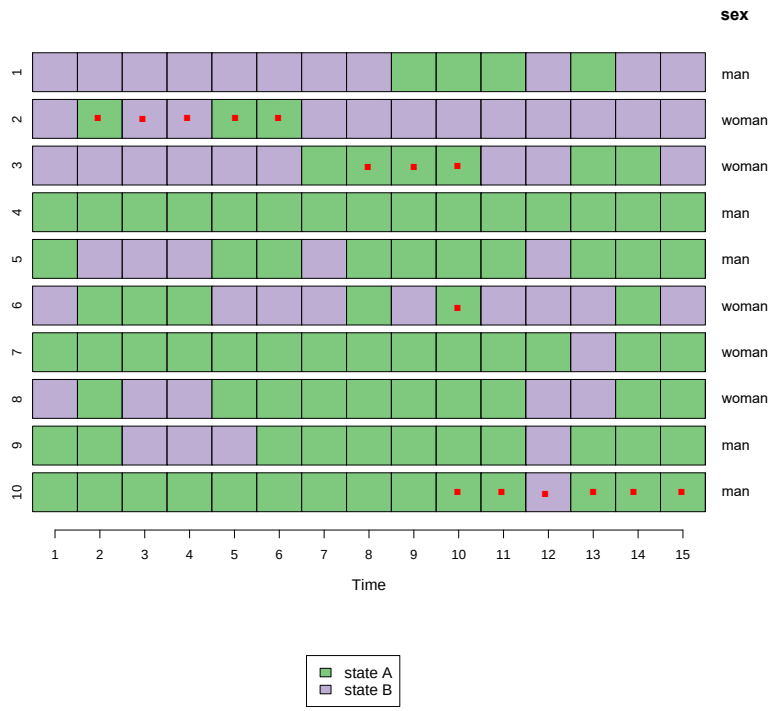



Figure 2.6: Completed dataset obtained by imputing the mode. Red dots localise the imputed values.

over single imputation is a better handling of the variability of the data, hence a lower risk of producing significant artificial effects during inferential analyses. The idea of multiple imputations is illustrated in Figure 2.8. The dataset with missing data is imputed a specified number of times (here, $M$). Suppose we are interested in the probability to change from one state to another (and its variance). These probabilities are computed on each completed dataset separately, then the results are finally aggregated. Rubin (1987) derived rules to combine the results from several completed datasets when a normally distributed parameter is the main focus. If this is not the case, the idea is to apply a transformation to make it normal and then apply Rubin's rule. Van Buuren (2018, p. 146) summarises the transformation towards normality of often-used statistics such as correlation, hazard ratio, or the coefficient of determination $R^2$. The combination of the results is not always straightforward. For example, when the objective is to perform a cluster analysis, small changes in data can lead to large differences in a grouping. Clusterings built on different completed datasets may not even have the same number of groups. Halpin (2012) proposed to apply cluster analysis to the pool of imputed datasets. Then, an observation is either assigned to the cluster which contains most of the imputed replications, or a degree of membership to each cluster is associated with each observation based on the share of imputed data assigned to each cluster.

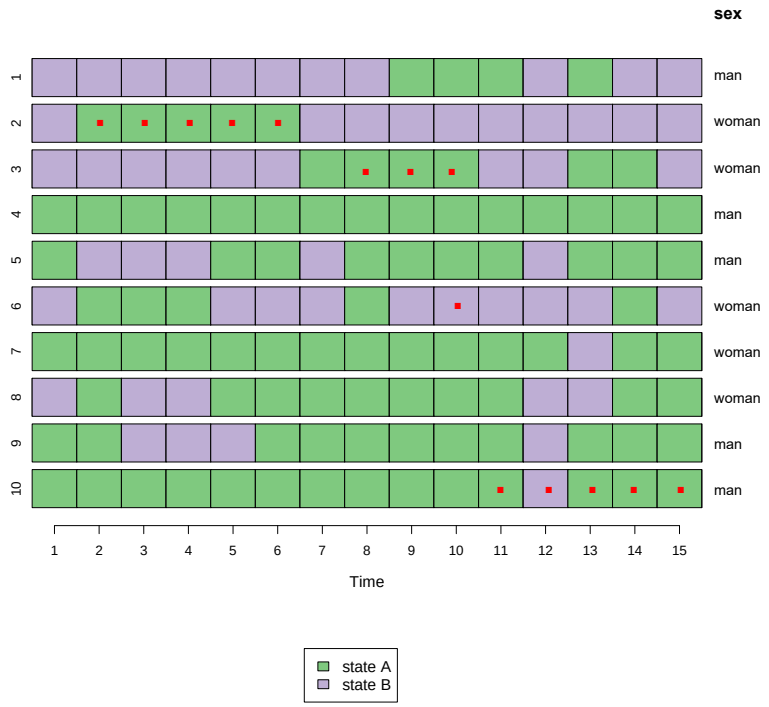



Figure 2.7: Completed dataset obtained with random draws from the observed distribution. Red dots localise the imputed values.

Early works (Rubin, 1987; Schafer and Olsen, 1998) suggested creating a low number of datasets (between 3 and 5). The main argument was that the gain obtained from increasing the number of imputations was not worth the computational burden. However, with the improvement of the calculation capacities of computers, the calculation time could be drastically reduced, and further analysis concluded that a larger number of imputations, more than 20, is generally better (Royston, 2004; Graham et al., 2007). Several authors suggested rules linking the number of imputations to the theoretical proportion of missing data in the population (e.g. Bodner, 2008; Von Hippel, 2020). Along this line, White et al. (2011) have shown that if the number of imputations is approximately 100 times the proportion of the variation explainable by missing data, some properties regarding the reproducibility of the results are satisfied. As a rule of thumb, they suggest setting the number of imputations to the percentage of incomplete cases, which approximates the fraction of missing information. Van Buuren (2018) suggested keeping the number of imputed datasets low when determining the imputation model and increasing it when computationally feasible for the final creation of the imputed datasets.

We discuss now the three main multiple imputation methods to deal with missing data in longitudinal categorical data, namely joint modelling, FCS and MICT.

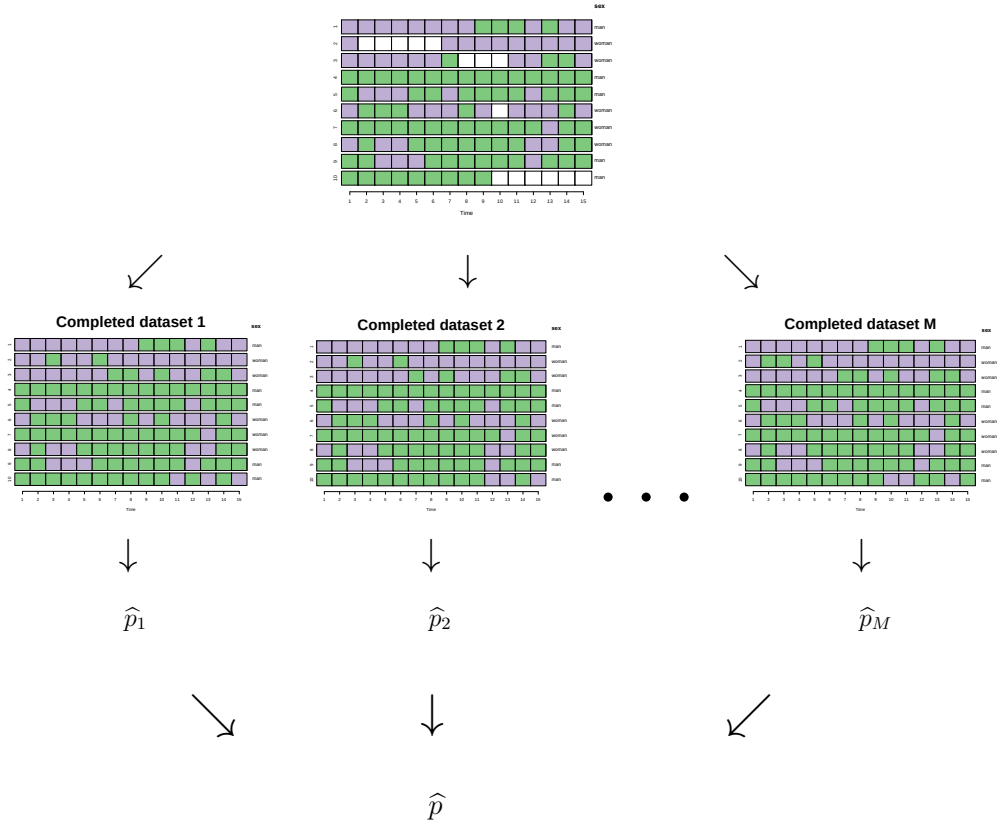



Figure 2.8: Illustration of a multiple imputation process.

**Joint modelling** imputation is closely related to likelihood and Bayesian methods, but differs in that it involves drawing imputations for missing values. This approach is useful when the statistical analysis is not directly linked to the estimated model, as is often the case in clustering analysis. First, we detail the application of the models outlined in the section on likelihood and Bayesian methods in this context. Next, we describe the use of a multivariate normal model, which is commonly employed, even in situations involving categorical data.

The different models detailed in the section about the likelihood and Bayesian methods are available for modelling the data. The difference is that imputations are finally drawn. For example, let us consider a Markov model of order 1 (i.e. the previous time point is enough to summarise the whole past). The second time point of the second trajectory is imputed based on the probabilities to transition from a state B to either a state A or a state B. Then, the third time point is imputed based on the probability of transitioning from the state imputed at time 2 to either of the two states until each gap is filled. The whole process of imputation is repeated several times.

Now we discuss the different strategies to apply the multivariate normal distribution to longitudinal categorical data. We first focus on the simplified case of a binary categorical

variable, before expanding to the general case. As suggested by Schafer (1997), one can first recode categorical binary variables as 0 and 1 (Figure 2.9). Then, it is supposed that a multivariate normal distribution generated the trajectories. In the example, a multivariate normal distribution of dimension fifteen (corresponding to the fifteenth time points) is supposed. The data-augmentation algorithm (Tanner and Wong, 1987) is typically applied. It alternates between drawing a value for the model's parameter from the complete-data posterior distribution and imputing the missing values. Schafer (1997) has detailed the algorithm for several statistical models. The data-augmentation algorithm is close to the EM algorithm, and it is also possible to apply the EM algorithm to obtain multiple imputation. However, it needs an extra step, where the data are imputed based on the posterior distribution. This is what is done in the R *Amelia* package (Honaker et al., 2011), which applies EMB (Honaker and King, 2010), a variant of the EM algorithm. $M$ bootstrap samples of the data are drawn, and the maximum likelihood is maximised on each of these samples with the EM algorithm. Imputations are then drawn based on the conditional distributions of the missing values on the observed ones. These conditional distributions are easily derived from the joint distribution in the case of multivariate normal models. For example, for the second trajectory, the conditional distribution of the time points 2 to 7, knowing that the times 1 and 8 to 15 have value 1, is computed. Based on this distribution, values (numerical) are drawn for the times 2 to 7. The imputed numerical values are rescaled as probabilities, and one of the two states is drawn based on these probabilities. For example, suppose a numerical value of 0.35 was set for the second time point of the second trajectory. In that case, the probabilities are 0.65 and 0.35 of imputing, respectively, states A and B. The imputation process is repeated several times to create several completed datasets.

The multivariate normal distribution is also applicable to a categorical variable with more than two categories. In this case, a categorical variable with $k$ categories is transformed as $k-1$ binary variables (Allison, 2001). For example, if we have a variable that has three categories: "full-time employed", "part-time employed", and "unemployed", it is recoded as two binary variables. The "full-time employed" category will correspond to a 1 in the first binary variable and 0 in the second, the "part-time employed" to a 0 in the first binary variable and 1 in the second, and the "unemployed" state to 0 in both variables. The multivariate normal model is usually fitted with the EM algorithm, and imputations are drawn for each of the $k-1$ variables. A value of one minus the values of the $k-1$ variables is set for the last category. Then, the category can either be chosen based on the imputed variable that has the largest values, or the $k$ imputed variables can be first transformed into probabilities (simply by scaling their sum to 1) before drawing a category based on these probabilities (Honaker et al., 2011). There are two main problems with these approaches. First, it is under question whether a Gaussian distribution, which is a continuous distribution, can effectively approximate the discrete binary variables. Then, the number of parameters to estimate increases quickly. For example, with a

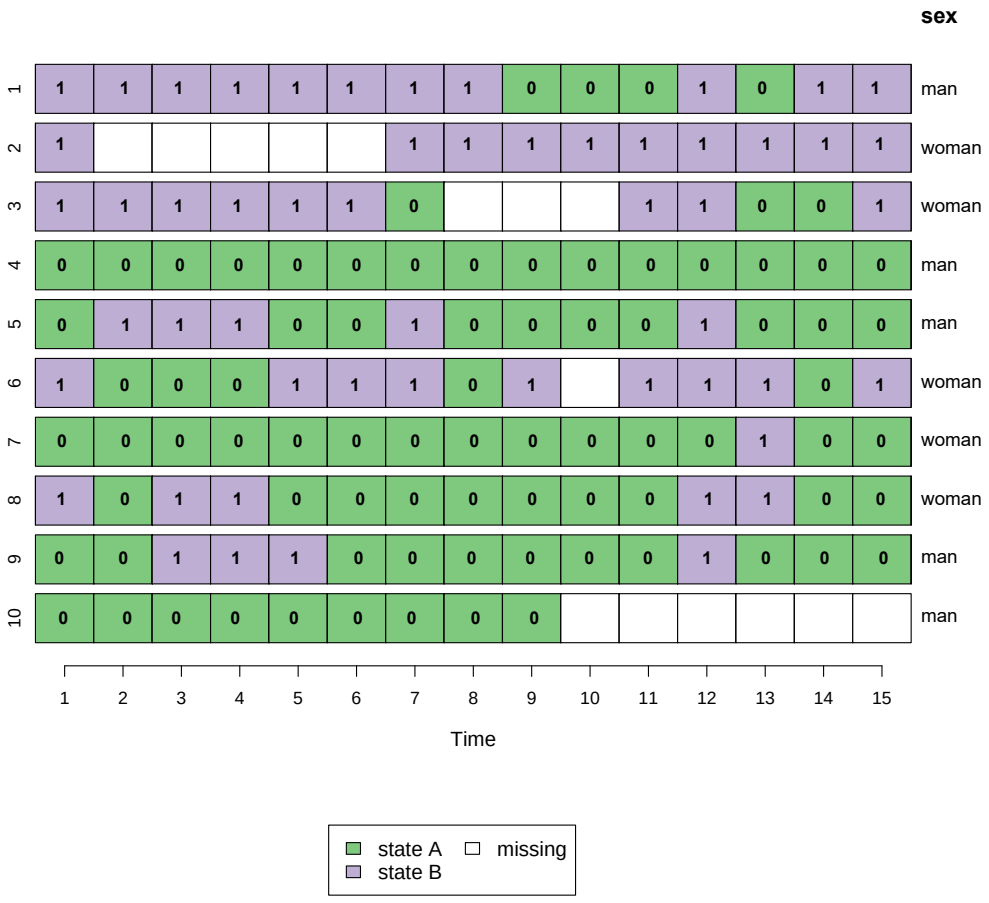



Figure 2.9: The values are transformed from categorical to numerical. Each state "A" is recoded as "0" and each state "B" as "1".

single categorical variable observed at ten time points and composed of four categories, there are 30 parameters related to the mean and a covariance matrix of size 30x30 to estimate.

**Fully conditional specification** (*FCS*), also known as chained equations, imputes missing data through the specification of a conditional distribution for each individual variable (Van Buuren et al., 2006). It works in the following way:

1. For each variable, missing data are first imputed from their marginal distribution.

2. An imputation model is defined for each variable, generally a linear regression for continuous variables, a logistic regression for dichotomous variables, and a multinomial regression for categorical variables with more than two categories. The predictors usually include every other variable.

3. For each variable, the imputation model is fitted from all available data; then, this model is used to predict a replacement value for each missing data. The algorithm goes through the variables until it reaches a predefined number of iterations.

4. The values obtained after the last iteration are either kept or replaced by the closest values appearing in the dataset, which is called predictive mean matching. The latter applies only to numerical values.

5. The whole process is run several times if multiple imputation is required.

The imputations are drawn from a joint distribution (Hughes et al., 2014) under the hypothesis that the conditional distributions are compatible and the margins are non-informative, but it is generally challenging to prove that these hypotheses are satisfied. Violation of the non-informative margins assumption could induce differences depending on the order in which the variables are imputed. However, in a simulation study based on a general location model, where margins are informative, order effects remain small (Hughes et al., 2014). Another simulation study done by Van Buuren et al. (2006) shows that even substantial violations of the compatibility assumption do not impact the statistical properties of multiple imputation through chained equations.

FCS applies to longitudinal data by considering repeated measurements as distinct variables. However, Kalaycioglu et al. (2016) have shown that fully conditional specification may face convergence issues in its standard form due to collinearity. Two-folds fully conditional specification (Nevalainen et al., 2009), which is a variation of fully conditional specification, tries to solve these convergence issues by limiting the predictors used in an imputation model. In addition, the algorithm imputes several times the variables at a given time point before skipping to the next time point.

FCS is implemented in the main statistical softwares, such as in *S-PLUS* (Van Buuren and Groothuis-Oudshoorn, 1999), the *R* package *mice* (Van Buuren and Groothuis-Oudshoorn, 2011), the STATA module *ice* (Royston, 2004), the SAS *MI* procedure (Yuan, 2010) and more recently in *SPSS* (SPSS, 2017).

Let's illustrate FCS with the example. Each missing data is first imputed by randomly drawing from the distribution at each time point. Then, FCS goes iteratively several times through all time points to improve the imputations. Since the first time point does not have any missing values, the second time point is the first to be considered. A logistic model is fitted with the second time point as the dependent variable and, generally, all other time points as independent variables. Then, a value is redrawn for the missing value on the second trajectory. The same process applies to each time point a predefined number of times. Finally, the whole process is repeated several times.

**The MICT algorithm** was created by Halpin (2012, 2013, 2016b). It considers missing data characteristics in life course sequences. This algorithm is central to this thesis since several extensions are developed.

The algorithm fills the internal gaps recursively from their edges, taking mainly into account the adjacent states. The algorithm works by distinguishing six missing data patterns and handles them sequentially. It then adopts a slightly different imputation method for each of them. Figure 2.10 presents six sequences taken from a built-up dataset composed of two states (states A and B) and fifteen time points, each illustrating one of these patterns. The imputation model uses a definite number of past and future observations. For the sake of the illustration, let us assume that we use two time points in the future ($nf = 2$) and the past ($np = 2$) for the imputation.

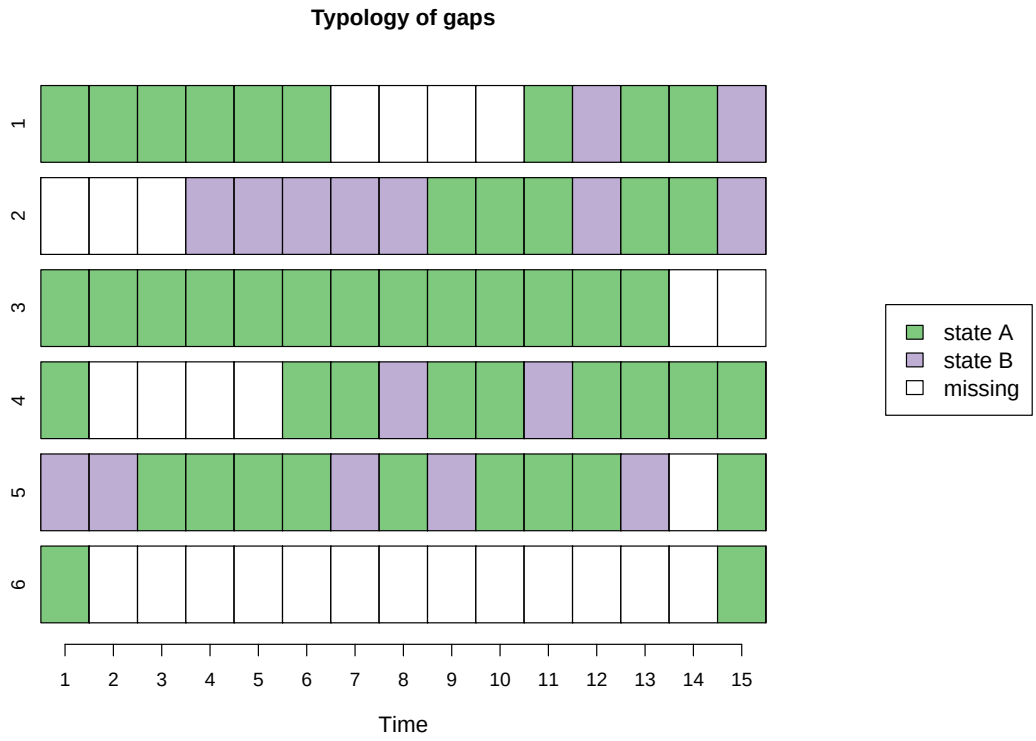



Figure 2.10: Typology of the different type of missing data gaps according to the *MICT* algorithm considering two predictors from the past and the future: 1. Internal gap, 2. Initial gap, 3. End gap, 4. Left-hand side gap, 5. Right-hand side gap, 6. Both-hand side gap.

Sequence 1 illustrates an *internal gap*. Here, sufficient information, i.e. at least two observations before and after the gap, is available to impute the missing data. In this case, the *MICT* imputation process fills gaps recursively from their edges. This strategy ensures that imputations are coherent and based on the closest observed values. Figure 2.11 illustrates the ordering of the imputations for two toy sequences with gaps of different lengths.

Concretely, the imputations are made using a multinomial model. In our example, this

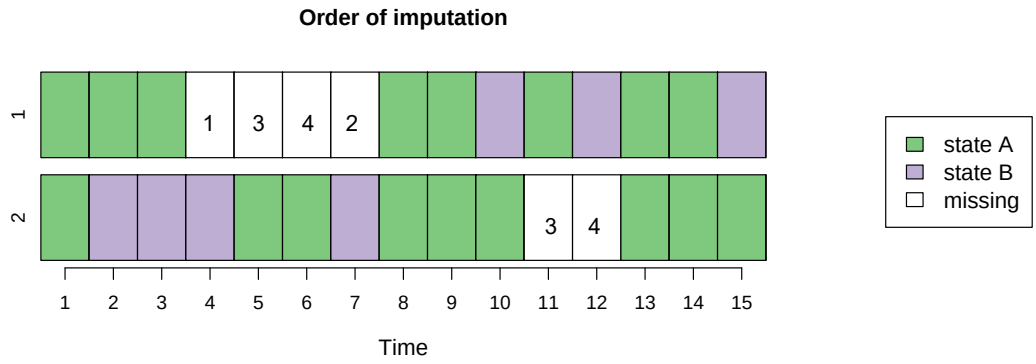



Figure 2.11: Imputation order for an example with two gaps of different lengths.

model uses predefined covariates (such as sex), the two previously available values in the sequence, either observed or previously imputed, and the two ($nf = 2$) subsequent values. The first missing data of Figure 2.11 uses the two values before the gap, i.e. the states at times 2 and 3, and the two following the gap, i.e. the states at times 8 and 9. This multinomial model is first estimated using similar fully observed patterns in the data. For example, the second trajectory would provide six observations: predicting the state at time 3 using the states at times 1, 2, 7, and 8 as predictors, the state at time 4 using the states at 1, 2, 8, and 9, until the state at time 10 using the states at times 8, 9, 14 and 15 as predictors.

Once all internal gaps are imputed, the process considers initial and terminal gaps. Sequences 2 and 3 of Figure 2.10 illustrate this type of gap. Since initial gaps have only one edge, imputations start here from the far right to the left and can use only predictors from the future. Indeed, there are no observed data points in the past. Here again, the imputation model is estimated based on similar fully observed patterns in the data, regardless of their location within the trajectories. The same (but reverse) strategy applies to terminal gaps.

Finally, the rarest cases, namely left-, right- and both-hand side gaps, are imputed at the end. Concerning left-hand side gaps, the imputation is split into several parts, depending on the number of states available before the gaps. Concretely, it starts by imputing the gaps that have one state available before (and that therefore start at the second time point), then two until imputing the gaps that have $np - 1$ states available before. The imputation is then identical to the internal gaps, the difference being that the number of past observations used in the imputation models is reduced. The imputation of right-hand side gaps is similar to left-hand side gaps; the process is split according to the number of observations available after the gaps. Finally, the imputation of both-hand side gaps is split according to the information available beforehand and afterwards.

In Stata, the module *MICT* (Halpin, 2015) and, in *R*, the package *seqimpute* (Berchtold et al., 2022) implement this strategy.

Multiple imputation, as with likelihood and Bayesian methods, induces unbiased results when the missing data are ignorable, provided the variables involved in the MAR mechanism are integrated in the imputation procedure and the imputation models are congenial to the statistical analysis (Rubin, 1987; Little, 1992). The latter implies that all variables used in the statistical analysis must be included in the imputation models. In contrast, multiple imputation may induce bias when the missing data is non-ignorable. Van Buuren (2018) provided guidelines if a non-ignorable mechanism is suspected. First, add more relevant variables in the imputation models so that the missing mechanism is closer to MAR. If it is not sufficient, perform a sensibility analysis, meaning testing different MNAR scenarios in order to determine how the results are impacted. Finally, explicit the non-response, in a similar manner as what was developed in the part about likelihood and Bayesian methods. However, one cannot generally differentiate between MAR and MNAR and unverifiable assumptions need to be made. Therefore, the most commonly used approach when dealing with missing data is to assume that the missing data are MAR (Van Buuren, 2015).

## 2.8 Sequence analysis

To study the life course, one can either focus on the transitions themselves, using Markov models or event history analysis, or on the trajectory as a whole (Piccarreta and Studer, 2019). The border between the two approaches is permeable, and some methods, such as Competing Trajectory analysis (Studer et al., 2018a), Sequence Analysis Multistate Model (Studer et al., 2018b), or Sequence History Analysis (Rossignon et al., 2018), link them together. In this thesis, we focus on the trajectories as a whole and, hence, on sequence analysis.

The core idea of sequence analysis is to consider life courses as a whole, with the idea that life events cannot be understood separately (Abbott, 1983, 1995). The situations experienced at different times are coded with a pool of states, creating a succession of states called a "sequence".

Sequence analysis was applied in many different areas of life course studies, such as the transition to adulthood (e.g. Schoon and Lyons-Amos, 2016), patterns of cohabitation (e.g. Di Giulio et al., 2019), fertility (e.g. Gemmill, 2019), work pathways (e.g. Struffolino et al., 2020), residential mobility (e.g. Coulter and Van Ham, 2013) or social isolation (e.g. Lay-Yee et al., 2021). Even if mainly used in the study of life courses, sequence analysis was used in other applications such as web log analysis (Gitinabard et al., 2019), building histories (Bradley, 2019), or fish movement histories (e.g. Lowe et al., 2020). Several reviews of the historical developments of sequence analysis and the available methodological tools have been published (Piccarreta and Studer, 2019; Ritschard and Studer, 2018b; Liao et al., 2022).

Hereafter, we first detail the classical sequence analysis. Then, we summarise alternatives

to the classical sequence analysis and ad hoc method to treat missing data. Finally, we develop the joint study of multiple sequences.

### Classical sequence analysis

According to Abbott and Tsay (2000), three steps compose a typical sequence analysis:

1. The data are coded as sequences.
2. A dissimilarity measure is chosen, and the pairwise sequence dissimilarities are computed.
3. The sequences are analysed based on their dissimilarities.

The coding phase could already raise some questions. For example, should the persons working at home or unemployed be described with different categories or gathered?

Introduced to social sciences by Abbott and Forrest (1986), optimal matching is commonly used to measure the pairwise dissimilarity between sequences. In this framework, the effort necessary to change one sequence into another is determined through the insertion, deletion, and substitution of states. Costs that can be state-dependent are defined for the operations of substitution and insertion-deletion. This offers a great flexibility for optimal matching. The range of dissimilarity measures within this framework extends from Levenstein II distance, which is optimal matching without substitutions, to generalised Hamming distance, which is optimal matching without insertion and deletion (Lesnard, 2010). Apart from optimal matching, many other dissimilarity measures have been provided. A large part of them are variations of optimal matching, such as dynamic Hamming distance, which is Hamming distance with time-dependent substitution costs (Lesnard, 2010), or optimal matching where insertion-deletion and substitution costs depend on the spell length (Halpin, 2010). However, dissimilarity measures are not restricted to variations of optimal matching. For example, Elzinga (2005) based the dissimilarity measure on the number of matching subsequences and Deville and Saporta (1983) on the distance, either Euclidean or $\chi^2$, between state distributions. Studer and Ritschard (2016) conducted a comprehensive review of sequence dissimilarity measures, offering guidelines for selecting an appropriate measure depending on the aspect of the sequences of interest, such as timing, duration, and sequencing of the states.

Regarding the analysis of sequences itself, a clustering is usually applied (Ritschard and Studer, 2018b) to identify groups in the data. The clustering can be the primary goal of the analysis in order to determine the main patterns among sequences, which was done, for example, by Jalovaara and Fasang (2017) to identify typical union trajectories that lead to childlessness or by Lorentzen et al. (2019) to study the transition from school to work in Finland, Norway, and Sweden. On the other hand, groups defined by clustering may be used in a regression analysis; either as the dependent variable, such as in Levy et al. (2006), where typical men's

and women's life course are compared, or as the independent variable, such as in Devillanova et al. (2019), where a clustering of employment trajectories is used as an independent variable in a regression about self-reported health in middle life. Clustering methods are typically unsupervised, which means that the number of groups is not known beforehand.

Two clustering algorithms are generally used in this context: partitioning around medoids (Kaufman and Rousseeuw, 1990) and hierarchical clustering. A predefined number of groups $k$ is specified for partitioning around medoids. It alternates between two phases; one that looks, for each group, for the medoid, namely the observation that minimises the distance from each group member; another that assigns each observation to its closest medoid. The process applies until no improvement is possible. With hierarchical clustering, no predefined number of groups is specified. It starts with each observation as a different group and recursively fuses the two closest groups until only one remains. The optimal number of groups is chosen afterwards. Several criteria, generally called "cluster quality index", are at hand to determine the best clustering among several possibilities (see e.g. Studer (2013) for a review and Arbelaitz et al. (2013) for a comparison of their behaviour in a general setting). However, the raw value of these cluster quality indexes is generally not informative. Studer (2019) introduced a framework based on parametric bootstrap to overcome this limitation.

Alternatives to the standard unsupervised clustering methods exist. First, model-based clustering even gets rid of the dissimilarity measure. The idea is that each observation belongs to a latent class and that a different probability distribution generates each class. The models commonly used are the independence model (Han et al., 2017), first-order Markov model (Melnykov et al., 2016), and hidden Markov model (Helske and Helske, 2017). Then, one can quantify the degree to which a sequence belongs to each cluster instead of assigning it to a unique cluster, called "fuzzy clustering" (Studer, 2018). Finally, to make the criteria used to determine the clusters more explicit, Studer (2018) introduced a property-based clustering, where properties that quantify timing, duration, and sequencing are used as separation rules in hierarchical clustering.

### Other approaches

Sequence analysis is not restricted to the classical three-step procedure, even if it is the most applied procedure. As mentioned, a key goal of sequence analysis is to extract the most salient characteristics of the sequences. Therefore, a large body of research develops visualisation and descriptive tools, which will be the subject of the next paragraphs. Another axis of research develops the statistical aspects of sequence analysis.

Concerning the visualisation of sequences, both the *TraMineR* package (Gabadinho et al., 2011) for the statistical software R and the *SADI* package (Halpin, 2017) for STATA provide helpful tools. Among the most common visual representations, we have:

- The chronogram (Billari and Piccarreta, 2005), which shows the state distribution at each time.
- The index plot (Scherer, 2001) displays the first sequences of the dataset.
- The sequence frequency plot (Müller et al., 2008) provides the most common sequences, where the bar widths are proportional to the sequence frequencies.
- The modal states plot (Gabadinho et al., 2010) provides the most common state and its frequency for each time.

These four visual representations are illustrated in Figure 2.12 with the completed dataset obtained with the last observation carried forward.

Moreover, several indicators were developed to describe sequences and their complexity. Among them, turbulence (Elzinga and Liefbroer, 2007), longitudinal entropy (Widmer and Ritschard, 2009), and complexity index (Gabadinho et al., 2010) quantify the variability of a sequence. These measures do not take into account the nature of the state. Therefore, several attempts were made along this line. Manzoni and Mooi-Reci (2018) introduced a measure of quality for binary sequences that can be coded as "success" and "failure", while Ritschard et al. (2018) modified the complexity index, which they called the "precarity index" to consider the type of transition (positive or negative) with partially ordered states. Ritschard (2021) reviewed indicators used to characterise sequences, studied their behaviour, and provided several extensions to overcome some limitations.

Another body of research is dedicated to expanding and strengthening the statistical aspects of sequence analysis. The interest is often to understand if the trajectories differ according to specific covariates, such as gender or social class. Several standard statistical tools were adapted to this aim. First, Studer et al. (2011) adapted the analysis of variance (ANOVA) framework for this purpose and, more particularly, the coefficient of determination R2, the F-statistic, and the Levene statistic. In the same article, the authors developed a tree-based procedure that sequentially splits the sample according to the covariates. Then, Liao and Fasang (2021) adapted Bayesian information criteria and likelihood ratio tests. An ongoing area of research is the application of sequence analysis to large samples (Liao et al., 2022), and note that with multiple imputation, the computational issues related to large samples is multiplied by the number of imputations considered.

### Ad hoc treatment for missing data

Some treatments for missing data are specific to sequence analysis. One can consider a missing state as an additional state and compute a dissimilarity as it was simply another state. However,

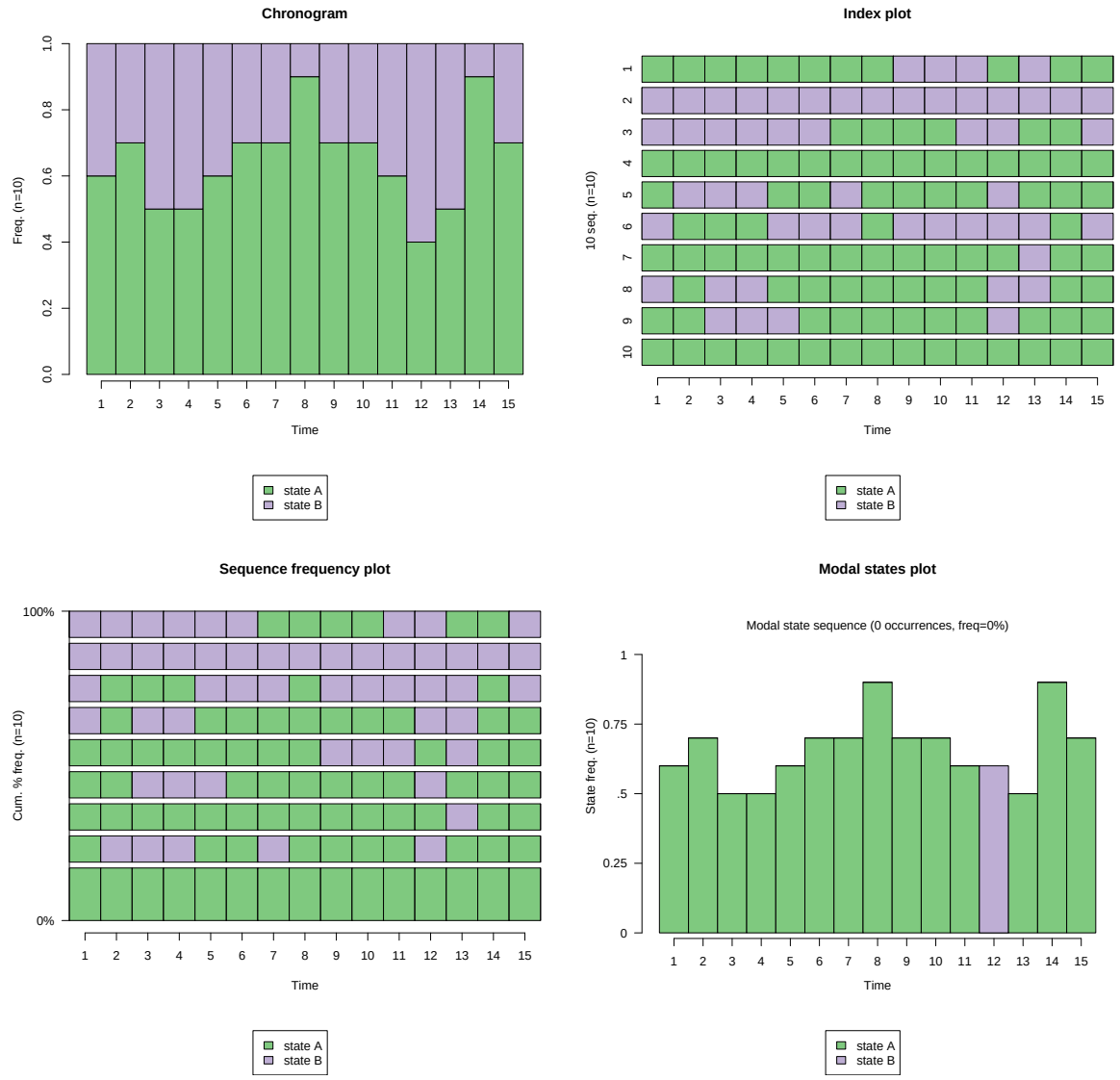



Figure 2.12: Visual representations of the complete dataset obtained with the last observation carried forward as a chronogram (top left), an index plot (top right), a sequence frequency plot (bottom left) and a modal states plot (bottom right).

this state could encompass heterogeneous situations, and a typology could be partially defined by missing data.

Halpin (2016a) proposed considering that a missing state has a maximum dissimilarity with any other state, including another missing state. The main drawback of this method is the more missing data a trajectory has, the more disconnected it is from other trajectories. In the extreme cases, we may have, in a clustering analysis, several clusters composed of only one trajectory. Anyway, this method cannot be completely discarded and further evaluation is needed on its performance. Therefore, this method will be evaluated as part of this thesis.

**Joint sequence analysis**

The focus is often on the simultaneous analysis of several trajectories. Two types of situations occur (Studer, 2015). On the one hand, these channels can be linked to the same domain. For example, Mattijssen and Pavlopoulos (2019) used income and labour market positions as two indicators of career trajectories, and Raab et al. (2018) characterised health with two channels, one related to physical health and the other to mental health. On the other hand, channels may relate to different domains. The idea is that resources, behaviours, and goals in one domain are linked with other domains' resources, behaviours, and goals (Bernardi et al., 2019). One of the most striking examples is that of work and family, where numerous studies have demonstrated, for example, the impact of the birth of children on women's occupational trajectories (e.g. Piccarreta and Billari, 2007; Widmer and Ritschard, 2009; Aisenbrey and Fasang, 2017). In a similar idea, the trajectories of individuals are influenced by those of others (and also influence theirs), called "linked lives" (Elder et al., 2003), such as parents and children (e.g. Fasang and Raab, 2014), couples (e.g. Möhring and Weiland, 2022) or siblings (e.g. Raab et al., 2014).

As with the standard sequence analysis, the contexts in which several sequences are considered simultaneously for each individual can be diverse. For example, Brum-Bastos et al. (2018) investigated the weather effects on human mobility, Roux et al. (2018) studied the patterns of care pathways, and Liu et al. (2022) studied time use and food-related geographic contexts.

The first simple idea to study these sequences is to apply the desired analysis to the domains separately and then combine the results. Along this line, Han and Moen (1999) and Widmer and Ritschard (2009) applied clustering separately to two domains, namely work and family trajectories. Then they studied the link between the two groupings. However, this is applicable only with a small number of domains. Moreover, it goes against the objective of the analysis, that is, to take their association into account. The other solution, joint sequence analysis, involves using a joint dissimilarity based on all domains (Piccarreta, 2017). The three main ways to do a joint analysis are:

- To compute the dissimilarities separately by domains and to combine them, usually by summing or averaging. The main drawback of this approach is that, like the combination of the domain-specific results, it does not consider the interplay between the different domains.

- To combine the pool of potential states, called the "alphabet", of each domain, giving a single channel with an extended alphabet.

- To extend optimal matching to multichannel sequences. This strategy is called "multichannel" or "multiple" sequence analysis (Pollock, 2007; Gauthier et al., 2010). Concretely, the substitution cost needed to align two multichannel sequences at a given time

point is defined as the mean, possibly weighted, of the substitution costs needed to align each channel. In the same way, the insertion-deletion costs of each channel are averaged.

The extended alphabet and multichannel sequence analysis are combinable: some channels could be first aggregated before multichannel sequence analysis. Another strategy, called "globally interdependent multiple sequence analysis" (Robette et al., 2015), combines optimal matching, multidimensional scaling, canonical partial least square, and clustering. The authors applied their method to study the patterns of transmission between generations. Several authors criticised this strategy. Fasang (2015) and Gauthier (2015) questioned its added value compared to multichannel sequence analysis, and Elzinga (2015) and Studer (2015) its use to assess the association between channels.

# Chapter 3

# Multiple imputation in longitudinal datasets

## 3.1 Introduction

In this chapter, we focus on the issue of missing data in longitudinal surveys. Specifically, we present a procedure, based on multiple imputation, to treat missing data in longitudinal surveys, paying particular attention to the challenges posed by logical missing data and categorical variables. The existing methods and guidelines are insufficient when it comes to these two specific issues. To demonstrate the effectiveness of our procedure, we provide an illustration using a real dataset obtained from colleagues who sought to address missing data concerns. This not only showcases the suitability of our approach but also emphasises the challenges it presents.

The existing approaches and guidelines fail to deliver complete satisfaction. Due to logical missing, standard multiple imputation procedures such as fully conditional specification (FCS) (Van Buuren, 2007) may create inconsistency in the imputed values and distort the association between the variables. If the variable that triggers the logical missing is not missing, we can specify beforehand the values that should be imputed. For instance, if age is not missing, only individuals who have reached the legal voting age would have variables related to voting imputed. On the other hand, when the variable that induces potential logical missingness is itself missing, as is the case with unit-level missing data, the situation is more challenging to handle. Indeed, the variable subject to potential logical missing will be imputed depending on the answer to another variable. Standard imputation procedures, such as FCS, cannot handle such situations. Therefore, applying these imputation algorithms could lead to impossible combinations of variables. The simplest way to tackle this issue would be to apply the imputation algorithm without taking into account logical missing and correct afterwards the values corresponding to a wrongly imputed logical missing. However, the imputation process uses these

nonsensical values. For example, suppose we have a question about a potential illness, injury, or health issue during the last year and, in the affirmative, another question asking if the individual still suffers from it. Suppose both these variables and self-rated health are missing for some individuals. If someone still suffers from the health issue, it will likely negatively impact his self-rated health, while if he does not suffer from it any more, it probably will not be the case. Therefore, still suffering or not from an illness or injury is likely a strong predictor of self-rated health during the imputation process. Imagine that it was imputed that someone did not have any potential illness during the last year but, on the other hand, still suffered from an illness at the time of the interview. By correcting the nonsensical imputed value for this variable afterwards, we change values used to predict the self-rated health and, hence, at least distort the relationship between the two variables.

Previous studies have focused on the multiple imputation of longitudinal datasets, but did not fully address the challenges related to categorical data and logical missing values. Among recent studies, Spiess et al. (2021) discussed the application of multilevel imputation models for panel data, with the idea that the individuals represent a cluster. Young and Johnson (2015) mostly focused on the impact of multiple imputation on the subsequent statistical analysis. Therefore, neither study focused specifically on the challenges induced by the specification of imputation models within longitudinal studies. On the other hand, Aßmann et al. (2017) discussed the issue of logical missing and proposed to handle it through classification and regression trees. However, they did not consider the challenges associated with categorical variables, particularly in relation to logical missing values.

In this study, we set ourselves in a broad scope of the imputation model, which allows multiple projects or analyses to use the same imputed dataset (Van Buuren, 2018). This scenario commonly arises when different researchers or groups handle the imputation and the analysis. In such cases, the specific form of the analysis is unknown during the imputation stage. The broad scope differs from the intermediate scope, where a single project estimates several similar quantities using multiple imputed datasets, and the narrow scope, which involves creating a separate imputed dataset for each analysis.

We demonstrate the process with a subset of the LIVES-FORS Cohort Survey. This sample is especially useful in illustrating the imputation process since it exhibits two main challenges we aim to address: categorical data and logical missing. The objective of using multiple imputation on this dataset is twofold. Firstly, it aims to prevent a reduction in statistical power. Since the subsample size is already small (849), retaining only individuals who did not miss any wave would result in a sample of 506, severely limiting the ability to detect statistical differences. Secondly, we may expect that missing data are not MCAR, and, hence, additional bias may be induced by simply deleting missing data. As is the case in most applications, we assume that the missing data are MAR. We discuss the implications of an MNAR mechanism in the

discussion section.

The remainder of this chapter is organised as follows. The methodology applied in this study is described in Section 3.2 and the sample used to illustrate the process in Section 3.3. Section 3.4 contains the results. A discussion ends the article.

## 3.2 Methods

We begin by outlining our systematic approach to address missing data, which comprises two distinct parts. In the initial part, we conduct a comprehensive analysis of the missing data within the sample. This step serves a dual purpose: firstly, to acquire an initial understanding of the missing data and their underlying patterns, and secondly, to identify variables that contribute to the mechanism of missingness. This information is crucial for incorporating these variables into the subsequent imputation process. The second part of the procedure focuses specifically on the imputation process itself.

### 3.2.1 Analysis of missing data

The initial phase of the missing data handling process focuses on analysing missing data, encompassing both unit-level and item-level cases. For unit-level missing data, our analysis involves comparing respondents and non-respondents to uncover characteristics linked to missingness. For item-level missing data, we focus on screening and examining the distribution of missing values. Additionally, we explore the interplay between item-level missing data and unit-level missing data. In the rest of this subsection, we provide a detailed account of how we investigate these three aspects more precisely.

Regarding unit-level missing data, respondents and non-respondents should be compared. The main goal is to identify characteristics related to missing data. We suggest building a logistic model for each pair of waves using several characteristics from the first wave as independent variables and the participation to the second wave as the dependent variable. Previous studies suggest that non-respondents tend to be male, younger, less educated, foreign, and unmarried (see e.g. Loosveldt and Carton (2001); Voorpostel (2010)). Additionally, individuals in vulnerable situations, such as those who are unemployed, have poor health, or lack social involvement, are more likely to be non-respondents (Rothenbühler and Voorpostel, 2016). Therefore, these variables should be considered in the logistic models. Any significant variables should be added later to the imputation model to justify the MAR hypothesis.

Regarding item-level missing data, one should first screen which values are really missing. First of all, logical missings are excluded. Then, special consideration is given to values such as "does not know". It is essential to discern whether such values should be considered as missing data. This ambiguity arises from the possibility that a person may genuinely not know the

answer to a question, or they may choose not to respond, which qualifies as a case of missing data. Consequently, a thorough examination of item-level missing data should be conducted on an individual and variable basis to identify problematic cases and variables with a high prevalence of missing data.

Finally, the link between item-level and unit-level missing data is under question. Item-level missing data may impact unit-level missing data and vice versa. Indeed, the number of item-level missing data and non-responses to difficult questions, such as income or health, could result in missing data in subsequent waves (Loosveldt and Billiet, 2002). Therefore, t-tests could be conducted between the number of item-level missing data and participation in the next wave for each pair of waves, and a chi-squared test could be performed between non-response to a difficult question and participation in the next wave. Additionally, individuals who were missing in a given wave may be more likely to have item-level missing data if they participated to the next wave. In this case, t-tests could be done between the number of item-level missing data and participation in the previous wave for each pair of waves.

### 3.2.2 Treatment of missing data

We present the multiple imputation process for addressing missing data, comprising two distinct steps: imputing item-level missing data and subsequently addressing unit-level missing data. This two-step approach aims to reduce the complexity by minimising the number of cases to be addressed. Indeed, combining the imputation of item-level and unit-level missing data simultaneously would induce numerous distinct patterns, complicating the process. To illustrate, let's consider an imputation model based on five variables, each of which is subject to item-level missing data. With 32 potential patterns of missing data among these five variables, it would require fitting 32 different imputation models.

Before focusing on the imputation process, we introduce a sequence of questions that we have developed. These questions are applied to both item-level and unit-level missing data, serving three main goals: identifying logical missing data, avoiding unnecessary imputation when certainty exists, and screening values that should not be imputed. We first present this sequence of questions. Then, we delve into the imputation of item-level missing data. Finally, we explore the imputation of unit-level missing data.

The sequence of questions, illustrated in Figure 3.1, begins by determining whether the missing value is due to a logical skip, meaning a question not asked to an individual based on its response to another question, or if it is indeed missing. If it is a logical skip, we differentiate between questions not asked because they were irrelevant and those for which the answer was already known. For the latter, we differentiate between questions that were not asked because the question did not make sense to the individual from the ones whose answer was known. On the other hand, when the information is truly missing, we investigate if the missing value is

retrievable anywhere in the dataset or from another source. If the value cannot be retrieved, we determine if the value is directly deducible from the values of other variables. If the value is not deductible, we must ask ourselves if the value should be imputed at all.

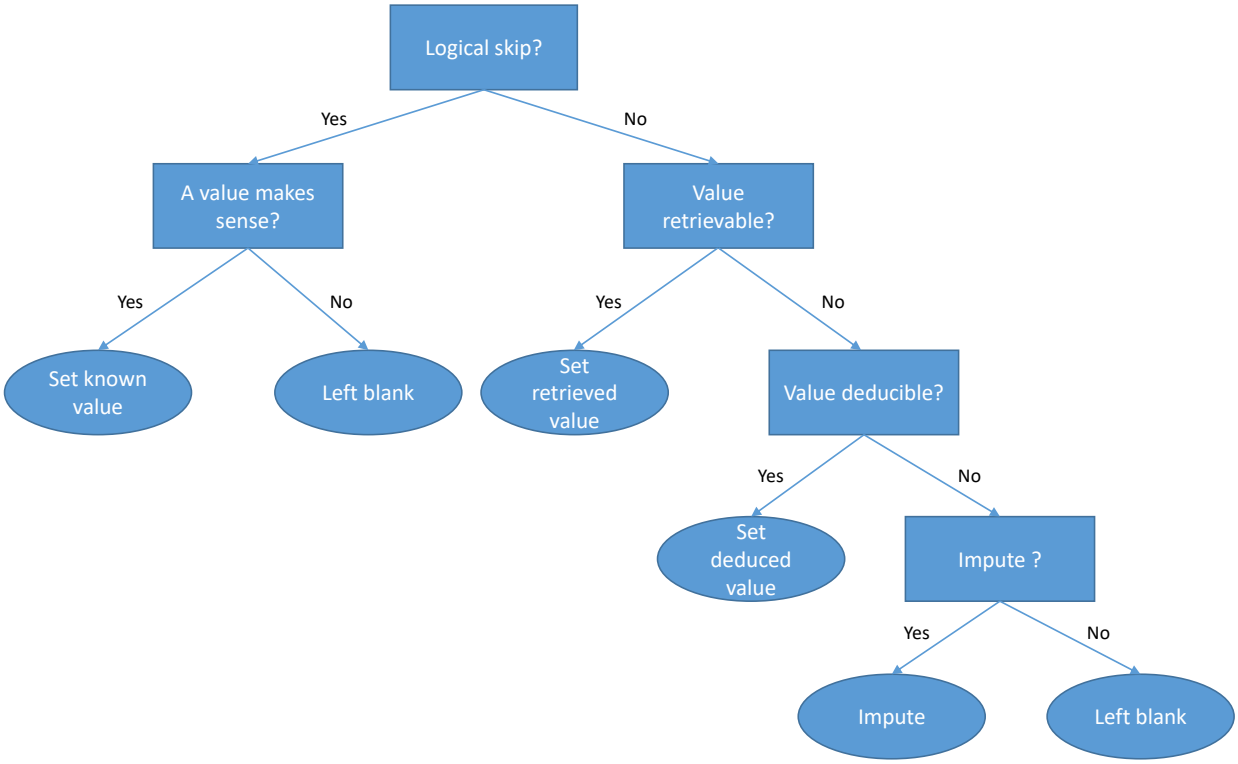



Figure 3.1: Questioning prior to a multiple imputation.

We will now describe the imputation process itself. We first focus on the treatment of item-level missing values before moving to unit-level missing data.

**Item-level missing data** We propose the following framework, which may serve as a guideline (each of these steps is developed below). This algorithm is designed to impute missing values in a logical and efficient way, using a combination of statistical tests and regression analysis to identify the most relevant predictors and impute missing values based on those predictors. The process takes the following form:

1. The variables should be ordered.

2. The algorithm iterates through each wave of data, starting with the first wave and proceeding to the last wave.

3. Within each wave, the algorithm goes through each variable and applies the following steps:

   (a) The sequence of questions (Figure 3.1) is applied to the variable.

   (b) The pool of potential predictors is identified and, if needed, recoded.

   (c) A test of association is done between the variable to impute and all sensible variables, including already imputed variables. A Fisher's exact test is applied when two categorical variables are involved, a correlation test with two numerical variables and an ANOVA with a categorical and a numerical variable. Significant variables are included in step (d).

   (d) A stepwise bidirectional regression is applied to explain the variable to be imputed, using Akaike's Information Criterion (AIC) as a comparison criterion. We use a logistic regression for dichotomous variables, a multinomial regression for multinomial variables, and a linear regression for continuous variables.

   (e) A bootstrap sample is drawn, the regression model is fitted using this bootstrap sample.

   (f) Missing values are imputed. For a logistic model (multinomial or not), predictions are drawn based on the estimated probabilities; for a linear model, predictions are drawn, and a randomly drawn error is added to them. When a numerical variable can take only some specific values, such as integers, the obtained value is rounded to the closest possible value.

   (g) If some missing values were not imputed at the last step, variables are removed from the model obtained at step (e) until every missing value is imputed.

   (h) The imputed values are checked.

We further detail and justify each of these steps. The variables are first ordered. A variable that may induce a logical missing in another variable must be imputed beforehand, just as a potential predictor of the variable's absence or presence.

The predictors are then selected. The pool of potential predictors can be considerable since the procedure is made to handle many predictors. The number of categories of a categorical variable should be reduced in order to avoid inflated variance. Moreover, some variable may need some recoding. For instance, categorical variables that have an inherent order may be recoded as numerical variables. A common example of this is a Likert scale question, where respondents provide their opinion on a scale ranging from "strongly disagree" to "strongly agree". Transforming such variables into numerical ones can be justified by the presence of a gradation among the responses. However, such a recoding is fully justified only when the gap

between successive categories is always similar (see e.g. Wu and Leung (2017) for a discussion on the coding of Likert scales in analysis).

The initial selection of variables in step (c) serves to decrease the number of potential variables in the model, as models have limitations in handling a large number of parameters, particularly when dealing with categorical variables. Additionally, including too many explanatory variables can lead to overfitting the model to the training dataset, resulting in poor imputations. Along this line, a simulation realised by Noghrehchi et al. (2021) showed that adding too many variables can lead to bias. Moreover, with logical skips, a model with many predictors could be fitted on only a small subsample of the observations. Since after the first step, the number of variables is often too large to test every subsample of variables, a stepwise procedure is applied to identify the most appropriate model (step (d)).

Distinct models are employed depending on whether the variable being imputed is numerical or categorical. We consider a variable to be of the numerical type when it is possible to perform computations with the different values of the variable, and the result has a numerical meaning. Otherwise, the variable is considered as categorical. When a numerical variable is the imputation's target, we apply a linear model for the imputation, regardless of the distribution of the variable. When the variable is not normally distributed, this could be under question. However, linear models often work well also when the distribution is not normal (Demirtas et al., 2008), and transforming a variable towards normality often leads to more biased results (Von Hippel, 2013). Another issue is linked to discrete variables, for example obtained with a Likert scale, or more globally to variables that only have a specific range of possible values. For example, if we have a variable that takes integer values between 0 and 10, an imputed value might not be an integer or even outside the range (such as negative values). Therefore, it is under question whether the imputed value should be rounded. Wu et al. (2015) have shown that not rounding performs well and that rounding afterwards could lead to biased results. However, the drawback of not rounding is that it could lead to imputing an impossible value, such as a negative percentage of work. We believe imputing values that do not make sense to a person is worse than potentially inducing some bias.

According to simulations run by Van Buuren (2018), bootstrapping and adding a random error (steps (e) and (f)) is a correct way to draw values since it leads to appropriate coverage. Failing to include bootstrapping or random error would result in underestimating the variance, which could lead to overly significant results and potentially impact the validity of statistical inferences.

Since, generally, missing data are not MCAR, we expect the distribution of the imputed values to be different from that of the observed values. However, as advanced by Van Buuren (2007), for example, substantial differences may signify that something went wrong. Therefore, it is advisable to compare the distribution of the imputed and the observed values and, if they

are clearly different, to look at the imputation model and the distribution of the predictors separately between imputed and observed values to explain such differences.

**Unit-level missing data** After treating the item-level missing data, unit-level missing data are considered. The following algorithm applies:

1. Identify the typology of missing gaps.
2. Impute middle gaps by recursively working from gaps of maximum length to those of length one, imputing the variables in the order determined during item-level missing data imputation by applying steps (a) to (h).
3. Impute initial gaps of missing data by recursively working from gaps of maximum length to those of length one, imputing the variables in the order determined during item-level missing data imputation by applying steps (a) to (h)
4. Impute end gaps of missing data by recursively working from gaps of maximum length to those of length one, imputing the variables in the order determined during item-level missing data imputation by applying steps (a) to (h).

The imputation process is divided into initial, middle, and end gaps because the information available for imputation differs depending on the gap type. For initial gaps, information from one or several following waves is available; for end gaps, information from one or several previous waves is available; and for middle gaps, information from both previous and following waves is available.

The first step is to identify the typology of missing gaps. As an illustrative example, we consider a situation with five waves. The typology depends on the number of previous and subsequent waves used in the imputation model. We consider that only information from the current and one previous and subsequent waves, when available, are used. Table 3.1 illustrates the typology of missing gaps, which includes, for middle gaps, one gap of length three, two of length two, and three of length one.

For middle gaps, the imputation process starts with the gap of length 3, where the first missing wave (wave 2) is imputed. As with the imputation of item-level missing data, the variables are imputed sequentially based on the chosen order. Once this wave has been imputed, gap (i) becomes like (ii). Then, the second missing wave of gaps of length 2 is imputed to realise a recursive imputation from the edges, ensuring longitudinal consistency. Gaps of length one are finally imputed. We considered the case of the use of only one wave before- and after-hand during the imputation process. By increasing the number of waves used, the typology becomes more detailed. For example, using two previous waves instead of one, the last type of middle gaps would be separated in two cases, depending on whether wave 2 is observed.

| Type of gap | | Waves | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 |
| Middle gap | i) | o | m | m | m | o |
| | ii) | | o | m | m | o |
| | iii) | o | m | m | o | |
| | iv) | o | m | o | | |
| | v) | | o | m | o | |
| | vi) | | | o | m | o |
| Initial gap | i) | m | m | m | m | o |
| | ii) | m | m | m | o | |
| | iii) | m | m | o | | |
| | iv) | m | o | | | |
| End gap | i) | o | m | m | m | m |
| | ii) | | o | m | m | m |
| | iii) | | | o | m | m |
| | iv) | | | | o | m |

Table 3.1: Typology of unit-level missing data. An observed wave is represented by "o", a missing wave is represented by "m" and a blank value means that the wave can either be observed or missing.

In a multiple imputation context, the whole process is repeated to produce several completed datasets.

## 3.3 Data

We use a subsample of the LIVES-FORS Cohort survey (LCS) that we received from colleagues to illustrate the imputation process. LCS is a yearly longitudinal survey focusing on young adults who grew up in Switzerland (Spini et al., 2019). The survey complements the Swiss household panel (SHP), an annual panel study launched in 1999 to observe social changes in Switzerland (Voorpostel et al., 2016). LCS almost shares the same questions and modules as SHP but differs in the target population and sampling procedure.

LCS comprises young adults whose parents have grown up in Switzerland and individuals called "secondos" whose parents arrived in Switzerland as adults. The main goal of LCS is to compare the transition to adulthood of these two populations. In the SHP, the secondos are too

few to perform statistical comparisons, hence the need for a specialised supplementary sample. The reference population of LCS is individuals born between 1988 and 1997, living in Switzerland on the 1st of January 2013, and schooled in Switzerland before age ten. Secondos, whose parents were born abroad and arrived in Switzerland after the age of 18, are overrepresented.

The SHP relied on the Swiss population register to create a probability-based sample. However, since the origin of parents is not available in this register, secondos were not directly identifiable (Herzing et al., 2019). Therefore, the LCS sampling scheme is a combination of stratified random sampling, screening and controlled network sampling. More precisely, 4000 individuals born between 1988 and 1997 were first randomly sampled. Individuals living in a region with a high share of residents born abroad or of individuals who come from or are born in countries from which the previous generation of workers tend to come from (e.g. Italy, Portugal, Spain or Serbia) were more likely to be selected. Then, they were contacted to determine the place of origin of their parents and whether or not they were schooled in Switzerland before the age of 10. Individuals that were not schooled in Switzerland before the age of 10 were discarded. Next, another sampling took place, oversampling individuals whose both parents arrived in Switzerland after the age of 18. At this stage, the sample was of 890 individuals. Finally, a network sampling took place. Individuals were asked to provide a list of their regular contacts and information about the origin of their parents. Another sample of individuals was randomly drawn from the regular contacts, with a higher probability for individuals whose parents arrived in Switzerland after the age of 18. A total of 1961 individuals participated in the study.

The LCS started in 2013 and has taken place yearly since then. The first wave consisted of a household questionnaire, which covers several areas linked to the household: the accommodation, the standard of living, the family, and the household's financial situation (Tillmann et al., 2016), and a life history calendar to collect retrospective information about living arrangements, family events, residence, couples relationships, professional and health trajectories. Starting from the second wave, in addition to the household questionnaire, individuals were asked about a variety of topics through traditional questionnaires: social origin, education, employment, health, politics and values, household and family, health and quality of life, leisure and media, and integration and networks.

The sample we received from our colleagues is composed of 849 individuals. It included the first five waves. Concerning the variables, in addition to sociodemographic characteristics, variables about education and work were kept. Appendix D details the complete list of variables.

The whole imputation process was repeated 20 times in order to produce a sufficient number of imputed datasets. The type I error was set to 5% for all analyses, and the R open source statistical environment was used for all computations (R Core Team, 2021).

## 3.4 Results

In this section, we illustrate the treatment process we have devised for handling missing data by applying it specifically to the LCS subsample. Following the structure outlined in the methods section, we proceeded with two steps. Firstly, we analyse the patterns of missing data in the subsample, considering both unit-level and item-level missing data. Subsequently, we demonstrate the sequence of questions and present the results of imputing selective variables. This serves the purpose of illustrating the imputation process while highlighting the challenges specific to our approach.

### 3.4.1 Analysis of missing data

We present the analysis of the patterns of missing data in the LCS subsample. We first focused on unit-level missing data. The articulation of missing data was first studied. Then, differences between respondents and non-respondents were tested. Afterwards, our investigation focused on determining if specific question responses might result in non-response during the subsequent wave. Finally, the impact of item-level missing data on unit-level missing data was investigated. Regarding item-level missing data, we first studied generally the patterns of missing data. Then, we screened the variables that displayed a higher tendency for item-level missing data and determined the underlying reasons for this behaviour.

#### Unit-level missing data

Since the first wave of the survey only consists of basic information that also appears in subsequent waves, such as age, gender, and achieved level of education, the first wave will not appear in the analysis of missing data or the imputation process. Waves 4 and 5 were more prone to missingness, with approximately 25% of unit-level missing data, while waves 2 and 3 were only missing in 7 and 10% of the cases. The patterns of unit-level missing data are displayed in Figure 3.2. Among the 849 individuals, 506 answered all waves. Having waves 4 and 5 missing is the most common pattern of unit-level missing data, with 100 individuals sharing it, while only one has waves 2 and 4 missing. However, not all potential patterns are represented in the data. Nobody has only the second wave observed, and individuals with all the missing waves were not included in the dataset. Like in all longitudinal studies, missing data often appear consecutively in this dataset. Indeed, chi-squared independence tests of unit-level missing data between pairs of consecutive waves were all significant, with the strength of the link increasing over time.

Then, we determined whether some socio-demographic groups were more prone to unit-level missing data. We considered the sex, the age, being a secondo or not, the highest level of education achieved and the marital status. On one hand, once the age, the sex and being

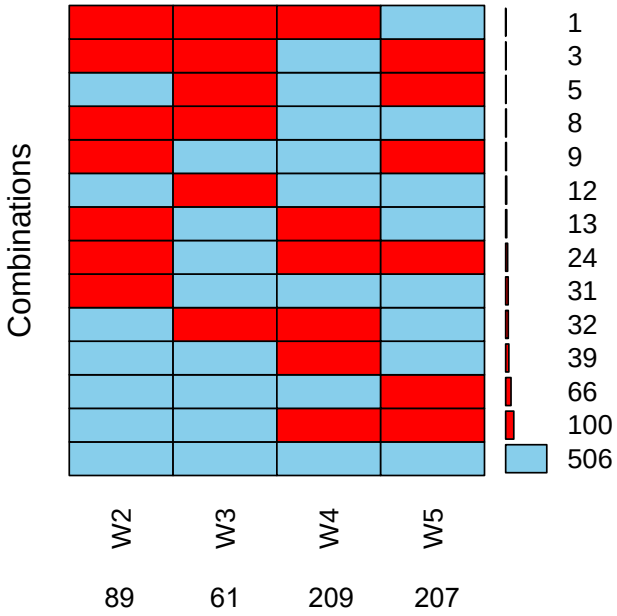



Figure 3.2: Patterns of unit-level missing data with their respective number. An observed wave is displayed in blue, while a missing wave is in red. In addition, the number of unit-level missing data per wave is displayed.

a secondo are observed, they are known for every other waves. On the other hand, when the marital status or the highest level of education achieved is missing, there could be uncertainty about the value. In this case, we used the value observed in another wave as a proxy. For example, if the wave 3 was missing for an individual, the marital status information from wave 2 was used instead. If wave 2 was also missing, the marital status information from wave 1 was used. In the event that data for waves 1 through 3 were missing, the marital status information from wave 4 was used instead, or the information from wave 5 if wave 4 was also missing.

Table 3.2 summarises the p-values of independence tests between a socio-demographic characteristic and the fact of having responded to a given wave. The age differs significantly between respondents and non-respondents in waves 4 and 5, while the achieved education level differs in waves 2 and 5. Concerning sex, being a secondo, and marital status, there are no significant differences between respondents and non-respondents. Boxplots for the age, separated between respondents and non-respondents, are displayed in Figure 3.3. The difference between the mean age of the respondents and the non-respondents is tiny, while, for wave 5, the median age of the non-respondents is slightly higher than the age of the respondents (25 vs. 24). Moreover, in waves 4 and 5, the variance is smaller across the non-respondents. Regarding education, individuals who had reached an upper secondary level of education are more prone to missing the second wave. Moreover, it is the case for those with a compulsory school level of education for the fourth wave.

| | W2 | W3 | W4 | W5 |
|---|---|---|---|---|
| age | 0.080 | 0.742 | 0.007 | $< 0.001$ |
| sex | 0.487 | 0.388 | 0.482 | 0.478 |
| education | 0.006 | 0.675 | 0.096 | 0.009 |
| secondos | 0.227 | 0.476 | 0.886 | 0.276 |
| marital status | 0.207 | 0.804 | 0.610 | 0.256 |

Table 3.2: P-values of the tests of independence between socio-demographic variables (age, sex, to be a secondo, and highest level of education achieved) and missingness in a wave (2 to 5).

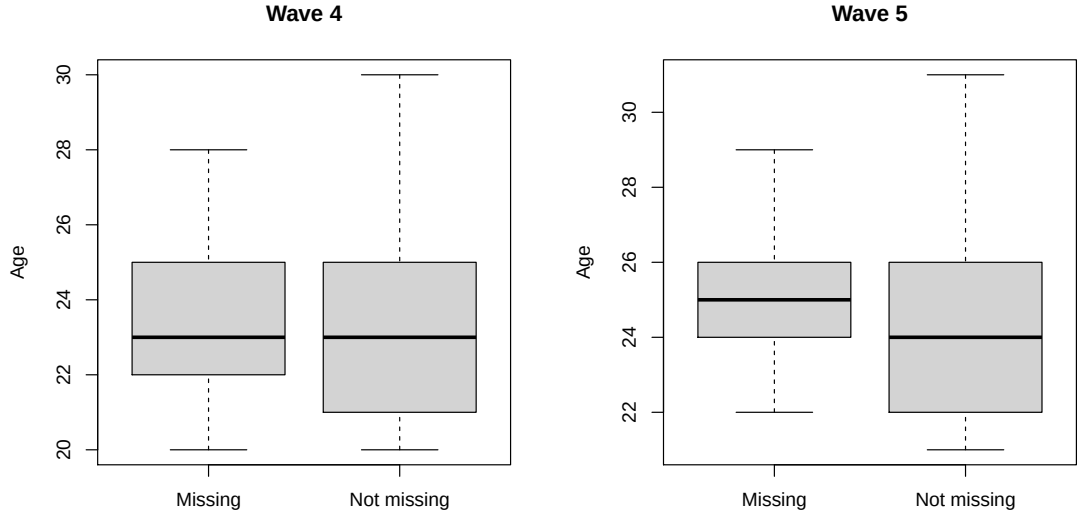



Figure 3.3: Boxplots of the age by response status at waves 4 and 5.

Next, we examined whether certain question responses could lead to non-response in the following wave. Consistent with prior research, we analysed variables related to employment, self-rated health, and social involvement, specifically "participation in clubs", "trust in people", and "political interest". Logistic models were fitted for each pair of successive waves, where the dependent variable is unit-level missing data of the second wave of the pair, giving three models. Although we anticipated no differences among these three models, we separated them to satisfy the independence assumption of logistic regression (Stoltzfus, 2011). Indeed, failing to do so would induce, for example, that individuals not missing any wave, would be included thrice.

Table 3.3 shows the results of the three regression models. There were no variable significant in all three models. In line with the tests realised above, age was significant in predicting miss-

ingness in waves 4 and 5, while gender, to be married, and to be a secondo, has no significant impact on the missing process. The effect of the highest level of education on the missingness status of wave 5 differs significantly between the levels of compulsory school and upper secondary. Even if not significant, the tertiary education level has a coefficient almost equal to the upper secondary level. Concerning social involvement, only the responses to the variables "participation in a club" and "trust in people" in wave 3 significantly impact the missing status of the subsequent wave. Higher confidence in people and participation in clubs decreased the probability of wave 4 being missing. Interest in politics and self-rated health have no significant relationship with the missing mechanism. These results must be carefully handled because only a few variables are significant, and there is almost no consistency across the waves. Therefore, significant results could be induced by randomness. Moreover, the overall performance of these models is very low since the Nagelkerke $R^2$ was between 0.06 and 0.08.

Finally, we considered either the number of item-level missing data at the previous wave or whether an individual had item-level missing data to predict complete nonresponse in the next wave, but the tests were never significant. Regarding potential sensible questions such as income, self-rated health, or whether the interviewee has suffered from a severe health issue during the last year, they have all very few missing data (less than two per wave). Therefore, we could not establish a link between non-response to theoretically sensible questions and a subsequent missing wave.

### Item-level missing data

In the second step, we focused specifically on item-level missing data. Approximately two thirds of the respondents in each wave have no item-level missing data, excluding logical missing. Only five observations, one in the second wave and four in the third wave, have more than ten missing values, and among them, two have a more significant number of item-level missing data (26 and 35) than all other respondents. These large numbers are difficult to explain since both individuals' attitudes were considered "friendly and cooperative", according to the interviewer. Moreover, they answered every wave, and the number of item-level missing data was limited on the other waves. Therefore, this may be due to technical issues or a lack of time for the participants.

Some variables are more prone to missing data than others. Table 3.4 displays the percentages of item-level missing data for the five most missing variables among the ones issued from the individual questionnaire. By looking into it, the situation is different depending on the variable. Indeed, most of the occupation missing values appear because no code from the Swiss-specific occupation codes corresponded to the respondent's answer. Alternatively, most of the missing for the company's restructuring is due to "does not know" modalities. For the percentage of part-time work and number of employees in a company, the missing are shared

| | W2–W3 | | W3–W4 | | W4–W5 | |
|---|---|---|---|---|---|---|
| | OR | p-value | OR | p-value | OR | p-value |
| age | 0.99 | 0.936 | 1.17 | 0.001 | 1.18 | 0.008 |
| gender (ref. male) | | | | | | |
| female | 1.00 | 0.993 | 1.07 | 0.698 | 1.07 | 0.796 |
| education (ref. compulsory school) | | | | | | |
| upper secondary | 0.53 | 0.095 | 0.88 | 0.664 | 0.42 | 0.032 |
| tertiary | 0.30 | 0.097 | 0.96 | 0.912 | 0.41 | 0.065 |
| married (ref. no) | | | | | | |
| yes | 1.80 | 0.392 | 0.87 | 0.718 | 0.51 | 0.204 |
| secondos (ref. no) | | | | | | |
| yes | 0.71 | 0.288 | 0.88 | 0.468 | 0.92 | 0.741 |
| participation club (ref. no) | | | | | | |
| yes | 0.73 | 0.392 | 0.64 | 0.032 | 0.78 | 0.357 |
| employment status (ref. active occupied) | | | | | | |
| unemployed | 1.10 | 0.876 | 1.03 | 0.923 | 0.74 | 0.589 |
| not in labor force | 0.32 | 0.021 | 0.73 | 0.172 | 0.48 | 0.052 |
| interest in politics | 0.91 | 0.096 | 0.94 | 0.087 | 0.93 | 0.078 |
| trust in persons | 1.05 | 0.431 | 0.88 | $< 0.001$ | 1.02 | 0.732 |
| self-rated health | 0.91 | 0.305 | 1.08 | 0.133 | 1.00 | 0.971 |
| N | 760 | | 788 | | 640 | |
| $R^2$ | 0.07 | | 0.08 | | 0.06 | |

Table 3.3: Logistic regressions predicting, for each pair of consecutive waves, unit-level missing data to the second wave (with non-missing as the reference category) of the pair using variables from the first wave of the pair. The number of observations used to estimate each model (N) and Nagelkerke $R^2$ are also displayed.

between "does not know" and lack of an answer.

Globally, no variable was significantly linked to the missing process.

### 3.4.2 Treatment of missing data

We demonstrate the application of the imputation process using the LCS subsample as an example. First, we illustrate the use of the sequence of questions with variables taken from the subsample. In most cases, it is not needed to apply this sequence of questions before the item-level imputation because most surveys have specific codes for logical missing. However,

| | W2 | W3 | W4 | W5 |
|---|---|---|---|---|
| restructuring of company | 5 | 5.7 | 4.2 | 6.2 |
| number of employees in company | 2.5 | 4.6 | 4.4 | 4.2 |
| percentage of part-time | 3 | 3.9 | 2.5 | 2.8 |
| occupation in main job | 0.7 | 1.9 | 3.9 | 2.8 |
| number of contractual hours | 0.9 | 1.4 | 1.3 | 2.5 |

Table 3.4: Percentages of item-level missing data per variable per wave for the five most missing variables on the whole dataset.

it is useful as an illustration since the sequence of questions will need to be applied during the unit-level imputation. We then detail the treatment of several variables during the unit-level imputation. As the imputation process for item-level missing data is similar to unit-level, we focus solely on examples of unit-level imputation. The objectives are two folds. First, illustrate the process of imputation. Then, highlight the difficulties linked to the application of our process. We focus on the case of an end gap of missing data with waves 4 and 5 missings and, more precisely, on the imputation of wave 4. Therefore, information on wave 3 is available. We have selected three variables corresponding to the work module's three layers:

1. The work status determines if the work module is to be imputed.

2. The type of employment represents basic information on the job.

3. The degree of interference of the work with the private life, for the different scales related to the job.

We illustrate each tree branch of the sequence of questions (Figure 3.1) with real cases issued from our sample. First, suppose that a value is lacking for the number of persons under supervision. We face a logical skip if the individual is not working or has no supervisory task. In contrast, we do not have a logical skip if the individual is working and has supervisory tasks. No value makes sense if we face a logical skip for this variable. An example of a logical skip for which a value makes sense would be the percentage of work, where a value of 100% can be set for the individuals working full-time. For the number of persons under supervision, a value not lacking due to a logical skip is neither retrievable elsewhere nor deducible from other variables. Moreover, there is no contraindication to the imputation of these missing values. We have spotted no cases of values that are lacking and retrievable elsewhere in the dataset. The work status is asked both in the questionnaire submitted to the individuals and the questionnaire answered by the household representative. Therefore, a work status lacking in

the individual questionnaire could be retrieved in answers to the grid questionnaire. However, in practice, this case does not happen. Concerning values that are deductible, it is the case, for instance, with an ordinal variable such as the highest level of education achieved. If the values at the previous and following wave are identical, then the missing value can be replaced by this value. Finally, the exact job denomination is a salient example of a variable that should not be imputed. Indeed, since most of the missing information of this variable appears because there was no correspondence in the Swiss-specific occupation codes, imputing a value with our method would lead to wrong values with certainty. Moreover, even if the value is missing due to a non-response, imputing could lead to very unlikely jobs for a given person.

Then, we focus on the imputation of the three selected variables from the work module, namely the work status, the type of employment and the interference of work on private life.

**Imputation of the work status**

This variable determines whether the other variables of the work module should be imputed. Therefore, imputing a wrong value to this variable will impact all the other variables of the work module. The work status is coded in three states: active occupied, not in the labour force, and unemployed. Initially, this variable was constructed after the interview based on the answers to a sequence of six questions (e.g. *Did you get paid for working, even if only for one hour, last week, either as an employee, self-employed or an apprentice?*). However, since the level of detail induced by these questions is not necessary, we directly impute this constructed variable.

We start by applying the sequence of questions before considering the imputation itself.

1. *Logical skip* No value is the consequence of a logical skip.

2. *Value retrievable* No value is retrievable with certainty elsewhere in the dataset.

3. *Value deducible* When an individual is outside the workforce, a variable captures the reason why. One of the possible answers is *Permanently disabled and/or unfit for work*. Even if an individual could join the workforce in a subsequent year, it should be rare. Therefore, we impute a not in labour force value for subsequent missing waves.

4. *Should it be imputed* There is no contraindication to the imputation of this variable.

We split the imputation process for this variable into three parts, based on the work status in the previous wave. This division is necessary because the available variables vary between active individuals and the two other situations. We provide a brief description of the imputation method for the latter two cases and present detailed results for the active occupied scenario. For individuals who are not active occupied, all the variables of the work module are logical missing. On the other hand, some questions are tailored to these situations (e.g. the reason not to be

working). For each of the 20 imputation, the imputation model consists only of an intercept for the individuals that were not in the labour force. Considering the unemployed individuals, the imputation models contain the age, the reason for unemployment, if the individuals were in training, and were registered as unemployed in a regional job centre.

Concerning the active occupied individuals, the pool of potential predictors consists of sociodemographic variables and most of the variables of the work module at the previous wave. We have discarded some variables due to their level of detail (e.g. the exact denomination of the job) or the few individuals concerned (e.g. the duration of the limited contract, only for individuals with such a contract).

Moreover, most categorical variables needed some pre-processing due to some relatively rare categories. For example, the position at work is composed of four categories: production position, supervision/training position, management position, and other, which are composed of 355, 43, 10, and 15 individuals. Since the three rarest categories represent some specific situations other than being employed, we merge them.

As a reminder, the entire imputation process was repeated 20 times to generate multiple completed datasets. Therefore, we analysed the results of these 20 completed datasets collectively rather than examining each imputation individually. Since the pool of predictors is large (39 variables), we initially applied a selection step using association tests. After this step, the number of predictors was reduced to 11 or 12, depending on the imputation number. In some cases, the percentage of work was included, while in others it was not. The p-value of the bivariate test between the percentage of work and work status was found to be close to 0.05. Consequently, depending on the imputed values during item-level imputation, this p-value either slightly exceeded or fell below 0.05. Subsequently, we employed the forward-backward stepwise procedure with AIC as the selection criterion, resulting in the selection of 8 or 9 variables. Notably, the percentage of work was included in thirteen imputations but excluded in seven. While the item-level imputation and bootstrapping introduced some randomness among the imputation models, all imputation models share the same characteristics except for the percentage of work. The odds ratios and their respective p-values are presented in Table 3.5 for the first iteration. It can be observed that:

- Job limitations in time induces a higher probability of becoming unemployed or not in the labour force, the coefficient being significant only in the not in labour force case. It is expected that individuals without a contract limitation are more likely to stay active occupied.

- The higher the % worked, the more likely the person is to be active occupied compared to those not in the labour force. Since the dataset is composed of young adults, it is likely that a low percentage encompasses individuals with temporary work in parallel to their studies and that are more likely to drop it.

| | unemployed | | not in labor force | |
|---|---|---|---|---|
| | OR | p-value | OR | p-value |
| (intercept) | 7.50e-18 | $< 0.001$ | 0.31 | 0.447 |
| age | 1.00 | 0.996 | 0.78 | 0.027 |
| level of education | 0.33 | 0.046 | 0.89 | 0.665 |
| % of work | 0.99 | 0.324 | 0.98 | 0.005 |
| intensity of work | 1.35 | 0.051 | 0.89 | 0.102 |
| position in firm (ref. other) | | | | |
| production | 1.41e06 | $< 0.001$ | 8.41 | 0.053 |
| job limitation in time (ref. yes) | | | | |
| no | 0.39 | 0.207 | 0.23 | $< 0.001$ |
| use of a computer (ref. yes) | | | | |
| no | 0.14 | 0.021 | 1.59 | 0.233 |
| security of job (ref. secure) | | | | |
| not secure | 8.00 | 0.004 | 0.24 | 0.184 |
| job with supervision task (ref. yes) | | | | |
| no | 1.01e11 | $< 0.001$ | 0.57 | 0.237 |
| N | 422 | | | |
| $R^2$ | 0.36 | | | |

Table 3.5: Imputation model for the work status at the first imputation. The reference category is *active occupied*. The level of education is a scale from 1 to 5, the % of work goes from 0 to 100 and the intensity of work is a scale from 0 to 10. The number of observations used to fit the model (N) and Nagelkerke $R^2$ are also displayed.

- The older the person, the less likely to get out of the labour force. As advanced with the coefficient related to the % worked, younger individuals may be more prone to have temporary work in parallel to their studies, which may differ from one year to another, while older people are really into the labour market.

- Individuals who estimated their job was not secure during the previous interview are more likely to become unemployed.

- The more intense the work was, the more likely the individual would be unemployed rather than active occupied.

- The more educated, the less likely to become unemployed.

This model estimates the probability of belonging to the three categories for each missing individual. Subsequently, values are drawn based on these probabilities. The distributions of the observed and imputed values are displayed in Figure 3.4. Generally speaking, unless the process is MCAR, we may expect the distributions to be different. However, considerable differences may be a sign of a poor imputation model. Notably, in 19 out of the 20 imputations, the proportion of imputed active occupied individuals surpasses that of the observed active occupied individuals. Conversely, the proportion of unemployed individuals is lower in 18 out of the 20 imputations. Individuals with a missing working status were more often in "other" positions in the firm (24% versus 16%) and "supervisory" tasks (38% versus 30%) during wave 4 than the individuals observed. Both scenarios suggest a high probability of being in an active occupied state, thereby elucidating the dissimilarities between the observed and imputed working status distributions.

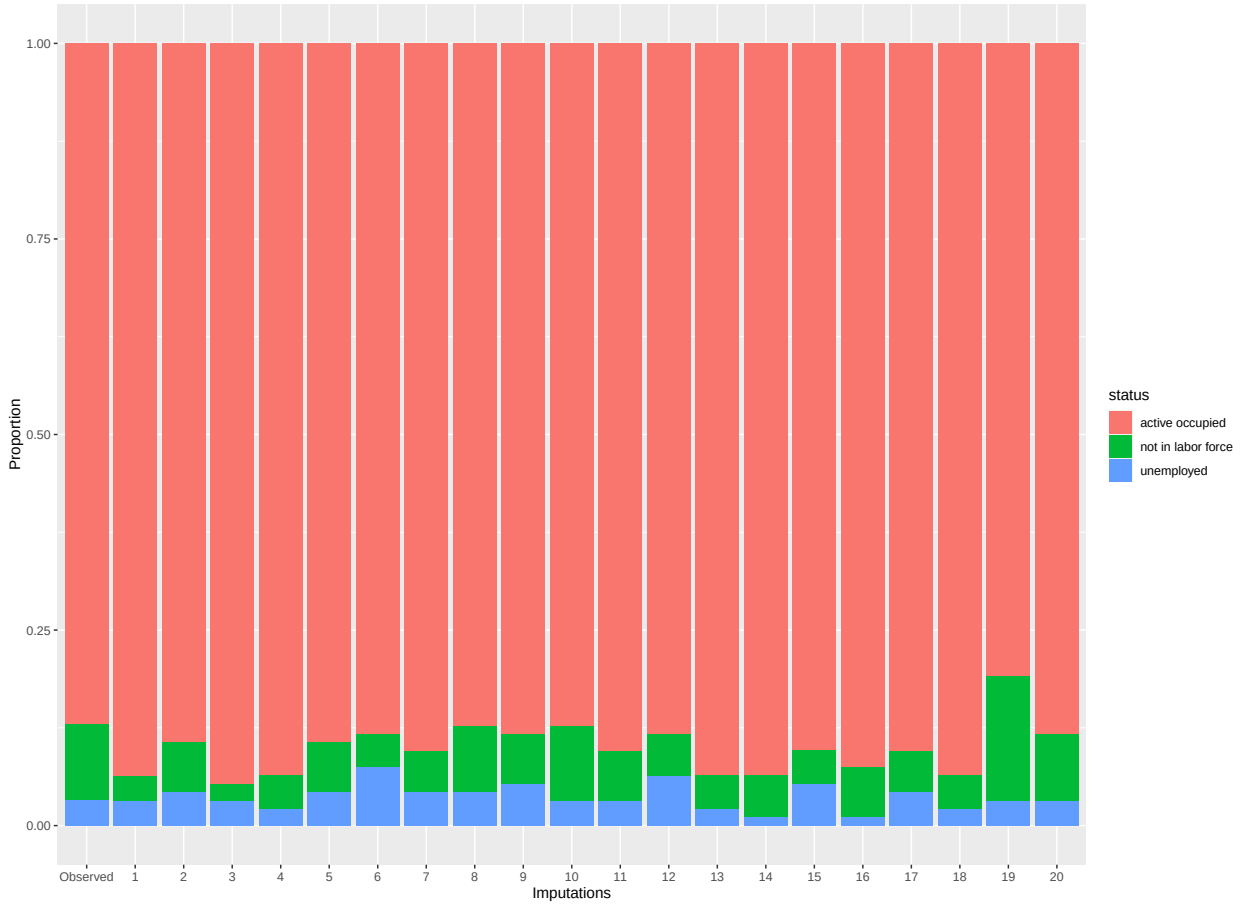



Figure 3.4: Distribution of the observed work status (first column) and the distributions on each of the twenty imputations (labelled as "1" to "20").

To summarise, differences between imputed and observed values are explainable from their different distributions in terms of independent variables. Moreover, the imputation model

makes sense interpretation-wise. Even if some coefficients have high values (i.e. supervising tasks and position in the firm), they correspond to situations linked to stable work positions.

**Imputation of the type of employment**

We focus on the imputation of the second illustrative variable, namely the type of employment. It is split in five categories: *employed by private household (houseworker, baby-sitter)*, *employee of own Public Limited or Limited Liability Company*, *self-employed*, *partner in his/her relative's firm* and *employee of another private firm or government organisation*.

1. *Logical skip* Individuals not in the workforce were not concerned by the question about the type of employment.

2. *Value retrievable* No value is retrievable elsewhere in the dataset.

3. *Value deducible* No value is deducible.

4. *Should it be imputed* There is no obvious contraindication to the imputation of this variable.

Regarding the imputation of the work status, we divided the imputation process based on the work status during the last wave. Specifically, we considered the unemployed and not in labour situations together. We explored the inclusion of the work status during the current wave as a predictor, but the results were inconclusive. Hence, the imputation model only includes the intercept for unemployed or not in labour cases.

For the active occupied status, simply applying our procedure leads to an over-parametrised model with some coefficients exhibiting high variance. This issue arises due to the presence of rare types of employment within the dataset. Among the 368 individuals belonging to the training sample, 347 were an employee of another private firm or government organisation, 10 were a partner in his/her relative's firm, six were self-employed, 1 was an employee of its own Public Limited or Limited Liability Company, and 4 were employed by a private household (houseworker, baby-sitter). Therefore, due to some very rare categories, we chose to only use the type of employment during wave 3 as a predictor in the model. Due to the rarity of certain employment categories, we made the decision to simplify the model by solely using the type of employment during wave 3 as a predictor. However, even with this simplified approach, a challenge emerged. Among the individuals requiring imputation, one had been an employee of his/her own Public Limited or Limited Liability Company during the previous wave, while none of the individuals who participated in waves 3 and 4 held this position. Consequently, it was not feasible to impute this missing value using the existing model. Even if it is not impossible that this individual was not an employee of his/her own Public Limited or Limited Liability Company any more, it is unlikely. Hence, we chose to impute the missing value as

an employee of their own Public Limited or Limited Liability Company. Additionally, this approach ensures that the sample's variability is preserved by not completely disregarding this particular employment scenario.

This example illustrates the difficulty of applying our process automatically.

**Imputation of the level of interference of work on private life**

The last illustrative variable is the level of interference of work on private life. It is a scale going from 0 to 10; the higher, the more interference there is. The sequence of questions was first applied. Since the conclusions are similar to the ones of the types of employment, we do not show it. As for the two other illustrative variables, the process was separated according to the work status during the last wave. We only detail the active occupied case.

Concerning the imputation itself, the pool of predictors consists of socio-demographics, variables related to the job at the previous wave, and already imputed work variables, for a total of 60 variables. Among them, between 28 and 30 are selected after the bivariate selection. Two variables from the previous wave, namely the percentage of work and the satisfaction with the workload, are sometimes included and sometimes not. Then, for each iteration, the stepwise AIC procedure leads to an imputation model with 11 independent variables. The linear regression model of imputation for the first iteration is displayed in Table 3.6, the models for the other iterations being very similar.

Individuals who worked on weekends, had unchosen variable working hours, had a full-time contract, were exhausted, or had difficulties disconnecting from work have, unsurprisingly, a higher interference of work on private life. Moreover, a higher interference of work on private life during the last wave induces a higher interference during the current wave. Since most individuals have kept the same job, this relationship is not surprising. Concerning the other variables issued from the last wave, the interpretation is not as straightforward. Indeed, both having a stressful job and possibilities of advancement have significant coefficients: having a job that was not stressful produces a level of interference that is, on average, 0.63 lower, and an increase of one point in satisfaction with opportunities of advancement induces a decrease of 0.10. On top of that, the age and satisfaction with the work conditions during the last wave also appear in the model but are not significant. Since these variables are not clear interpretation-wise, it could be sensible to remove them afterwards from the model.

Not all missing values were imputed with this model. Indeed, some individuals were neither concerned by opportunities for advancement or the type of employer (private vs public). Therefore, we discarded these variables to impute these missing values and built simpler models.

Figure 3.5 shows the density of the observed values and the densities of the imputed values for each imputation. First, the density of the observed values shows two maximums at 0 and 5, while most of the shapes of the density of the imputed values are closer to Gaussian distri-

| variable | coefficient | p-value |
|---|---|---|
| (intercept) | -0.68 | 0.658 |
| age | 0.09 | 0.13 |
| **variables from the previous wave** | | |
| work conditions | 0.21 | 0.009 |
| satisfaction with opportunities of advancement | -0.08 | 0.079 |
| interference of work on private life | 0.33 | $< 0.001$ |
| stressful job (ref. yes) | | |
| no | 0.55 | 0.058 |
| **variables from the current wave** | | |
| satisfaction with workloads | -0.30 | $< 0.001$ |
| exhausted after work | 0.30 | $< 0.001$ |
| difficulty to disconnect from work | 0.15 | 0.003 |
| private or public employer (ref. private) | | |
| public | -0.58 | 0.022 |
| part-time or full-time (ref. part-time) | | |
| full-time | 0.87 | 0.001 |
| work on weekends (ref. yes) | | |
| no | -0.93 | $< 0.001$ |
| type of working hours (ref. same each day) | | |
| variable and you do not decide | 0.61 | 0.041 |
| variable but you decide | -0.10 | 0.760 |
| N | 302 | |
| $R^2$ | 0.42 | |

Table 3.6: Linear regression model to impute the level of interference of work on private life at the wave 5, for the first of the 20 imputations. Some variables are issued from the previous wave and some other from the current wave. Work conditions, satisfaction with opportunities of advancement, interference of work on private life, satisfaction with workloads, exhausted after work, and difficulty to disconnect from work are all scales going from 0 to 10. The number of observations used to fit the model (N) and Nagelkerke $R^2$ are also displayed.

butions. Even if not perfect, the Gaussian distributions are a relatively good approximation of the distribution of the observed values. Then, imputed values tend to be higher than observed ones. This is explainable by the different characteristics of the missing individuals compared to the observed ones. In Figure 3.6, the densities of the main numerical predictors in the imputation model are compared between observed and missing individuals. The interference of work on private life at wave 4 only has one density concerning the missing individuals because they took part in wave 3. In contrast, the three other variables concerned wave 4, and hence were missing.

On the one hand, we do not observe any clear difference between observed and missing individuals regarding satisfaction with workload and exhaustion after work. On the other hand, the "difficulty in disconnecting from work" and "interference of work on private life" present higher values for the missing individuals. These higher values explain the higher values for the interference of work on private life compared to the observed ones.

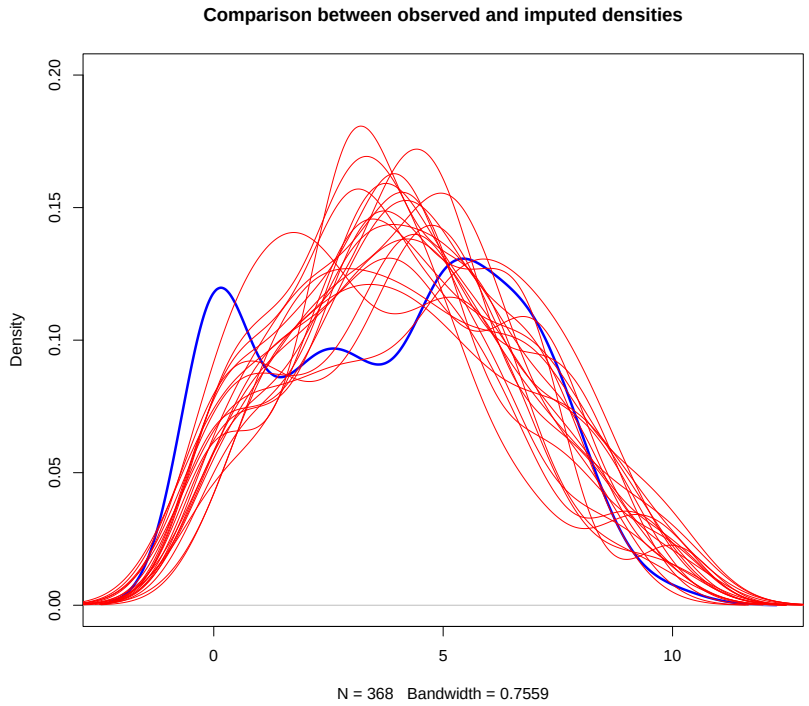



Figure 3.5: Comparison between the observed density of the interference of work on private life (blue) and the densities of the twenty imputed datasets (red). The densities are represented as continuous lines even if only integer values between 0 and 10 are possible.

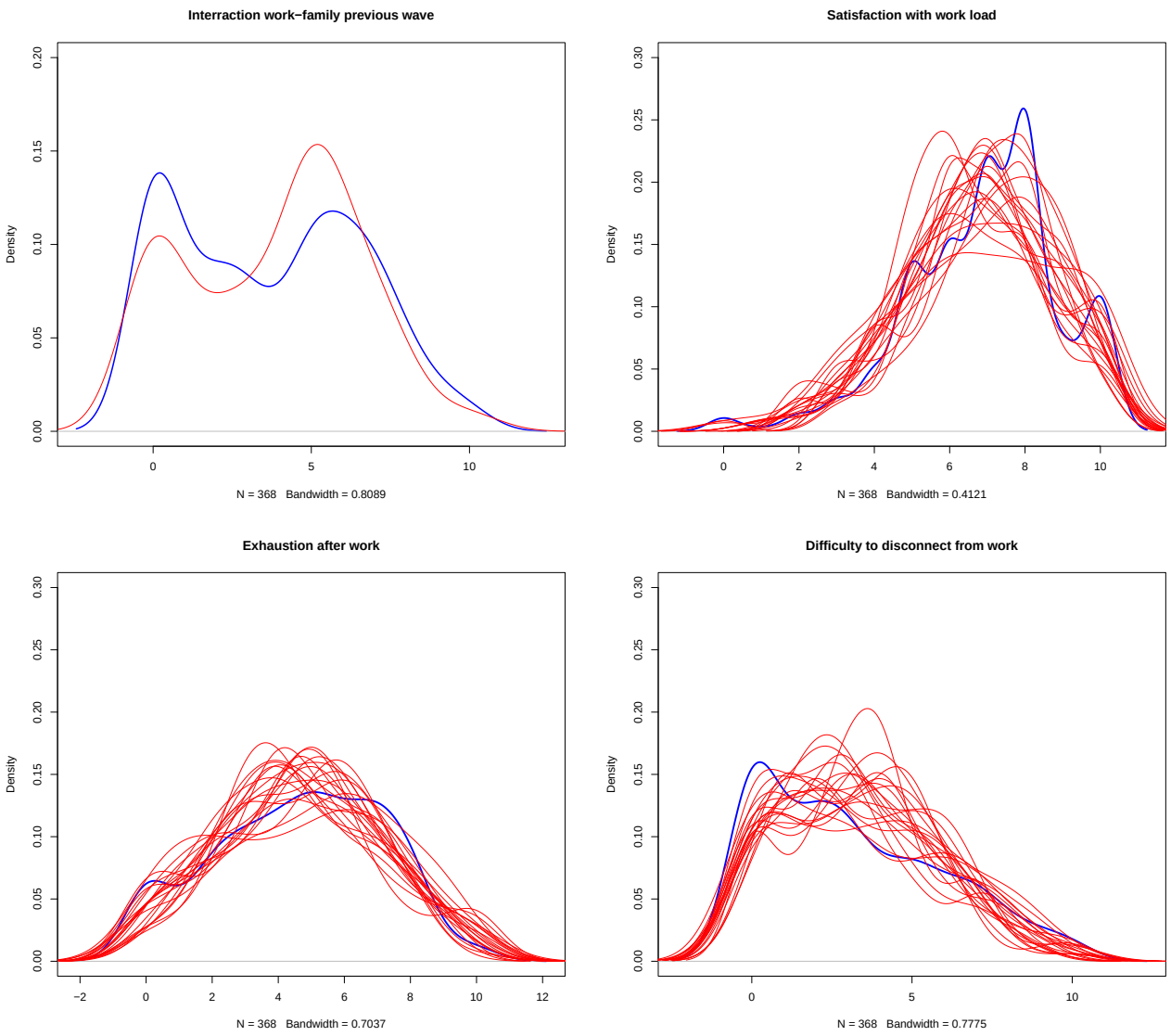



Figure 3.6: Comparison of the density of the interference between work and family at wave 3 between observed (blue) and missing (red) values, and comparison between the observed density (blue) of, respectively the satisfaction with the workload, exhaustion after work and difficulty to disconnect from work, and the densities of the twenty imputed datasets (red). The densities are represented as continuous lines even if only integer values between 0 and 10 are possible.

## 3.5 Discussion

In this research, we have addressed the issue of imputing missing data in a longitudinal multivariate dataset. Our main contributions include a series of questions for identifying missing data and values to be imputed and a framework to treat missing data. The process was applied to a sample received from colleagues. This dataset was particularly suitable for illustration purposes, given its longitudinal nature and the presence of categorical data and logical missing

patterns. Imputing this dataset was essential, as it consisted of 849 individuals, with only 506 individuals having participated in all waves. Excluding individuals with missing data in any wave would significantly compromise the statistical power of subsequent analyses. Moreover, we may expect, as it is generally the case, that missing data are MAR and additional bias may be induced by complete case analysis.

While our proposed procedure serves as a valuable baseline, it faces several challenges that warrant further consideration. These challenges primarily involve the effective handling of categorical data, dealing with a relatively small dataset, and effectively addressing filtering questions. In order to provide a more comprehensive examination of these challenges, we delve into their implications and explore potential solutions. To illustrate these challenges, we present examples derived from the illustrative subsample.

Imputing categorical data is often more complex than imputing numerical variables, as rare categories can cause issues at various stages of the process. When the variable being imputed has a limited number of categories, the number of predictive variables included in the model should be restricted to avoid an ill-defined multinomial logistic model. Additionally, when the imputed variable is a predictor in the imputation model, it can lead to estimated parameters with high variance, especially when some categories are rare. In some cases, a category may appear in observations that fit the model but not in those used to train it. For example, one individual that had wave 4 missing was an employee of its own public limited or limited liability company during wave 3, but none of the individuals that answered both waves 3 and 4 was in this situation in wave 3. Therefore, when the type of employment at wave 3 is included in an imputation model, this model cannot be applied for this particular individual. Bootstrapping the observations can also result in infrequent categories of predictive variables being excluded from the bootstrap dataset. We considered combining categories to avoid these issues. However, caution must be exercised when fusing the categories. While automating the imputation process, one might be tempted to merge all the smallest categories and retain the largest one, resulting in just two categories. However, such an approach could potentially merge categories that are conceptually distinct. An example of that would be the highest level of education achieved, that could take values “compulsory school”, “upper secondary” and “tertiary”. Even if “compulsory school” and “tertiary” were the rarest categories, it does not make much sense to fuse them, due to the grading between the three categories. Furthermore, the decision to merge categories may depend on the subsequent analytical model employed. For instance, in the case of illustrating the imputation of work status, we combined the categories of “supervision/training position”, “management position”, and “other” situations. Nevertheless, this grouping is problematic if the research aims to examine differences specifically between management positions and non-management positions.

The number of observations limits the number of variables that can be included in a multi-

nomial model. De Jong et al. (2019) demonstrated that both the sample size and the ratio of observations in the smallest category to the number of predictors impact the performance of a multinomial logistic model. Adding too many predictors can lead to overfitting and poor results on observations being imputed. Therefore, in some cases, the model chosen by the AIC stepwise procedure may not be optimal due to the high variance of some estimated parameters. In these instances, we advise selected the most significant variables and those that are intuitively linked to the dependent variable.

The final challenge concerns the inclusion in an imputation model of variables that may be logically missing. For categorical variables, one approach is to consider missing values due to filtering as an additional category. However, this can be problematic when the variable inducing the missing values is also in the model. It leads to high variance in the regression model estimates due to near-collinearity. In this situation, combining the variables into a single variable with categories such as "not employed", "employed - private employer", and "employed - public employer" may be a better approach. Alternatively, using a classification tree rather than a logistic model (Burgette and Reiter, 2010) may be a suitable solution. However, it demands that each situation is explicitly considered. For numeric variables, this is not possible. Therefore, we fit a simpler model without certain variables to impute observations with missing values due to filtering questions.

In this research, we have presented a process for imputing missing data in a longitudinal and multivariate dataset, including a sequence of questions for identifying missing data and values to be imputed and a framework for implementing multiple imputation. While our proposed process can serve as a guide, it is not a one-size-fits-all solution due to the complexity of survey data. The researcher's knowledge and expertise in the data are essential for successful imputation. Based on our analysis, we recommend the following steps when imputing a longitudinal dataset with filter questions:

- Analyse the missing data to identify which situations are more prone to missing values. Consider variables typically related to missing data. If any of these variables stand out, they should be included in all imputation models to support the a priori MAR assumption.

- Apply the sequence of questions outlined in this research to each variable to determine the "true" missing values.

- Determine how filtering affects the dataset and order the variables accordingly for the imputation. A variable that filters another one should be imputed first.

- For the imputation process, start by addressing item-level missing data to avoid too many cases. Going from the first to the last wave, impute variables in the determined order, using the three-step imputation process as a base. However, the specific approach will

depend on whether the variable being imputed is numerical or categorical. For both types of variables, potential predictors may need pre-processing. Categorical variables should be reduced to a small number of categories. For numerical variables, the imputation process can generally be applied without restrictions. It is advisable to examine the resulting imputation model for categorical variables and remove variables with inflated coefficients or high variance.

- Identify the typology of unit-level missing data to determine what information is available and how the imputation process should be structured. For example, in our case, we started by imputing the middle gaps, before moving on to the end gaps and, finally, the initial gaps.

We made the assumption, following the standard practice, that the missing mechanism was MAR. However, it is crucial to acknowledge the possibility that the mechanism may have been MNAR. Therefore, when there is a suspicion of a MNAR missing mechanism for one or more variables, it is essential to exercise caution. Van Buuren (2018) suggested three steps to be applied when such a mechanism is suspected. The first step is to add variables that are suspected to be predictive of missingness in the imputation model. If this is suspected not to be enough, it is advised to perform a simulation to determine the magnitude that a MNAR mechanism must have to influence the results. If this is also not enough, it is suggested to test non-ignorable imputation models, which is called a "sensitivity analysis" (see also Van Buuren (2018) for an example of a sensitivity analysis).

In this study, we set ourselves in a broad scope of imputation, which typically arises when the imputer and the analyst are different individuals. However, if the analysis model is known at the moment of the imputation process, it is recommended to adopt a narrower scope. Specifically, the imputation model should be compatible with the analysis model, incorporating all variables, interactions, and non-linearity (Van Buuren, 2018) that will be used in the analysis. Our procedure remains applicable in this scenario, but it is crucial to ensure that the variables included in the analysis model are integrated into the corresponding imputation models.

This research has some limitations. One challenge with our proposed procedure is how to combine the results afterwards. For example, suppose we are interested in the relationship between the type of working hours and job satisfaction in a specific wave. In that case, we might conduct a regression with job satisfaction as the dependent variable and type of working hours and other control variables as independent variables. The regression parameters and variance are typically calculated on each completed dataset and combined using Rubin's rule. However, if an individual has unit-level missing data for the wave of interest, some completed datasets might include an imputed active occupied status while others do not. In the first case, it would be included in the calculation of the parameter of interest, while in the second it would not, since both variables of interest would be missing due to a logical skip induced by a non-active

occupied state. As a result, the parameter calculation would be based on different samples depending on the iteration, which could impact the overall results. Then, our method is not automatically applicable. The researcher still has to intervene in the imputation process. This approach guides the researcher but does not replace it. Finally, we decided to limit the number of predictors to add to the imputation models. This limitation was done for two reasons. First, to avoid overfitting the data and then to cope with logical skips. Indeed, as already mentioned, adding a variable to the imputation could reduce the sample used to fit this model. However, there is no agreement in the literature.

# Chapter 4

# Comparison of imputation methods in the case of life course data[1]

## 4.1 Introduction

This chapter aims to review multiple imputation methods proposed so far for life course data, and to assess their practical relevance using real data on which we simulate missing data. By doing so, we aim to provide clear methodological guidelines and to strengthen missing data handling in life course research. In the meantime, we also develop the *MICT-timing* algorithm, which is an extension of the *MICT* algorithm. This innovative multiple imputation method improves the quality of imputations in trajectories with time-varying transition rates.

The life course paradigm has gained increasing importance in the social sciences over the last decades. It has proved its contributions in numerous disciplines ranging from sociology, demography, gerontology, and medicine to psychology (Elder et al., 2003; Bernardi et al., 2019). This paradigm insists on the need to study not only the situation at a given time point but also its evolution over the life course in the medium or the long run. These trajectories are then often described with categorical data. For instance, the school-to-work literature focuses on professional integration trajectories following compulsory education, distinguishing between education, employment, or unemployment (e.g. Brzinsky-Fay and Solga, 2016).

This life course perspective, therefore, implies the use of longitudinal data over the medium to the long run. This data requirement is highly sensitive to missingness because it multiplies missing data occasions and retrospective questions tend to be more challenging to answer. The lack of a commonly accepted solution to handle missing data is one of the significant challenges faced by life course methodology (Piccarreta and Studer, 2019).

[1]This chapter is a joint work with Matthias Studer from the University of Geneva and André Berchtold from the University of Lausanne. The chapter appears under the form of a self-containing article, because it has the objective to be published under this form. As a result, there is some overlap with the introduction.

Several missing data-handling strategies have been suggested. Their relevance typically depends on the missing data mechanism (e.g. Little and Rubin (2019)). When data are missing completely at random (MCAR), observations with missing value(s) can be removed, a strategy called complete case analysis or listwise deletion. This strategy leads to unbiased estimates, but it can drastically diminish the statistical power of the analysis depending on the number of cases with missing data. Since they tend to be frequent in longitudinal analysis, this is a wasteful strategy. In practice, missing data are more frequently missing at random (MAR), i.e. missing data occurrence is linked to specific profiles. For example, panel attrition, i.e. individuals leaving a longitudinal survey, is linked to vulnerable situations such as unemployment, migration background, or poor health (Rothenbühler and Voorpostel, 2016). When data are missing not at random (MNAR), meaning that the probability to be missing depends on the missing value itself, standard strategies can lead to bias, and a model for the missing data generation process is needed.

In addition to the deletion of missing data, we can distinguish three strategies to handle missing data: weighting methods, likelihood-based approaches, and imputations, whether simple or multiple (Molenberghs et al., 2014). First, the fully observed trajectories can be weighted to make the sample more representative of the target population, at least in terms of socio-demographic characteristics. These weights are often based on the characteristics used to build the original sample, such as gender, age, or occupation. Second, likelihood-based approaches rely on a hypothesised distribution of the complete data, often the multivariate normal distribution. Provided the missing data mechanism is either MCAR or MAR, an unbiased estimate of the distribution parameters can be obtained. Finally, imputation methods work by replacing missing values with probable values. The simplest way to do this is to replace each missing value with a single value, resulting in a complete dataset. However, the inherent uncertainty of missing values is ignored. The multiple imputation framework aims to solve this issue by randomly imputing the missing values $M$ times, resulting in $M$ datasets (Rubin, 1987). The analysis is then performed separately on each dataset before aggregating the results. Generally speaking, multiple imputation is a highly efficient and flexible strategy (see e.g. Molenberghs et al. (2014)).

The multiple imputation framework requires an imputation method. The two most common ones are fully conditional specification (*FCS*) (Van Buuren et al., 2006), also called chained equations (Van Buuren and Groothuis-Oudshoorn, 2011), and joint modelling (*JM*) (Schafer, 1997). *FCS* uses a separate imputation model for each incomplete variable. The algorithm then iterates over each variable to impute the missing values in a predefined order. The whole imputation process is performed several times until a convergence in the distribution of the imputed variables is achieved. *JM* is based on a multivariate model fitted to the data, which generally assumes a normal distribution. The imputations are then randomly drawn from this

model.

*FCS* and *JM* apply with longitudinal data by treating repeated measurements of a variable over time as different variables. However, in their standard form, they do not directly consider the (usually strong) relationships between successive observations of the same variable. For this reason, several variants of these algorithms have been proposed, including two-fold *FCS* (Nevalainen et al., 2009). This algorithm relies only on the previous, current, and subsequent time points and not on all the time points. In addition, the algorithm runs several times in a row at the same time point before moving on to the next.

Several imputation algorithms were developed specifically for categorical longitudinal data. Gabadinho and Ritschard (2016) proposed to use a variable-length Markov model (*VLMC*) to impute missing values. Halpin (2016b) proposed the "Multiple Imputation for Categorical Time Series" (*MICT*) method. It works by imputing missing data gaps from their edges, which ensures longitudinal consistency of the imputed values.

Previous studies have compared different multiple imputation methods on longitudinal data. Kalaycioglu et al. (2016) conducted an in-depth comparison of Bayesian multiple imputation, multivariate normal joint modelling, and different versions of *FCS* using real and simulated data of a numerical outcome with different types of explanatory variables. De Silva et al. (2017) conducted a simulation study to compare *FCS*, two-folds *FCS*, and *JM* in the presence of a time-varying covariate with a non-linear association with time. Huque et al. (2018) used 12 algorithms, which were different variations of *FCS* and *JM* imputation algorithms, and compared their impact on both a linear regression and a linear mixed-effect model. All three studies focused on estimating regression parameters with a numerical outcome. The first study found that two-folds *FCS* should be preferred when successive data points are highly correlated, while the two others concluded that *FCS* and *JM* generally perform well. However, none of these comparisons involved a categorical data. As a result, *MICT* and *VLMC* were not evaluated. Furthermore, these studies have focused on regression coefficients, whereas other methods, such as classification, are also extensively in life course research. Therefore, we lack a comparison of imputation methods in this context.

Longitudinal categorical data generally shares several characteristics requiring a specific missing data treatment:

- Because of their longitudinal structure, missing values usually appear as gaps, i.e. consecutive missing observations. This may be due, for example, to individuals momentarily leaving a survey or periods not filled in a retrospective life history calendar.

- Their categorical coding makes it more difficult to use standard tools for handling missing data. Categorical data are not normally distributed, making the application of likelihood-based methods and joint modelling challenging.

- The imputation of categorical data (except for ordinal variables) is either correct or false, with no gradation between these two extremes.
- Although life course data tend to have few transitions, their longitudinal consistency is crucial. Life course trajectories are often viewed as a whole, not as a time-ordered juxtaposition of separate events (Piccarreta and Studer, 2019).

This study aims to evaluate the relative strengths and weaknesses of several imputation algorithms for life course data. Based on these results, it aims to provide recommendations to guide life course researchers in their choice of the imputation method and on how to set their associated parameters. For this reason, we develop a simulation framework specifically designed for life course research. The framework relies on six datasets, each of them highlighting common longitudinal data characteristics. We then simulate missing data in these datasets, according to three different missing data models. These models aim to reproduce common missing data patterns in life course research. Finally, the imputations are evaluated according to three key aspects when analysing longitudinal categorical data in a life course perspective, namely the timing, the duration and the sequencing of the processes (Studer and Ritschard, 2016).

Alongside these simulations, we propose two extensions of the *MICT* algorithm. First, we propose the *MICT-timing* extension aiming to better take the *timing* aspect of the trajectories into account. The *MICT* algorithm implicitly assumes that the transition rates are *homogeneous* over time. However, this assumption is unrealistic in many life course research, which might lead to poor imputations. Second, we consider the use of random forests instead of a multinomial model. Random forest is commonly used for missing data imputation (Burgette and Reiter, 2010; Shah et al., 2014; Doove et al., 2014). It can handle non-linear relationships, interactions effect and it is immune to irrelevant predictors (Friedman et al., 2001). All these features might be crucial for longitudinal data. Indeed, specific state combinations might trigger long-term effects. Furthermore, random forest is generally not too sensitive to the choice of its parameters, and works pretty well with its standard parameters (Cutler et al., 2012).

The remainder of this chapter is organised as follows. First, the various imputation methods being compared are introduced. We then present the simulation framework, including the datasets used, the missing data generation processes, and the evaluation criteria. Finally, we delve into the results obtained and conclude with a discussion of the findings and by issuing general recommendations.

## 4.2 Imputation methods

This section presents the algorithms that are compared, including the new *MICT-timing* algorithm, together with their tested parameterisations. We integrate four methods in our comparison: *FCS*, *MICT*, *MICT-timing* and *VLMC*. To allow for a meaningful comparison, we use,

when possible, similar parameterisations across the different methods. The only exception is the imputation with *VLMC* models that function differently from the other algorithms.

Complete case analysis, which consists in keeping only the trajectories that have no missing data, which is still a standard strategy in social sciences (Berchtold, 2019), is used as a baseline.

### Fully conditional specification (FCS)

Fully conditional specification (*FCS*) imputes missing data through the specification of a conditional distribution for each variable (Van Buuren et al., 2006). It works in the following way for categorical data:

1. For each variable, missing data are first imputed from their marginal distribution.

2. An imputation model is defined for each variable using a multinomial regression or a random forest for categorical data. All the other variables are commonly used as predictors in the imputation model.

3. The algorithm runs then through the variables. For each variable, it fits the imputation model from all available data. Then, it uses it to predict a replacement value for each missing data. The algorithm goes through the variables until it reaches a predefined number of iterations.

4. The values obtained after the last iteration are kept.

5. The process is executed several times if multiple imputations are required.

The algorithm applies to longitudinal data by considering repeated measurements as distinct variables. However, as pointed out by Kalaycioglu et al. (2016), it might face convergence issues in its standard formulation due to collinearity issues.

Two-fold fully conditional specification aims to solve this issue by limiting the predictors to the ones observed at the same time points and a limited number of previous and future measurements of the variable to impute. In addition, the algorithm runs several times through the variables at each time point before moving to the next. However, this last point does not apply to univariate longitudinal data, as in our case, since there is only one variable at each time point. Therefore, for simplicity, we will call *FCS* the algorithm even if all variables are not used as predictors.

We test *FCS* both with multinomial and random forest as imputation models. The number of predictors applied was one or five predictors in the past and one or five in the past and future.

### Multiple Imputation for Categorical Time Series (MICT)

The algorithm "Multiple Imputation for Categorical Time Series", i.e. *MICT*, was introduced in Halpin (2012) and Halpin (2013). It is specifically designed to handle missing data in longitudinal categorical datasets. The imputation themselves are based on a statistical model. This model can include time-invariant covariates, such as gender, or time-variant covariates, such as the number of children. Furthermore, a user-defined number of previous or future time points can also be taken into account to ensure the longitudinal consistency of the imputations. *MICT* relies on a multinomial model in its first formulation. We have added the possibility to use random forest models instead.

The original algorithm works by distinguishing six missing data patterns that are handled sequentially. It then adopts a slightly different imputation method for each of them. Figure 4.1 presents six sequences taken from a built-up dataset composed of two states (states A and B) and fifteen time points, each illustrating one of these patterns. For the sake of the illustration, let us assume that we use two time points in the future ($nf = 2$) and the past ($np = 2$) for the imputation.

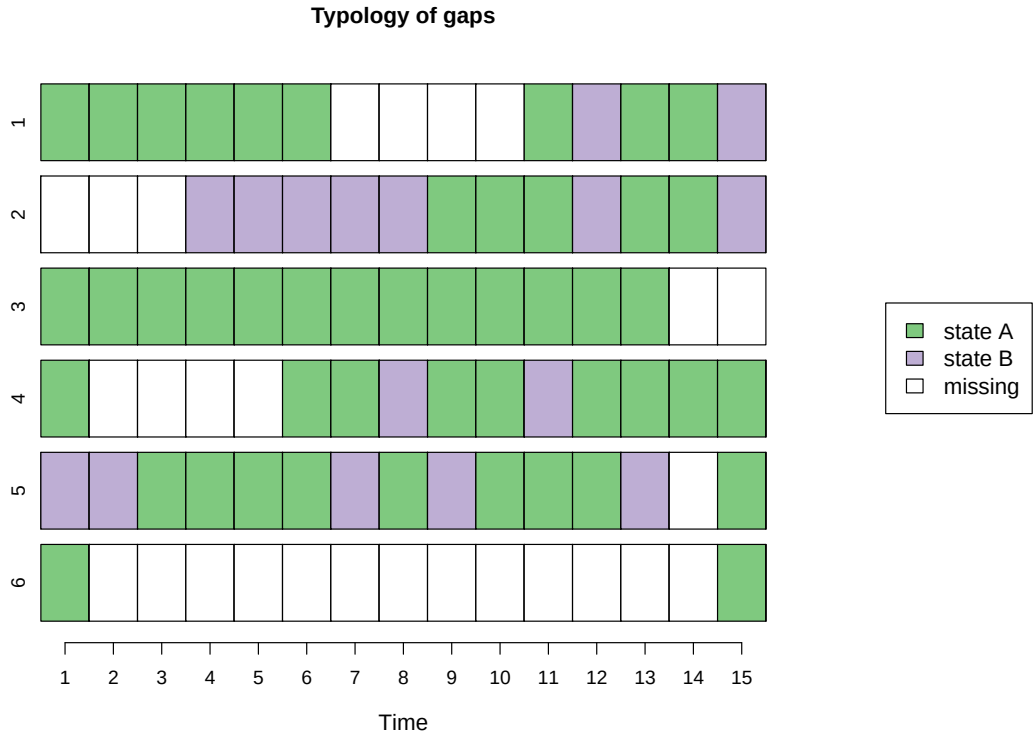



Figure 4.1: Typology of the different types of missing data gaps according to the *MICT* algorithm considering two predictors from the past and the future: 1. Internal gap, 2. Initial gap, 3. End gap, 4. Left-hand side gap, 5. Right-hand side gap, 6. Both-hand side gap.

Sequence 1 illustrates an *internal gap*. Here, sufficient information, i.e. at least two ob-

servations before and after the gap, is available to impute the missing data. In this case, the *MICT* imputation process fills gaps recursively from their edges. This strategy ensures that imputations are consistent and based on the closest observed values. Figure 4.2 illustrates the ordering of the imputations for two toy sequences with gaps of different lengths. Concretely, the imputations are made using a multinomial model. In our example, this model uses predefined covariates (such as sex), the two previously available values in the sequence, either observed or previously imputed, and the two ($nf = 2$) subsequent values. In our example, the first missing data of Figure 4.2 uses the two values before the gap, i.e. the states at times 2 and 3, and the two following the gap, i.e. the states at times 8 and 9. This multinomial model is first estimated using similar fully observed patterns in the data. For example, the first trajectory would provide one observation: predicting the state at time 10, using the states at time 8, 9, 14, and 15, while the second trajectory would provide six observations: predicting the state at time 3 using the states at times 1, 2, 7 and 8 as predictors, the state at time 4 using the states at 2, 3, 8 and 9, until the state at time 10 using the states at times 8, 9, 14 and 15 as predictors.

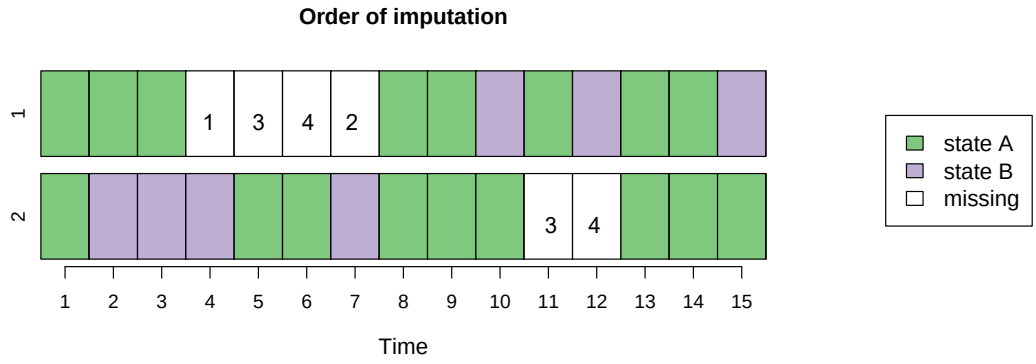



Figure 4.2: Imputation order for an example with two gaps of different lengths.

Once all internal gaps are imputed, the initial and terminal gaps are considered. These gaps are illustrated using sequences 2 and 3 of Figure 4.1. Since initial gaps have only one edge, imputations start here from the far right to the left and can use only predictors from the future. Indeed, there are no observed data points in the past. Here again, the imputation model is estimated based on similar fully observed patterns in the data. The same (but reverse) strategy is used for terminal gaps.

Finally, the rarest cases, namely left-, right- and both-hand side gaps, are imputed at the end. Left-hand side gaps (sequence 4 of Figure 4.1) have enough observations for the imputation model after the gap but not before (only one in our example). Here, the algorithm only considers one time point in the past instead of two. The right-hand side gaps used the reverse strategy, considering only one time point in the future. Similarly, both hand-side gaps use only one observation in the past and one in the future.

We test *MICT* both with multinomial and random forest as imputation models. The number of predictors applied was one or five predictors in the past and one or five in the past and future.

### MICT-timing

The *MICT* algorithm estimates an imputation model using all fully observed patterns similar to the missing data to impute, regardless of their position in the trajectory. However, this assumption is not verified in many life course applications. For example, the transition rate between education and work strongly varies over time, with individuals staying in education during childhood and mostly transitioning from education to work between the ages of 16 and 30. In such situations, using a constant transition rate might lead to wrongly imputing transitions to work during childhood, which is impossible.

To overcome this limitation, we developed the *MICT-timing* algorithm. It modifies the original *MICT* imputation process in two ways. First, the imputation of the gaps of a specific length is made separately, according to *when* they happen. Second, only similar fully observed patterns in a time frame around the missing data to impute are used to estimate the imputation model. As a result, the transition rates are specific to a given time frame. An additional parameter specifies the radius of the time frame. Using a radius of length 0, only the observations occurring at the same time points as the missing data to impute are used. Using a radius of one, patterns occurring one time point before or after are used in the process. Finally, using a radius of the length of the sequence minus one is equivalent to the original *MICT* algorithm. As with the original *MICT* algorithm, the algorithm can apply either multinomial or random forest models.

Figure 4.3 illustrates the difference between the two algorithms. The *MICT* algorithm simultaneously imputes gaps of the same length, regardless of their place. For instance, only one imputation model was built to impute the missing values labelled “3” and “4”. With the MICT-timing algorithm, two separate models are fitted, one to impute the missing value labelled as “3” and another one for the missing value labelled as “4” since they do not occur at the same time point. Furthermore, only the observations included in the time frame of the predefined radius are used to fit the imputation models. For example, with a radius of length 1, the second sequence only provides three observations (the times 3, 4, and 5 and their predictors), instead of the six with the *MICT* algorithm, for the imputation of the first missing data (labelled as “1”), and the first trajectory does not provide any observation.

As for *FCS* and *MICT*, we test *MICT-timing* both with multinomial and random forest as imputation models. The number of predictors applied was one or five predictors in the past and one or five in the past and future. Moreover, radius of 0 and 5 are considered for the length of the time frame. With a time frame of radius 0, the algorithm account for the potential time-heterogeneity since only the observations arising at the similar time-point as the missing

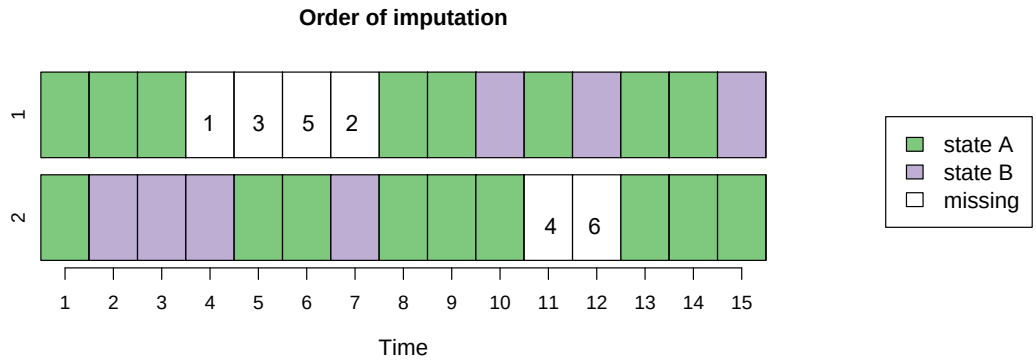



Figure 4.3: Order of the imputation of the *MICT-timing* algorithm.

data to impute are used to fit the imputation model. However, the predictive performance of the multinomial model (De Jong et al., 2019) and, to a lesser extent, the random forest (Luan et al., 2020), depends on the sample size. Therefore, a time frame with a radius of 5 may represent a trade-off between the desire to account for time-heterogeneity and the need to have a sufficient sample size for accurate predictions.

### VLMC

Variable-length Markov chains (*VLMC*) are a type of Markovian model that do not consider a predefined constant number of time points to predict a current situation. On the contrary, the number of past states required to summarise the whole past depends on each situation. Two algorithms are available to fit a *VLMC* model: Learn-PSA (Ron et al., 1996) and the context algorithm (Rissanen, 1983). Starting from subsequences of maximum length, both algorithms compare the conditional distribution of a subsequence with its suffix of maximum length, which is the subsequence without the first state. If the conditional distributions are sufficiently close, the conditional distribution of the suffix is kept instead of the one of the whole subsequence. The two algorithms differ in the criteria used to compare the distributions. The criterion implemented by Learn-PSA is based on the ratios between the individual conditional probabilities. The underlying idea is that if at least two probabilities are sufficiently different, the conditional distribution of a given subsequence cannot be approximated by that of its suffix. The criterion used in the context algorithm is based on the differences of deviance between the conditional distribution induced by the subsequence and its suffix of maximum length. If the difference exceeds a given $\chi^2$ quantile, the distribution of the current subsequence is chosen.

*VLMC* models can be used to impute missing data. Missing data gaps divide sequences into subsequences. For example, the first sequence in Figure 4.1 is divided into two subsequences, one of length six and one of eleven. A *VLMC* model is fitted with the dataset consisting of these two observed subsequences. The missing data gaps are then filled from the left based

on the probabilities induced by the subsequence preceding the missing data to be imputed. In practice, one value is drawn based on the probabilities induced by the three-state subsequence between the times 1 to 3. Then, the second missing value is imputed based on the subsequence consisting of the three observed states and the imputed value at the time 4. For initial missing data gaps, such as the one shown in sequence number 2 in Figure 4.1, the first missing data is imputed based on the distribution of states in the dataset. The process is then similar to that used for other types of gaps.

We have chosen *VLMC* models instead of the standard Markov models because they are more flexible. With complex sequences, we may have a high-order dependency and, hence, may need high-order Markov models, which have many parameters.

Concerning the parameterisations considered for this comparison, we built *VLMC* models either with the Learn-PSA or the context algorithm. For Learn-PSA, we set the threshold value to either 1, 1.05, 1.1, 1.2, or 1.5. We set the quantile values for the context algorithm to either 0.1, 0.05, 0.04, 0.03, 0.02, 0.01, or 0.001. In both cases, we chose the imputation model having the best AIC. It is worth noting that, unlike *MICT* and *FCS*, *VLMC* uses only past information. Therefore, the other algorithms were fitted with past predictors only to allow for a fair comparison.

## 4.3 Simulation Framework

As a recall, this chapter aims to assess the strength of each imputation method using simulations. Each step of these simulations is explicitly linked to the life course theory to ensure the relevance of the resulting recommendations. Generally speaking, the simulation framework is designed as follows. We first randomly insert missing data into an existing dataset using three models. The simulations are based on different datasets, each presenting different characteristics encountered in life course research. The missing data-generating models are also designed to generate patterns often encountered in longitudinal studies. Then, these datasets with missing data are imputed with the imputation methods and parameterisations, introduced before. Finally, the quality of the imputations is evaluated with criteria based on the recovery of key characteristics for life course research.

We organise the presentation of the simulation framework as follows. We first present the choice of the datasets and their characteristics before the missing data generation models. We finally discuss the evaluation of the imputations.

### Data

The life course paradigm emphasises understanding how trajectories unfold over time, and the data aims to describe these trajectories. Such data can be characterised by different aspects

depending on the research questions and the data at hand. In this chapter, we rely on six complete real-world datasets to encompass the different configurations of these characteristics.

First, some datasets are characterised by very strong timing aspects, which means that some states or transitions typically occur at specific time points. For instance, most people live by their parents at age 10, which is generally not the case at 40. In other databases, the timing aspect is less prominent. For instance, in professional trajectories, change in working status can occur anytime between 20 and 40 years old. This is often the case when trajectories are described using calendar time instead of process time (typically the age).

Second, trajectories typically differ by the characteristics of their transitions. While some processes show very few transitions, others are more volatile. In addition, some processes are strongly ordered and seldom return to a previously visited state. This behaviour often occurs with developmental trajectories or when some transitions, such as from dead to alive, are impossible.

Third, the coding of the process itself might vary. While the time is sometimes measured on a monthly scale, generally resulting in longer sequences, it is often measured yearly. Furthermore, the level of detail used to describe the possible states occurring in trajectories can vary. We typically expect the imputation of more complex trajectories to be more difficult since there are more possibilities for incorrect imputations. These states can be ordered or unordered. With unordered states, imputations are either correct or not, while an imputation might be more or less correct with ordinal states. Finally, the number of cases typically varies between studies. Overfitting is more likely to occur in small datasets.

In this study, we selected six datasets to illustrate the frequent configurations of the above-mentioned aspects of life course research. These datasets and their main characteristics are presented in Table 4.1. Aside from a short description and the data source, Table 4.1 regroups information on the number of cases, the time unit, sequence length, the possible states, and their overall frequencies. The table also provides information on the overall percentage of transitions between time points and the average number of visits to visited states (ANV), which measure the recurrence of previously visited states (Pelletier et al., 2020). It varies from 1, when return to previously visited states never occur, to half the sequence length, when the process strictly alternates between two states at each time point.

Figure 4.4 regroups the chronograms of trajectories in each of the six datasets. These chronograms present the proportion of individuals in each state at each time point. Among others, it illustrates the timing regularities in trajectories and the overall frequencies of each state. This information is also summarised in Table 4.1.

| Trajectory | Data Source | Timing | $n$ | Time unit | Length | ANV | Transition (in %) | States | Freq. (in %) |
|---|---|---|---|---|---|---|---|---|---|
| Professional | SHP (retrospective) | Strong/ Process time | 3382 | Year | 26 | 1.33 | 10.1 | full-time work<br>part-time work<br>non-working<br>education | 50.1<br>10.3<br>12.6<br>27 |
| Cohabitational (4 states) | SHP (retrospective) | Strong/ Process time | 3710 | Year | 26 | 1.08 | 9.8 | with child<br>with partner, no child<br>with parent(s)<br>other | 34.3<br>16.6<br>32.2<br>16.8 |
| Cohabitational (8 states) | SHP (retrospective) | Strong/ Process time | 3710 | Year | 26 | 1.08 | 11.1 | living alone<br>with both parents<br>with one parent<br>with partner, without child<br>with partner and child<br>with child, without partner<br>with relative(s)<br>other | 13<br>27.1<br>5.2<br>16.6<br>32.7<br>1.6<br>1.1<br>2.7 |
| Civil Status | SHP (panel) | Weak/ Calendar time | 2324 | Year | 21 | 1.01 | 1.6 | married<br>separated<br>divorced<br>widowed<br>single, never married | 59.1<br>1.1<br>6.8<br>3.1<br>29.9 |
| Health Satisfaction | SHP (panel) | Weak/ Calendar time | 1259 | Year | 21 | 3.16 | 36.5 | low<br>average<br>high<br>very high | 3.7<br>11.4<br>49.3<br>35.6 |
| School-to-work transition | MVAD | Strong/ Process time | 712 | Month | 72 | 1.27 | 3.6 | school<br>further education (FE)<br>higher education (HE)<br>training<br>employment<br>joblessness | 8.5<br>16.2<br>11.7<br>10.3<br>44.7<br>8.6 |

Table 4.1: Datasets and their main characteristics. For each of the six datasets, the source of the data, the timing aspect, the number of trajectories (n), the time unit, the length of the trajectories, the average number of visits to visited states (ANV), the percentage of transition, the detail of the states (which are mutually exclusive) with their frequencies in the dataset, are shown.

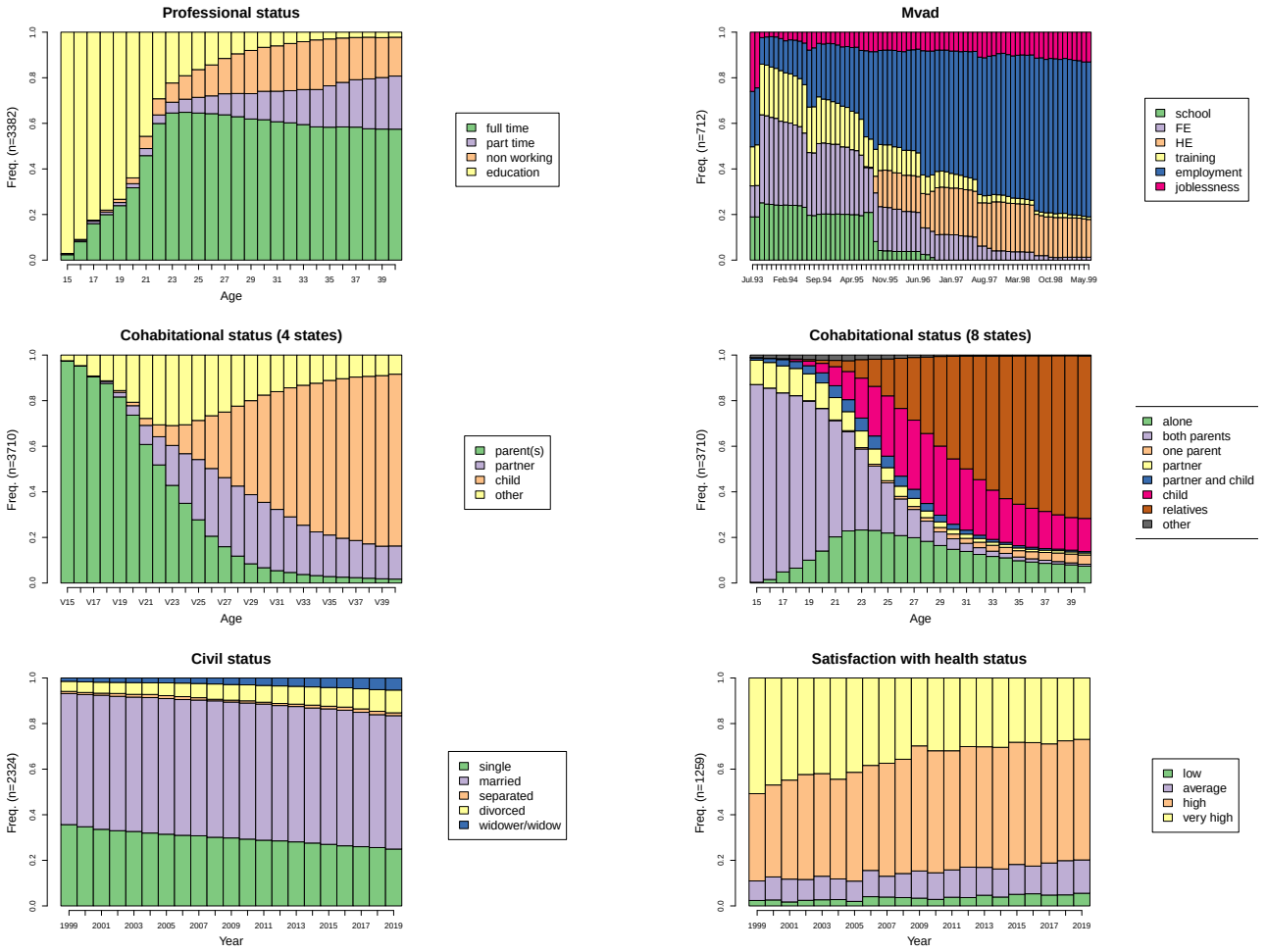



Figure 4.4: Chronograms of the datasets.

We now provide a brief description of how these datasets were constructed, with further details available in Appendix E.5. We constructed three datasets using a retrospective life history survey of the Swiss household panel (SHP) (Voorpostel et al., 2016). They all code trajectories measured in process time and are characterised by strong timing regularities as illustrated by Figure 4.4. Professional and cohabitation trajectories in Switzerland share similar characteristics, except that return to previously visited states is more frequent in professional trajectories. The two coding of cohabitation trajectories, with four and eight states, were taken to illustrate how the coding impacts imputation. These three databases regroup the largest number of cases.

We built two datasets using the prospective survey of the yearly Swiss household panel from 1999 to 2019. These trajectories are measured on calendar time with much weaker timing regularities than the previous ones. The first database encodes civil status over time, while the second focuses on health status satisfaction. These trajectories differ by their general transition rates. While civil status trajectories are highly stable, health satisfaction is the most volatile

(see the percentage of transition displayed in Table 4.1).

Finally, we also considered the MVAD dataset (McVicar and Anyadike-Danes, 2002) coding trajectories measured monthly. As a result, the sequences are longer and more stable, because transition rates to other states tend to be lower with smaller time units. This dataset is also characterised by a strong timing, with school transitions only occurring during the summer months.

To summarise, we selected six different datasets to represent the diversity of data characteristics encountered in life course research. We aimed to capture differences in timing regularities, overall transition rates, the possibility of visiting previously visited states, time measurement (monthly or yearly data), and coding detail. Although sample size also varies between datasets, this aspect is further investigated in the simulation as it might influence any of the above configurations of data characteristics.

## Missing Data Generation

In this section, we present the models used to generate missing data in the six complete datasets. The goal of these models is to simulate realistic missing data patterns. We used three different models, each representing a common situation in life course research. Each of these models might also differently affect the performances of each algorithm. Concretely, these models simulate a MAR mechanism, attrition, and a MAR mechanism in a small sample size setting.

Rubin (1987) distinguishes three missing data mechanisms. The data might be missing completely at random (MCAR) when no systematic difference exists between observed and missing data. This assumption is generally unrealistic. The data are said to be missing at random (MAR) when observed characteristics explain their missingness. Finally, the data might be missing not at random (MNAR) when missingness arises because of the value itself. While multiple imputation might be used to improve missing data handling for the first two cases, they cannot be used directly for MNAR, where the missing mechanism should be accounted for. In this chapter, we, therefore, focus on the MAR case.

The rest of the presentation is organised as follows. We start by describing the overall simulation framework before presenting each model separately. Finally, we present descriptive statistics of the generated missing data patterns to ensure that the models are performing as intended.

In all models, we start by randomly selecting 60% of the observations, where missing data will be inserted. This step ensures a sufficient number of complete cases in every simulation. Then missing data are incorporated in these selected sequences according to one of the following three models. Finally, the sequence is checked and the missing data generation model is restarted if the sequence has more than 75% of missing data. This procedure ensures that all sequences have enough information. Indeed, sequences with too many missing data points are

generally not included in the analysis in practice.

**MAR model** This model simulates a MAR case, where missing data depends on the previous state and therefore on an observed characteristic.

As already discussed, missing longitudinal data tends to occur by gaps, i.e. successive missing information. To mimic this process, we first simulate the occurrence of the start of a gap, which depends on the previous state. Then, we generate the gap itself in the sequence. More precisely, the probability of starting a gap equals 0.06 for the first time point (where no previous situation is available). For the subsequent time points, the missingness probability depends on the previous state. For some predefined states (see Table E.1 for the list of states), the probability of starting a gap equals 0.20, while it equals 0.03 for the other states.

Once a gap is started, its length is generated by considering that the probability to continue a gap equals 0.66 (and therefore 0.34 to end the gap).

Overall, this simulation aims to evaluate the ability of each algorithm to handle a longitudinal MAR case.

**Attrition model** This model mimics an attrition process, i.e. when individuals stop answering a prospective survey. Attrition then induces missing data to all subsequent waves, and therefore until the end of the sequence. Sequence three of Figure 4.1 is an example of attrition. The individual stopped answering the survey in time 14 and never participated again.

The model works as follows. Starting from the middle of the sequence, the probability of starting attrition depends on the previous state, as with the MAR model. For the sake of simplicity, we used the same predefined states list as before. The probability to stop answering the survey is then set to 0.10 if the previous state is in the list and 0.015 otherwise. These probabilities ensure an overall amount of missing data in the simulations similar to the MAR model.

The attrition model is included in the simulation for two reasons. First, this is a highly frequent missing data pattern in life course research. Second, it might impact each imputation method differently. The *MICT* algorithm will only make use of past information. However, *FCS* uses the previously imputed value of future states. Such a strategy might be hurtful if these first imputation were bad. Finally, *VLMC* should not be impacted by attrition since it only uses past information.

**Small sample** The last set of simulations aims to study the impact of sample size on the performance of each method. A small sample size is expected to impact each imputation method differently. Indeed, the predictive performance of multinomial models is directly linked to the sample size (De Jong et al., 2019). However, random forest works well with small samples (Biau and Scornet, 2016). Second, the *MICT* imputation algorithm uses comparatively more

observations to estimate the imputation model than the *FCS* or the *MICT-timing* algorithm. Therefore, *MICT* may be more robust to a reduction in sample size.

This simulation randomly selects 200 cases from a dataset before generating missing data, according to the MAR model.[2]

The missing data generation models aim to simulate the MAR missing mechanism. The models further aim to document attrition, a typical longitudinal data pattern, and the behaviour for a small sample size. Table 4.2 provides the (average) percentages of complete sequences and the overall (average) percentages of missing data generated with these models. The overall percentages of missing data are relatively consistent among the datasets for the MAR and small sample models. The percentages of complete sequences depend on the length of the sequences, with lower percentages for longer sequences. The process of attrition creates longer gaps of missing data than the two other mechanisms. The percentage of missing data is lower for the professional and health satisfaction datasets because the states with a higher probability of triggering a missing value are rare at the end of the trajectories. However, we kept it as it was for consistency with the first process.

## Evaluation Criteria

Our study aims to evaluate the relative quality of imputation methods for life course research. To do so, we need to evaluate and quantify the quality of the resulting imputations. A first and direct method to do so is to compute the accuracy of the imputation, i.e. the ability of the imputation method to retrieve the original value. We compute it here as the percentage of missing states that are correctly imputed. While this is a good indicator of imputation quality, it also has limitations. We are not interested in predicting the correct value for a specific sequence in a multiple imputation framework. Indeed, in this case, non-random imputation would certainly lead to better results. We want to retrieve our data's key dimensions without missing data biases.

The question is, therefore, what the key dimensions we are interested in when analysing longitudinal data from a life course perspective are. Studer and Ritschard (2016) identify three key aspects of interest based on their review of life course literature. We describe these three aspects and we measure them. The formulae are provided in Appendix E.3.

First, the timing of the process — i.e. when an event or a state occurs — is a central aspect, as the impact of a situation is typically thought to depend on its timing. For instance, experiencing unemployment at 15 or 50 years old typically has very different later-life consequences.

[2]The simulations were also ran with subsamples of size 100 and 500. Since the ranking of the algorithms was mostly the same and for sake of simplicity, we only present here the results showing the most important differences between the algorithms, i.e. with subsample size of 200.

| Process | Dataset | % incomplete sequences | % missing data | mean length gap | % gaps of length 1 |
|---|---|---|---|---|---|
| MAR | Professional | 51.4 | 10.0 | 2.8 | 35.2 |
| | Cohabitational (4 states) | 55.4 | 12.8 | 2.6 | 37.6 |
| | Cohabitational (8 states) | 51.8 | 11.1 | 2.8 | 36.1 |
| | Civil status | 42.1 | 10.6 | 2.6 | 37.3 |
| | Health Satisfaction | 38.3 | 7.3 | 2.6 | 37.4 |
| | MVAD | 58.8 | 11.5 | 2.9 | 34.6 |
| Attrition | Professional | 14.8 | 4.8 | 8.5 | 5.3 |
| | Cohabitational (4 states) | 40.3 | 12.9 | 8.4 | 4.5 |
| | Cohabitational (8 states) | 29.6 | 10.7 | 9.4 | 3.3 |
| | Civil status | 23.6 | 8.4 | 7.5 | 5.2 |
| | Health satisfaction | 17.6 | 5.7 | 6.9 | 7.1 |
| | MVAD | 39.1 | 14.2 | 26.1 | 0.9 |
| Small sample | Professional | 51.3 | 10 | 2.8 | 35.5 |
| | Cohabitational (4 states) | 55.4 | 12.9 | 2.7 | 37.2 |
| | Cohabitational (8 states) | 51.8 | 11.1 | 2.8 | 36.2 |
| | Civil status | 41.9 | 10.6 | 2.7 | 37.3 |
| | Health satisfaction | 38.2 | 7.4 | 2.6 | 36.7 |
| | MVAD | 58.8 | 11.7 | 2.9 | 34.7 |

Table 4.2: Average percentage of incomplete sequences, missing data by dataset, mean length of the gaps of missing data and percentage of gaps of length 1 by dataset and missing data generation process.

As this is a central aspect, we, therefore, would like the imputation to retrieve the original timing dimension of the data. We evaluate this dimension by computing the mean absolute differences of state frequencies between the imputed dataset and the original one at each time point. We call it the *timing* indicator. A high value indicates that the imputation was unable to retrieve the original data's timing dimension. In contrast, a value close to zero is linked with high-quality recovery.

The duration is the second aspect of interest in life course research. It refers to the time spent in each spell, i.e. the length of the consecutive time points in the same state. Duration also has substantial consequences as it typically captures the effect of exposure to a given

situation, such as long unemployment spells. We use a second indicator to measure the ability of imputation methods to recover the *duration* aspect. It is computed as the mean absolute differences between the average spell length in each state computed in the imputed and original datasets. A high value indicates that the average spell lengths are very different (i.e. shorter or longer), while low values indicate good recovery of the duration aspect.

Finally, the sequencing — i.e. the ordering of different states — is the last key aspect of interest in trajectories. It summarises their dynamics, which might also have long-lasting consequences. For instance, the sequencing employed–unemployed is expected to have different consequences than the unemployed–employed ordering. Our last indicator, therefore, aims to measure the ability of imputation methods to recover the overall ordering of states. The *sequencing* indicator is computed using the sum of the absolute differences between the transition matrix computed on the sequences of distinct states, weighted according to the state distribution on the original dataset. A low value on this indicator indicates that the method can reproduce the original transition matrix between states, while a high value indicates that new ordering was inserted.

Summarising, we identified four key indicators to evaluate the quality of the imputations. While *accuracy* measures the quality of predictions, the three others—namely *timing*, *duration* and *sequencing*—aims to measure the ability to recover essential dimensions of the original data in a life course perspective.

Except for the accuracy, the raw values of these indicators are not informative. Therefore, to improve the comparability, the four indicators were standardised. Since the magnitude of the values typically depend on the dataset and the missing data mechanism, this standardisation is made separately for each dataset and simulation model. By using standardised values, we adopt a comparative perspective where the results of each algorithm is compared to the other ones. The detail of the formula is presented in Appendix E.4.

To facilitate the reading of some of the results, we also computed a so-called "Total" indicator by summing the four previously presented indicators. The Cronbach's $\alpha$ computed on these four indicators equals 0.8 justifying this choice. However, we also present the value of the indicator separately when they do not lead to the same conclusion.

All computations were made using the `R` statistical environment. The *mice* (Van Buuren and Groothuis-Oudshoorn, 2011), *seqimpute* (Berchtold et al., 2022) and *PST* (Gabadinho and Ritschard, 2016) packages were used to apply, respectively, the *FCS*, *MICT* and *VLMC* imputation algorithms. The evaluation criteria were computed using the *TraMineR* package (Gabadinho et al., 2011).

## 4.4 Results

We present the results in two stages. We first discuss the parametrisation of each algorithm and identify the best-performing ones. Then, we compare the algorithm with their best performing parametrisations.

### Algorithm parametrisation

In this subsection, we compare the result of the different parametrisation of each algorithm. Since all indicators generally lead to the same conclusion, we present the results using their sum, i.e. our "Total" indicator. To avoid overloading our presentation, the results are presented in Appendix E.5.

#### FCS multinomial

Figure E.1 in the Appendix presents the different results of *FCS* multinomial using the "Total" indicator. The best performing parametrisation depends on the missing data processes. For the *MAR* process, using one predictor both in the past and future works best for the professional and both cohabitational datasets. However, it lags behind other parametrisations in the three other datasets, where using five predictors both in the past and future is best.

In the attrition simulation, the use of future observations diminishes the quality of the imputation. This is an expected result. Future observations are not observed but randomly drawn according to the marginal distribution at the beginning of the imputation process. This information might then mislead the *FCS* multinomial model, which is unable to improve these imputations in the subsequent imputations. As a result, using five predictors in the past generally gives the best results. It lags slightly behind the use of a single predictor in the past for MVAD and both cohabitational status datasets. However, using five past predictors is better on the satisfaction with health status dataset.

Even if it is punctually beaten by using five predictors both in the past and future, using only one predictor in the past feature among the best-performing parametrisation for the small sample simulations. Again, this is expected as multinomial regressions are affected by small sample size and even more when it includes many covariates (De Jong et al., 2019).

To summarise, the best parametrisation of *FCS* multinomial depends on the situation. In presence of attrition process or small sample size, using only one predictor in the past is the most suitable choice, while in the other situations, using one predictor both in the past and future provides better performances.

### FCS random forest

The results for the parametrisation of *FCS* random forest are presented in Figure E.2 of the Appendix.

Using a large number of predictors generally leads to better performances, which can be found with five predictors in the past and future or all the other time points. As with *FCS* multinomial and probably for the same reason, the use of future time points deteriorates the performance, except for the civil status dataset. However, even in this case, using five predictors in the past and future or all observations do not lag far behind. Random forest is thus less impacted by the first random imputations than multinomial regression. As a result, we suggest using *FCS* random forest with all the potential predictors.

### MICT multinomial

Figure E.3 presents the results for *MICT* multinomial. The best results are found when using five predictors both in the past and the future. It provides the best results on the health status dataset, which is the most subject to transitions. However, using only one predictor both in the past and future provide slightly better performance in the small sample simulations.

The performances of *MICT* multinomial with or without the use of future time points in the attrition simulations are identical. This result was expected. Indeed, the *MICT* algorithm only makes use of future time points if some of them are available. This is an interesting feature of the algorithm, as it relieves us from the choice on including or not future time points.

Summarising, we suggest using *MICT* multinomial with five predictors in the past and the future.

### MICT random forest

The results of *MICT* random forest are generally in line with those of *MICT* multinomial according to Figure E.4. The best parametrisation is found with five predictors in the past and future, even if it lags behind on the *small sample size* simulations. However, it clearly outperforms the others for the satisfaction with the health status. Here again, the use of future time points does not change the results in the attrition process.

For all these reasons, we suggest using *MICT* random forest with five predictors in the past and the future.

### MICT-timing multinomial

Figure E.5 presents the results of the *MICT-timing* multinomial algorithm.

As a recall, the *MICT-timing* method extends *MICT* by providing an additional parameter, the radius, controlling the time frame used to estimate the imputation model. We observe

better performance with a smaller radius and one predictor both in the past and future for the datasets with a strong *timing* structure, such as the professional or MVAD datasets, but to a lesser extent with the cohabitation dataset. These differences are stronger in the MAR than in the attrition simulations. Logically, a closer look at the results show that the improvement is particularly important for the timing indicator. In these cases, using only one predictor both in the past and future appears as the best parametrisation.

However, this does not hold true for the small sample simulations and satisfaction with health status. In the case of small sample simulations, reducing the time frame yields poorer results, especially when using five predictors in the past and future. This outcome can be attributed to the imputation model's estimation based on fewer observations with a smaller radius, leading to poor results if the sample size is already low and the number of predictors is high. However, it should be noted that even in the small sample simulation, the best results for the MVAD dataset are found with a radius of zero. For the satisfaction with health status, better results are obtained when increasing the number of predictors, while the radius of the time frame does not have much impact. A trade-off between the sample size, the number of predictors and the timing structure of the dataset is therefore to be made.

For all these reasons, we consider one predictor in the past and the future with a radius of zero, or five predictors in the past and future with a radius of five.

#### MICT-timing random forest

Figure E.6 presents the results of the *MICT-timing* random forest algorithm. Generally speaking, the best parametrisation is found with five predictors in the past and future and a time frame of length five for the MAR and small sample simulations. Even if the results for a single predictor in the past and future are slightly better for the small sample, five predictors in the past and future lies closely behind. However, using a radius of zero slightly improve the results of most attrition simulations.

To summarise, using five predictors both in the past and future and a time frame of length five is the most suitable choice for *MICT-timing* random forest. Even if it is not the best performing parametrisation in all cases, it is close to the best performance.

#### VLMC

In most situations, the results obtained with both gain functions are close (Figure E.7). Since the first criterion is marginally better in some cases, we selected it for comparison with the other algorithms.

## Comparison between the imputation algorithms

We now compare the imputation algorithms using the best parametrisations identified above. The aim is to derive recommendations on the most suitable(s) algorithms in life course research. Contrary to the previous section, we present and discuss the indicators separately (instead of aggregating them) to highlight the differences between the algorithms as well. These results are presented in the four plots of Figure 4.5. The plots presents for each simulation and dataset the boxplots of the standardised value of the indicators. A high-performing algorithm would show the highest values for all simulations, datasets and criteria. Furthermore, such an algorithm would also provide stable performance, and therefore shows little variations among the repetition of the same simulation. This can be identified by looking at the width of the boxplots.

Aside from the best parametrisation of the algorithms, the results include the value of the indicators for complete case analysis, i.e. when considering only fully observed sequences. These results are presented for the three longitudinal criteria, but not for accuracy, as it cannot be computed in this case.

We now turn to the discussion of the simulation results by looking successively at each algorithm.

Complete case analysis is outperformed by any imputation algorithms in most simulations. Unsurprisingly, the timing is the most impacted criterion (Figure 4.5d). Indeed, in all simulations, some states have a higher chance to trigger a missing data gap. These states therefore tend to be under-represented in the remaining complete sequences. The duration (Figure 4.5b) and sequencing (Figure 4.5c) criteria are sometimes higher for complete case analysis than for *FCS* or *VLMC*. However, it only outperforms *MICT* multinomial with the duration criteria in the small sample simulation with the professional status dataset. Comparatively, complete case analysis provides the best results in the small sample size simulations, but it generally lags behind imputation methods such as *MICT* multinomial in most simulations.

*MICT* shows very good performance, but with different strengths depending on the imputation model. *MICT* multinomial features among the best-performing algorithms for timing, duration, and sequencing in every scenario. On the other hand, *MICT* random forest is often better in terms of accuracy. The most prominent example can be found in the attrition simulation for the satisfaction with health status. A trade-off between these two aspects is therefore to be made.

*MICT-timing* multinomial improves the results of *MICT* multinomial on the datasets with a timing compound for the MAR and attrition simulations. As expected, the difference is especially marked in terms of timing. However, we observe no improvement for the small sample simulations, except for the MVAD dataset, which has the strongest timing structure. The results for the datasets with a weak timing structures are mixed. *MICT* and *MICT-timing*

multinomial show similar performance for civil status, but *MICT* multinomial performs better on the satisfaction with health status simulations, especially in terms of accuracy.

*FCS* random forest performs poorly when compared to the other in almost all criteria and considered scenarios. Even if it is sometimes close to the best-performing algorithms, such as in the attrition simulation based on professional status, it is way worse in the others, such as the small sample missing mechanism applied on the civil status or the attrition of MVAD. As with the other algorithms based on random forest, it is better to predict the individual values accurately than to keep the longitudinal consistency.

As a recall, the best parametrisation was not homogeneous for the *FCS* multinomial. We therefore kept two parametrisations: one predictor in the past and one predictor both in the past and future. The conclusions differ according to the missing data generation process. First, both parametrisations lag behind *MICT* imputation methods in the small sample size simulations. In the attrition simulations, *FCS* multinomial using one predictor in the past features among the best performing algorithms, except for the satisfaction with health status. It even beats *MICT* multinomial on the MVAD dataset and the cohabitational status coded as eight states (but not *MICT-timing* multinomial). As identified before, using one predictor in the past and future is not working well in the attrition simulations. Finally, in the MAR simulations, using one predictor both in the past and future features among the best algorithms on the professional status and the two cohabitational status datasets. However, on the three other datasets, *MICT* multinomial performs better. In summary, except for the attrition process, on which it shows similar performance, multinomial *FCS* is outperformed by multinomial *MICT* and *MICT- timing*.

Finally, *VLMC* features among the best algorithms in the attrition simulations. In the two other simulations, *VLMC* works well for imputing satisfaction with health status but can also be among the worst algorithms (e.g. the MAR missing data process applied to the professional status or both cohabitational status datasets). In the case of attrition, *VLMC* is on par with the other algorithms, as only the past is available. In contrast, on the other missing data generation process, the other algorithm can use future observations, while *VLMC* cannot.

## 4.5 Discussion

In this research, we compared several imputation methods for univariate longitudinal categorical data, and evaluated their best parametrisations. We further proposed two extensions of the *MICT* algorithm. First, we considered the use of random forest instead of multinomial imputation models. Second, we proposed the *MICT-timing* algorithm aiming to improve the imputation in presence of heterogeneous transition rates.

First, the *MICT-timing* algorithm achieved his goal and is a suitable alternative to the

*MICT* algorithm when the trajectories have time-varying transition rates. In the early stages of this study, we also considered including the timing of the trajectories by including the time elapsed since the beginning of the trajectory as a covariate in the *MICT* algorithm. However, no improvement of the results was obtained. For the sake of simplicity, the results were not included in this study. Second, in the examined scenarios, the random forest does not provide any additional benefits compared to the multinomial imputation models. While they are better at predicting the correct value (i.e. accuracy), it is often at the detriment of keeping the longitudinal consistency of trajectories.

This comparison of imputation model was based on six complete real-world datasets, representative of different situations encountered in life course research. We then defined three different missing data simulations. The first mimic a MAR process, where missingness occurs as gaps, i.e. consecutive time points with missing data. In addition, some states were set to have a higher probability to trigger a missing data gaps. This is typically observed in life course research, where some situations, such as unemployment or poor health, have a higher risk to be linked to missingness. The second set of simulations aims to evaluate missing data due to attrition. This is also a very common situation in prospective longitudinal survey. In presence of attrition, only past information is available, which raise specific difficulties for some algorithm. The last simulations aimed to evaluate the performance of each algorithm in presence of (very) small sample sizes. Here again, it impacted each algorithm differently.

From each complete dataset and simulation, we generated 100 datasets with missing data. These simulated missing data were then imputed with the different algorithms, namely *FCS*, *MICT*, *MICT-timing*, both with multinomial and random forest imputation models, and *VLMC*. We further included several parametrisations of each of these algorithms. The algorithms and their parametrisation were then evaluated. We used four criteria for assessing, not only prediction accuracy, but also three key characteristics of trajectories for life courses research, namely timing, duration, and sequencing of successive events.

*MICT* and *MICT-timing* multinomial stand out as the best imputation methods in these simulations. Even if they were not always the best in terms of accuracy, they were the best at producing imputed datasets that had the most similar characteristics in terms of duration, timing, and sequencing to the original datasets. There is a link between the sample size, the degree of time heterogeneity of the dataset and the performance of the *MICT-timing* algorithm. Indeed, when the sample size is sufficiently large, which is the case for the *MAR* and *attrition* simulations, *MICT-timing* multinomial showed better performances on datasets that have a timing structure, while having slightly lower performance on time-homogeneous datasets. Moreover, the stronger the timing structure is, the stronger the improvement of *MICT-timing* multinomial over *MICT* multinomial. However, by reducing the sample size, the gain provided by *MICT-timing* multinomial over *MICT* multinomial is mitigated by the difference of the sub-

samples used to fit the imputation models. Indeed, *MICT* uses all configurations similar to the missing data to impute. However, in *MICT-timing* restrict those similar configurations to those occurring within a predefined radius. The estimations of the multinomial regressions are therefore based on a larger number of observations in *MICT* than in *MICT-timing*. Hence, it provides a more robust estimation with a limited number of observations, even if it should be noted that we used a very small sample size of 200. However, when the time heterogeneity is strong, which was the case with the transition from school to work trajectories, *MICT-timing* is better even on small samples.

The use of random forest had mixed results. It showed a tendency, both for *MICT* and *MICT-timing*, to work well in terms of accuracy, but to lie behind the multinomial model for the three other criteria related to the longitudinal characteristics of the trajectories. However, it is worth noting that random forest, and hence *MICT* random forest, is more robust to a reduction of the sample size. Therefore, on small sample sizes, the results are closer to the ones of *MICT* multinomial than on the two other missing data generation process.

The *FCS* algorithm with a multinomial imputation model still features in some of the simulations among the best performing algorithms (e.g. the attrition process or the imputation of the professional status trajectories). However, it has three main drawbacks. First, it performs poorly on datasets with only a few transitions, such as civil status. One possible explanation for this behaviour is that the *FCS* algorithm starts by imputing missing values randomly. It might then have trouble to improve these first imputations in long gaps of successive missing data. Second, it is clearly outperformed by the *MICT* and *MICT-timing* imputation algorithms on small datasets. Finally, the optimal parametrisation of *FCS* differs from simulations to simulations. Even if it can sometimes present good performance, *FCS* random forest is never the best performing one. Moreover, it sometimes shows bad performance, such as with the imputation of the civil status dataset. Therefore, *FCS* appears to be a less efficient imputation algorithm than *MICT* with datasets typical of the life course study.

*VLMC* is suitable for datasets with more transitions and to impute attrition. However, its performances were generally poor on most datasets typical of life course research. In particular, it showed poor performances in the presence of time-heterogeneous transition matrices, even when compared to other algorithms using only the past.

Complete case analysis, i.e. keeping only sequences without missing data, is clearly outperformed by the *MICT* and *MICT-timing* algorithm. Complete case analysis showed especially poor results on the *timing* criteria. Indeed, for the three considered sets of simulations, some states had a higher probability to trigger a missing data gap. These states were then under-represented in the remaining complete sequences. Situations linked with missingness are typically also linked to vulnerability. Restricting to complete cases therefore tends to under-represent precarious situations, which is clearly not desirable. Therefore, complete case analysis

is a poor strategy to deal with missing data.

With respect to the criteria for comparing imputed data, we sometimes observed a trade-off between accuracy and criteria derived from the life course analysis. This highlights the fact that an imputation method cannot be evaluated solely on its ability to perfectly reconstruct missing data values, at least in a longitudinal case.

The six datasets used for our comparison were carefully selected to encompass the variety of situations arising in life course research. Before concluding, we briefly discuss the influence of the characteristics of these situations. First, the included datasets differed according to their timing structure, or in other words, a different degree of time heterogeneity of the transitions between the datasets. As already underlined, the stronger the timing structure is, the better the performances of *MICT-timing* are in comparison with *MICT*. Except with attrition, *VLMC* was also impacted by time heterogeneity. Indeed, gaps of missing data break the sequences into subsequences that are realigned, regardless of their temporal situation to fit the model. With the attrition process, the sequences are potentially reduced in length but never broken into subsequences, hence impacting less *VLMC*. There does not seem to be a clear impact of the timing on *FCS*.

Second, the datasets are distinguished by the characteristics of their transitions. With highly stable trajectories, the imputation is unsurprisingly easier and less information is required for the imputation. *MICT* and *MICT-timing* algorithms are especially better than *FCS*, which has the tendency to create artificial transitions in the trajectories. The imputations were generally more difficult in the more volatile datasets. A detailed look at the imputation also showed that, with all algorithms, impossible transitions were sometimes created, including for instance a transition between the *divorced* and *single, never married* civil status. For *MICT* and *MICT-timing* this is, however, only possible if some of these impossible transitions occur in the original dataset. This highlights the importance of checking the consistency and more specifically the transitions occurring in the original data. Whenever appropriate, these impossible transitions should be corrected *prior* to the imputation. Furthermore, some transitions might be impossible at certain time points. For instance, one cannot transition from school to university at 9 years old. *MICT* and *VLMC* could generate such impossible transitions and situations. Indeed, *MICT* uses every available situation similar to the one to impute, while *VLMC* realigns subsequences broken by gaps of missing data. However, the *MICT-timing*, with a small radius, and *FCS* should not suffer from this limitation. These impossible transitions could be corrected afterwards, but this is burdensome especially for multiple imputation. Furthermore, it might impact the longitudinal consistency of the whole imputed trajectories. The timing structure of the studied trajectories as well as (time-specific) impossible transition are aspects to be taken into account when making a choice about the imputation algorithm. Furthermore, this discussion illustrated the need to clean the data beforehand.

Third, the dataset differed according to the possibility to return to previously visited states as illustrated with difference between professional and cohabitational status trajectories. This data characteristic does not seem to have an impact on the quality of the imputations and, hence, of the algorithms. It should therefore not be taken into account when choosing an algorithm.

Fourth, the detail of the coding varied between the datasets. The cohabitational status was coded either with four or eight states. The aim was to study the impact of the level of detail used to describe the states on the imputation. Logically, a more detailed coding of states results in less precise imputations. Moreover, we observe that the differences between the algorithms are magnified when using a more detailed coding of the states. However, it did not change the ranking of the algorithms. For all these reasons, we recommend restrict to the most required detail level when imputing such process.

Finally, our simulations showed the impact of sample size. The results showed that *MICT* is more efficient on (very) small sample sizes than the other algorithms. *MICT-timing* was still very efficient in the presence of a strong timing structure, and this penalty can be lowered by increasing the radius parameter. A trade-off is therefore required here. However, it should be noted that our simulations tested very small sample sizes and the algorithms were quite robust with the other datasets with sample sizes ranging from 710 to 3710.

### Guidelines

The aim of this chapter is to propose guidelines to impute missing data in longitudinal categorical databases. Our previous discussion highlighted that these guidelines should concern data preparation and the choice of an imputation algorithm.

The data should be carefully prepared *before* imputation. We highlighted two aspects requiring a special attention. First, one should avoid too detailed categories, as it strongly impact the quality of imputations whatever the algorithm. We therefore recommend restricting them to the most important distinctions that are required by the research question.

Second, the longitudinal coherency of the trajectories should be carefully checked and corrected. Indeed, any errors in the data, such as the occurrence of impossible transitions, might be reproduced by the imputation algorithm. It is therefore highly important to correct these errors prior to the imputation.

Once the data have been prepared, an imputation algorithm should be chosen. Our results showed that complete case analysis should not be the default choice, as it is currently the case in most studies (Berchtold, 2019).

The choice of the imputation algorithm should then be ground on the data characteristics and its timing structure.

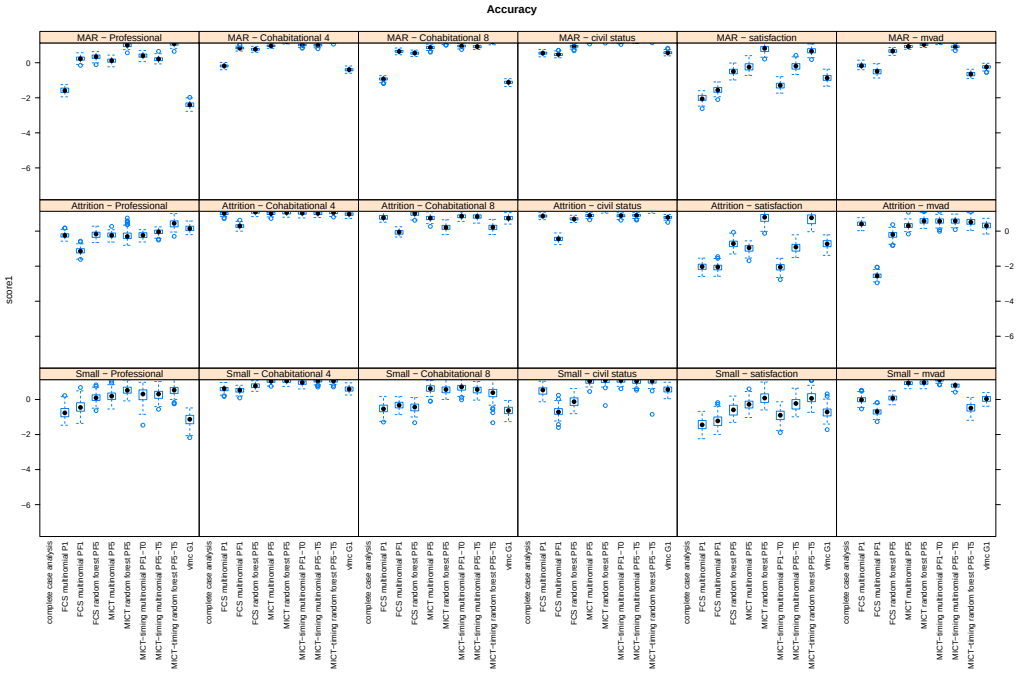
When the data do not have a strong timing structure, *MICT* multinomial provided good

results. In our simulations, using five predictors both in the past and the future provided the best results, but this might be lowered in a small sample.

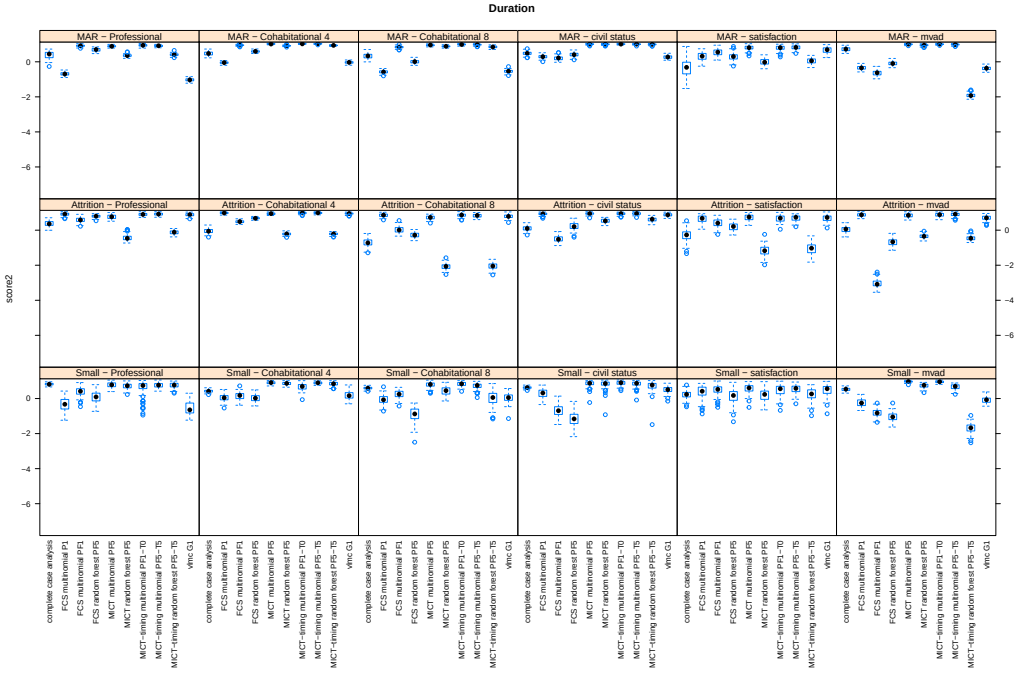
When the dataset has a strong timing structure, *MICT-timing* multinomial should be used. A radius of zero and one predicator in the past and future provided good results in most simulation, but increasing the number of predictors improved some of the results. In presence of small sample size, one should increase the radius or even use *MICT* with very small samples (fewer than 200 observations).

This study involved different several datasets, imputation algorithms and simulations, but it also has limitations. First, we focused here on univariate data. However, as pointed out by Bernardi et al. (2019), the different life domains must be considered as interdependent. For instance, family and professional trajectories of women are strongly intertwined in many countries. These trajectories should therefore probably be jointly considered and imputed. We used three separate missing data generating process in our simulation. However, in practice, they most probably occur altogether, involving for instance attrition and MAR missing data. However, they proved sufficient to reveal some of the differences between the imputation algorithms. Then, we only considered past or future time points in the imputation models. However, taking into account other elements, such as pertinent covariates, could improve the performance of the imputation models. Notably, the trajectories of occupations unfold differently for men and women, as previously mentioned. In particular, adding more information may increase the applicability of random forest as imputation models. Indeed, it handles a large amount of information and non-linear effects well.

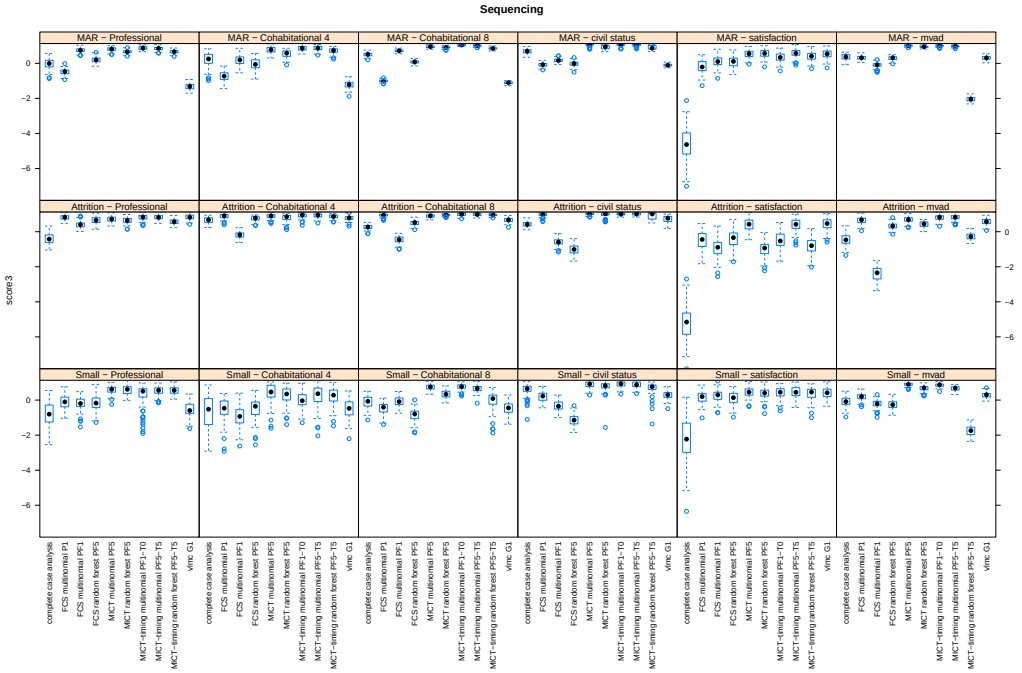
In this study we reviewed algorithms for missing data handling in univariate categorical longitudinal data. However, more work is still required. Indeed, the interrelations between imputation algorithms and the used longitudinal method to be used should also be considered. This includes the creation of typology of trajectories, or the estimation of multistate or hidden Markov models for instance. Second, as already discussed, the presented algorithms should be extended to handle multivariate longitudinal categorical data. *FCS* naturally extends to this case, but *MICT* algorithm cannot handle it, even if it features among the best algorithms for univariate data.



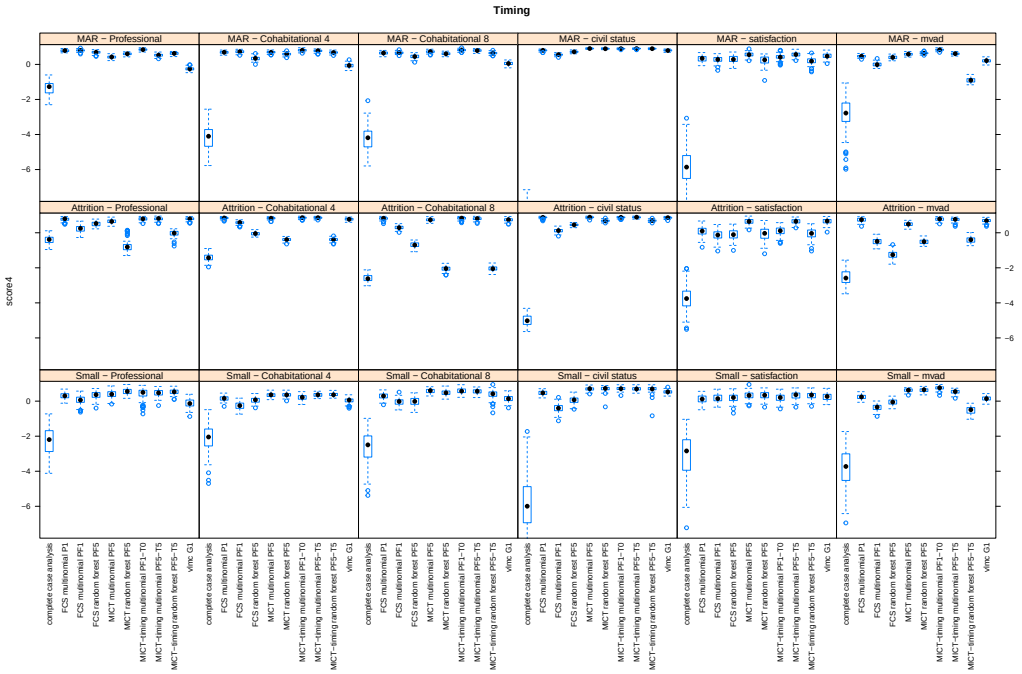

(a) Accuracy




(b) Duration




(c) Sequencing




(d) Timing

Figure 4.5: Simulation results of the best parametrisations of each algorithm using the (a) accuracy, (b) duration, (c) sequencing and (d) timing evaluation criteria. The included parametrisations are complete case analysis, *FCS* multinomial with either only one predictor in the past or one predictor in the past and future, *FCS* random forest with five predictors in the past and future, *MICT* multinomial with five predictors in the past and future, *MICT* random forest with five predictors in the past and future, *MICT-timing* multinomial with a radius of zero and one predictor in the past and future, and radius of five and five predictors in the past and future, *MICT-timing* random forest with a radius of five and five predictors in the past and future, and VLMC with the gain function.

# Chapter 5

# Comparison of two approaches in multichannel sequence analysis using the Swiss household panel[1]

## 5.1 Introduction

Life course analysis is concerned with the many events that punctuate the lives of individuals from birth to death. The focus is often on several supposedly interrelated domains, the idea being that resources, behaviours, and goals in one domain are linked with the resources, behaviours, and goals of other domains (Bernardi et al., 2019). Therefore, to fully understand a given life domain, its linked domains are considered simultaneously. One of the most striking examples is that of work and family, where numerous studies have demonstrated, for example, the impact of the birth of children on women's occupational trajectories (e.g. Piccarreta and Billari, 2007; Widmer and Ritschard, 2009; Aisenbrey and Fasang, 2017). A life domain can encompass different dimensions. For example, the family life domain may encompass a couple dimension, a children dimension and a place dimension.

Sequence analysis, a central tool in the study of life courses, aims to determine their most important features (Abbott, 1995; Ritschard and Studer, 2018a). Possible situations occurring during the life course are represented by a finite set of mutually exclusive states, whose succession over time is called a sequence. These sequences are considered as a whole, with the idea that events cannot be isolated from each other (Piccarreta and Studer, 2019). Sequence analysis has often been applied in the life course study in such areas as transition into adulthood (Oris and Ritschard, 2014; Lorentzen et al., 2019), work pathways (Malin and Wise, 2018; Wahrendorf

[1]This chapter has been published in the same form, with minor changes made only to spelling and grammar, in Longitudinal and Life Course Studies: Emery, K., & Berchtold, A. (2022). Comparison of two approaches in multichannel sequence analysis using the Swiss household panel. Longitudinal and Life Course Studies, 1-32.

et al., 2018), and union trajectories (Jalovaara and Fasang, 2017). A standard sequence analysis is typically conducted by computing the pairwise dissimilarities between the sequences of different individuals before applying clustering to identify a typology (Abbott and Tsay, 2000). Optimal matching, which was first applied to social sciences by Abbott and Forrest (1986), is often used to compute these pairwise dissimilarities. In this framework, the minimum effort to transform one sequence into another through the insertion, deletion, and substitution of states is determined. Since the introduction of optimal matching to social sciences, many variations of optimal matching and other types of dissimilarities have been discussed (Studer and Ritschard, 2016). However, basic optimal matching remains the most often used approach.

Multiple sequences, also called channels, are sometimes considered simultaneously. This happens in broadly two types of situations (Studer, 2015). First, these channels can be related to the same domain. For example, income and labour market positions are two indicators of career trajectories (Mattijssen and Pavlopoulos, 2019). Then, channels from different life domains can be considered simultaneously, either to summarise the association between life domains (Spallek et al., 2014), or to reduce the amount of information before further analysis (Müller et al., 2012). With multiple channels, joint sequence analysis, which is an extension of standard sequence analysis, can be used. Joint sequence analysis involves the computation of dissimilarities based on all channels (Piccarreta, 2017). The two most common strategies are the extended alphabet (EA) approach and multichannel sequence analysis (MSA) (Gauthier et al., 2010). With the former, the states of each channel are combined to build a single set of super-states, called an extended alphabet, each super-state being defined by combining one state from each of the original channels. Pairwise dissimilarities are then computed as if it was a single channel. On the contrary, MSA extends optimal matching to the multidimensional case. Concretely, the substitution cost needed to align two multichannel sequences at a given time point is defined as the mean, possibly weighted, of the substitution costs needed to align each channel separately. Insertion/deletion costs are generally averaged over the different channels. Moreover, these two strategies are combinable: some channels can be first aggregated before applying MSA. Whatever the approach, after the computation of pairwise dissimilarities, a clustering is often used to identify the most typical patterns in the data. However, the application of other sequence analysis tools such as pseudo-ANOVA (Studer et al., 2011) and regression trees (Studer, 2018) is also possible.

Then, channels from different life domains can be considered simultaneously, either to summarise the association between life domains (Spallek et al., 2014), or to reduce the amount of information before further analysis (Müller et al., 2012).

Even if MSA is more commonly used than EA, the latter can also lead to interesting results. For example, considering the extraction of a joint typology of work and family, which is the most common application of a joint analysis, each of the two methods was successfully applied.

Aisenbrey and Fasang (2017) used MSA to compare the interplay of work and family between Germany and the United States, Lacey et al. (2016) applied EA to match work-family sequences to ideal types extracted from theory in order to study the impact of work-family patterns on well-being, while Schwanitz (2017) chose a mixed approach: three channels relative to the family were first combined into an extended alphabet before applying MSA to family and professional sequences.

No in-depth comparison was made between MSA and EA. Gauthier et al. (2010) realised a simulation involving two randomly generated channels and concluded that MSA was the most representative technique. There are also some informal guidelines. For example, PiCCArretta (2017) argued that EA is only sensible if the number of combined states is not too large. However, Lacey et al. (2016) still applied EA with an extended alphabet of size 36. There is therefore no established rule to choose between MSA and EA.

The goal of this article is to compare empirically, via cluster analysis, EA with MSA the idea being to understand the differences in behaviour between EA and MSA, and to determine which method is the most suited depending on the context. To the best of our knowledge, this has never been done before using real data. Concretely, we considered four channels with various interrelations between them in order to have a broad range of situations. Based on the last methodological developments in multichannel analysis, we introduced a framework to compare clusterings and, hence, MSA and EA. The remainder of this article is organised as follows. The empirical dataset is described in Section 5.2 and the methodological tools used to compare EA with MSA are presented in Section 5.3. The results are presented in Section 5.4, and a discussion ends the article.

## 5.2 Data

We used data from the Swiss household panel (Tillmann et al., 2016), a yearly panel study that started in 1999. People living in Switzerland are interviewed on different topics such as family, work, and health. In 2013, an additional sample of 4093 households comprising 9945 individuals was added into the panel. In addition to the standard questionnaires, a retrospective life history calendar was used (Morselli et al.) to investigate life domains such as residential trajectory, living arrangements, partner relationships and changes in civil status, family events, professional activities, and health issues from birth to 2013. Here, we considered sequences of individuals between the ages of 20 and 45 who had answered questions in the domains of professional activities, health issues, living arrangements, and family events for a total of 1707 respondents. Each domain was then summarised in the form of a single channel:

- Child: 0 to 4 years old, if at least one child between the age of 0 and 4 lives in the same household, 5 to 18 years old, if there is at least one child between 5 and 18 but no child

between 0 and 4, and No otherwise.

- Cohabitational status: living with Both parents, living with One parent, living with a Partner, living Alone, and Other situations.
- Professional status: Education, Full-time employment, Part-time employment, and Non-working.
- Health issues: Yes if the person has suffered an illness/accident or undergone surgery or psychological issues during the considered year, and No otherwise.

Cohabitation and child trajectories are expected to be highly interrelated and are often considered in the literature as a single trajectory. They can be seen as two indicators of the family trajectory. The simultaneous analysis of occupational status and family sequences is a typical application of MSA, especially for women. The definition of what a health issue might be is so broad that we do not expect a link between health and any other channel. Even if not linked to other channels, the health one is useful because it allows testing the behaviour of both algorithms when channels are unrelated. Indeed, when channels are not interrelated, algorithms are expected to not produce significant results. Although we did not intend to draw conclusions about the Swiss population, the 1707 respondents were weighted using the sample weights provided by the Swiss household panel, adjusted to match the size of the subsample. This was done to better account for selection bias and allows us to work with a more realistic sample.

## 5.3 Methods

A multichannel analysis consists broadly of three phases. As pointed out by Gauthier et al. (2010), a joint analysis is suitable only if the channels are associated. Therefore, the first step is to check whether their degree of (linear) association is sufficient to justify a joint analysis, even when the channels are supposed to be indicators of the same domain. Then, choices should be made about how the joint analysis is conducted. This includes the choice of the approach (EA vs. MSA), the choice of a dissimilarity measure (i.e. a way of determining how different two sequences are) and the clustering algorithm. Finally, the clustering is performed and a final grouping is chosen. We detail hereafter the tools that are considered at each of the three steps and the choices made for this article.

**Association between channels** Piccarreta (2017) extended the Cronbach's $\alpha$ and principal component analysis (PCA) approaches to determine the degree of associations between channels. Both measures depend on the prior choice of a dissimilarity measure on each individual

channel. Optimal matching is a standard choice for the dissimilarity. However, it also seems sensible to use the dissimilarity measure that will be used for the joint sequence analysis. Consider p dissimilarity matrices $D_1, \ldots, D_p$ containing the pairwise dissimilarities computed on p channels. One can then consider $d_1, \ldots, d_p$, the vectors of the respective upper triangular values of the matrices $D_1, \ldots, D_p$, as p measurements of the same concept. Cronbach's $\alpha$ can then be applied to assess the similarity between the measurements. PCA can also be applied to the set of vectors $d_1, \ldots, d_p$. If the first principal component is enough to summarise the information, then the channels are associated. When more than one component is necessary, the loadings shed light on the relations between the channels.

**Clustering tools** The first choice concerns the method to perform the joint analysis. In this article, we apply both MSA and EA in order to compare them. Then, the way to compute the dissimilarity is under question. When MSA is applied, the standard optimal matching is the most popular choice, either with substitution costs set to 1 (e.g. Pasteels and Mortelmans, 2015; Aeby et al., 2019) or substitution costs determined from the transition rates (e.g. Aisenbrey and Fasang, 2017; Arpino et al., 2018). However, other dissimilarity measures, such as dynamic Hamming distance (McMunn et al., 2015; Sirniö et al., 2017) are punctually used. Concerning EA, the two most common strategies are optimal matching with costs based on transition rates (e.g. Piccarreta and Billari, 2007; Lesnard, 2008) and dynamic Hamming distance (e.g. Eisenberg-Guyot et al., 2020; Lacey et al., 2016). All these dissimilarity measures are variations of the optimal matching. In this context, the minimum cost to transform a sequence into another with substitutions, insertions or deletions of states, is determined. Based on the literature, we apply the following dissimilarity measures:

- Standard optimal matching: The substitution costs are all set at 1 and the insertion/deletion costs are set at 0.5

- Optimal matching with substitution costs based on transition rates: For two states a and b, the substitution cost between these two states is set at 2-p(a,b)-p(b,a), where p(a,b) is the transition rate from a to b in the whole dataset. The insertion/deletion costs are set at half the largest substitution cost.

- Hamming distance: The substitution costs are set at 1 and no insertion, nor deletion are possible.

- Dynamic Hamming distance: As with the Hamming distance, no insertion nor deletion are possible. The substitution costs are time-dependent and based on the transition rates at a given time point. Hamming distance and its variations are more sensitive to differences in terms of timing of the states, while standard optimal matching is less

sensitive to timing differences but is, in return, more sensitive to differences in terms of time spent in each state (Studer and Ritschard, 2016). To derive a typology, Ward Jr (1963) hierarchical algorithm is commonly used with sequences. However, the original algorithm was built for Euclidean distance and there are two different algorithms in the literature that claim to apply Ward's hierarchical algorithm (see Murtagh and Legendre (2014)). In this research, we use the version named "Ward2" in Murtagh and Legendre (2014), which actually applies Ward's clustering criterion, unlike the other algorithm.

**Determining the final clustering** The average silhouette width (ASW) (Rousseeuw, 1987) and Hubert's C index (HC) (Hubert and Levin, 1976) are standard criteria for selecting the number of clusters. For each element, the silhouette value is a comparison between the cohesion of this element in its assigned cluster and its separation from other clusters. The ASW is the mean of these values. It ranges from -1 to 1; the higher the value, the better it is. When the data are weighted to account for selection bias, the weighted version of the ASW (ASWw) is used (Studer, 2013). HC compares the sum of the obtained within-cluster distances with the minimum possible value with the same distance and number of groups. Contrary to the ASW, a smaller value is better for HC. The standard way to determine the optimal number of clusters is to find the partition that gives the best values for the criteria or local extrema. This could lead to several potential partitions and, anyway, with the use of both EA and MSA with different dissimilarity measures, several clusterings are created. Therefore, we identify characteristics that are of interest when studying and comparing clusterings of several channels. We do not start from the idea that there is a "true" clustering of data but rather that there is a clustering allowing extracting a maximum of useful information from data. Our whole strategy of searching for the best clustering tends to identify a clustering that takes into account the association between channels, that makes sense interpretation-wise, whose groups are sufficiently distinct from each other, and sufficiently large so that they are not the simple reflection of extreme and rare situations. Therefore, the criteria of comparison are:

- association between channels taken into account
- channels summarised equally
- channels summarised efficiently
- clusters' separation and homogeneity
- size of the clusters
- shaping aspect
- interpretability and rootedness in theory

We detail the criteria that are used to assess these characteristics.

*Association between channels taken into account*: On the one hand, a clustering provides results even with fully dissociated channels. On the other hand, even with channels that are supposedly interrelated, some clusterings could take accordingly their link into account, while others may not. Studer (2019) extended the work of Hennig and Liao (2010) and Hennig and Lin (2015) to study the behaviour of a clustering quality measure with similar but unstructured sequence data using permutation tests. In the case of multichannel sequences, the clustering quality measures computed by clustering the empirical data are compared to the ones obtained on data generated by a null model, which keeps the structure of the individual channels, but without the association between them. To generate data from this null model, one channel is kept fixed and the others are randomly permuted. A clustering is then applied on this generated dataset and a cluster quality index is computed. This process is repeated many times, usually 1000 times. We can compare the value obtained by clustering the data with that from clustering unstructured data. In this research, we will consider that a clustering takes into account the interrelation between the channels if it is outside the interval containing 95% of the values obtained by clustering the data generated by the null model. We apply the same quality measures as for the selection of the best partitions, namely ASWw and HC.

When more than two channels are involved, this procedure can also be applied to each individual channel in order to determine if the association between this channel and the other ones is taken into account. In this context, the sequences of each channel are fixed, except for those of the considered channel which are randomly permuted.

*Channels summarised equally and efficiently*: The channel-specific R2 (Piccarreta, 2017) gives the share of the total pairwise dissimilarities of a channel explained by a clustering. Therefore, this measure allows us to determine if individual channels are summarised equally and efficiently by a clustering. If the R2 computed on an individual channel is low, this channel is not summarised satisfactorily by the clustering, and if the R2 is unbalanced, the clustering is more driven by some channels than others.

*Clusters' separation and homogeneity*: The ASWw can also be computed independently for each cluster to determine its homogeneity and separation from other clusters. A small value could mean that the cluster is either heterogeneous, not well separated from the other clusters, or both. Size of the clusters: The percentage of sequences classified into each cluster allows spotting clusters that are too small. Shaping aspect: Studer and Ritschard (2016) identified three central aspects structuring sequences, namely timing (i.e. the age of an individual in a specific state), sequencing (i.e. the ordering of the states) and duration (i.e. the overall time spent in a state). We use chronograms to determine which aspect drives the clustering.

*Interpretability and rootedness in theory*: It is crucial that a clustering makes sense from

the point of view of interpretation and that it corresponds to the theory. A clustering can have sound statistical properties but be useless in terms of sociological interpretation (see e.g. Piccarreta and Studer (2019) for such an example). This criterion is more abstract and depends on the analysis that is realised.

Some of these criteria are more important than others. As just stated, the interpretability and rootedness in theory are crucial. In addition, since the goal is to realise a joint analysis, the association between channels should be taken into account. When this is not the case and if the goal is to realise a clustering to reduce the information for a further analysis, one could consider doing it on each channel individually, while when the extraction of a typology is the goal, one may have to accept that it is not possible to extract a joint typology. As for the other criteria, their importance depends on the analysis. For example, having small clusters may not be a problem when the goal is to reduce the information, while it could become a problem when we want to obtain a typology capturing the most common situations in a population.

We compared MSA and EA regarding the ability to produce meaningful results in different contexts. To do so, we first determined if the channels were linearly linked using Cronbach's alpha and PCA. Since we did not have any clear-cut value to decide whether a joint analysis is sensible or not and since neither measure is meant to detect non-linear relationships, we studied a large range of possibilities. We then looked for a clustering of these channels. To do so, we applied both MSA and EA with several dissimilarity measures screened from the literature, namely optimal matching both with unit substitution costs and substitution costs based on transition rates, Hamming distance and dynamic Hamming distance. We compared the results based on the criteria introduced before.

Since at least the professional status trajectories proved to be different between men and women as well as, potentially, their relationship with the other domains, the analyses were run separately by sex. All the computations were performed within the R statistical environment (R Core Team, 2021). The TraMineR (Gabadinho et al., 2011) and WeightedCluster (Studer, 2013) packages were used for most of the analyses.

## 5.4 Results

As pointed out by Levy et al. (2006) and Widmer and Ritschard (2009) among others, the professional status trajectories differ between men and women. An interrelation could exist between the child and professional status trajectories for women since their working rate often decreases when a child is born, whereas the same is barely observed for men. This is confirmed by the chronograms computed separately for women and men (Figure 5.1). Indeed, the professional status trajectories of men are mainly characterised by full-time work, while women are more prone to nonworking and part-time work, and these states seem synchronised with

the arrival of a child in the household. Moreover, women have children slightly earlier than men. These findings motivated us to run all the analyses separately by sex. Detailed results for women are provided hereafter, and results concerning men are provided in Appendix C

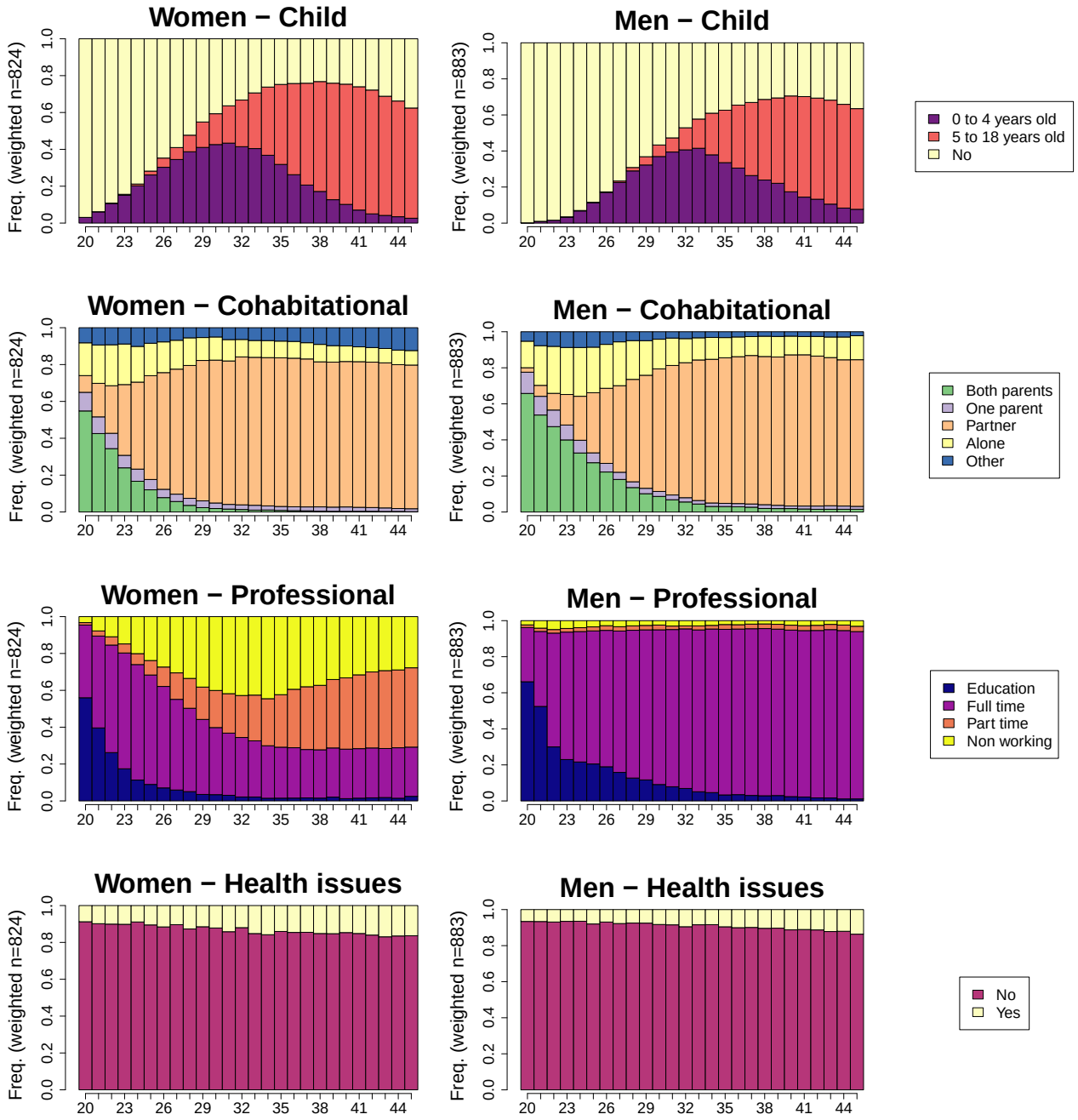



Figure 5.1: Chronograms of the individual domains computed separately by gender.

We first determined the linear association between the channels using Cronbach's $\alpha$ and PCA. We considered the values obtained with each of the dissimilarity measures, but since the conclusions did not vary between them, we only detail the results obtained with optimal

matching with a substitution cost of 1. For women, the health issues trajectories were disconnected from all other channels. Each pair of channels involving health issues had a Cronbach's $\alpha$ smaller than 0.04. Therefore, the health channel is not associated, at least linearly, to the other channels. The conclusion was different for the professional status channel. Although the cohabitational status and child trajectories, which had a Cronbach's $\alpha$ of 0.54, were still the most (linearly) interrelated channels, the professional status channel was somehow linked to them. Taking the three channels together, we obtained a Cronbach's $\alpha$ of 0.51, while the pairs of professional status and, respectively, child and cohabitational status channels gave values of 0.39 and 0.26. These results were confirmed by the results of the PCA.

We still applied clustering to the pair of channels involving the health channel to determine if MSA and EA had the expected behaviour in the case of supposedly unrelated channels, but none of the clustering of pairs of channels involving the health proved satisfactory, neither with MSA nor EA. On the other hand, a joint typology of the cohabitational status, child and professional status could prove suitable. Moreover, we also analysed the two pairs of channels professional status/child and cohabitational status/child. Concerning the cohabitational and professional status trajectories, they seemed mainly linked through the child channel, because individuals are more likely to have a child when living with a partner and, as previously highlighted, working rates of women tends to decrease when a child is born. It was therefore not sensible to extract a joint typology of cohabitational and professional status without taking the child channel into account.

**Child/Cohabitational status/Professional status** We first applied MSA with standard optimal matching and EA with substitution costs based on transition rates, which are the standard parametrisations Figure 5.2 shows that the solution in two groups built by MSA gives the best values both in terms of ASWw and HC, while the solutions in six and eight groups are local extrema. The two-group solution built with MSA (Figure 5.3) is characterised by a first cluster of women having a child, mainly living with a partner, which are more prone to part-time working or non-working, and a second cluster of women not having a child with a variety of living statuses and working mainly full-time. The six-cluster solution is composed of four clusters of women having a child and two clusters of women not having a child (Figure 5.4). Among the women that have a child, the largest group is composed of women that stopped working when they had a child and mostly remained outside the labour force while the child was growing up. Women of the second group had a child earlier than those of the first group and either worked part-time or stopped working when they had a child and mostly worked part-time while the child was growing up, women of the third group mostly continued to work full-time and women of the fourth group had a child later on and were more prone to stay in the labour force. The two remaining clusters are mainly composed of women that did not have

a child, and they differ according to the work status in the forties. The eight-cluster solution splits the two largest clusters, the first one according to the timing of the child birth, and the second mostly according to the cohabitational status (Figure 5.5).

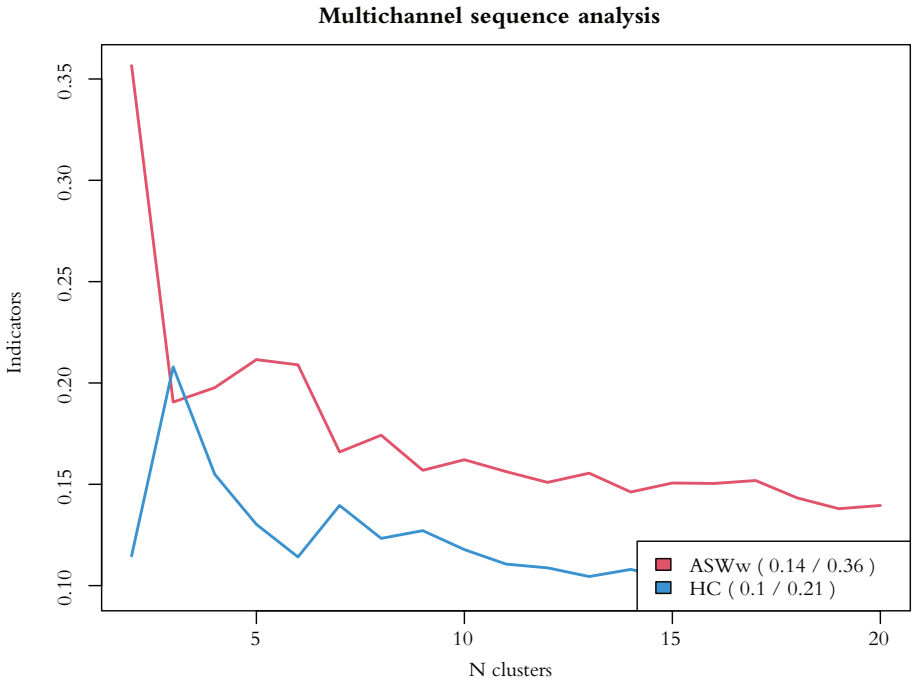



Figure 5.2: Evolution of the ASWw and HC cluster quality indices according to the number of clusters when MSA is applied with standard optimal matching.

We compared these three clusterings based on the criteria we introduced before. The three clusterings are more driven by the child channel, the difference being less acute for clusterings in six and eight groups, though. The two-group solution is composed of two groups that have a relatively good trade-off between separation and homogeneity (ASWw values by group of 0.36 and 0.35), while clusterings in both six and eight groups have some clusters that are not well separated. Based on this analysis, the clustering in eight groups lags behind the two other clusterings. Moreover, even if the clustering in two groups has good statistical properties, it is not interesting from the point of view of the analysis. Therefore, the clustering in six groups appears as the most suitable choice.

For EA with substitution costs based on transition rates, the best value, both in terms of ASWw and HC, is obtained with the thirteen groups clustering and the four groups clustering is a local extremum. The four groups clustering (Figure 5.6) is composed of three relatively homogeneous and well-separated groups and one large residual group containing approximately one third of the sequences. The cluster with the highest ASWw by group is composed of women having a child, living with a partner and that stopped working when the child enters

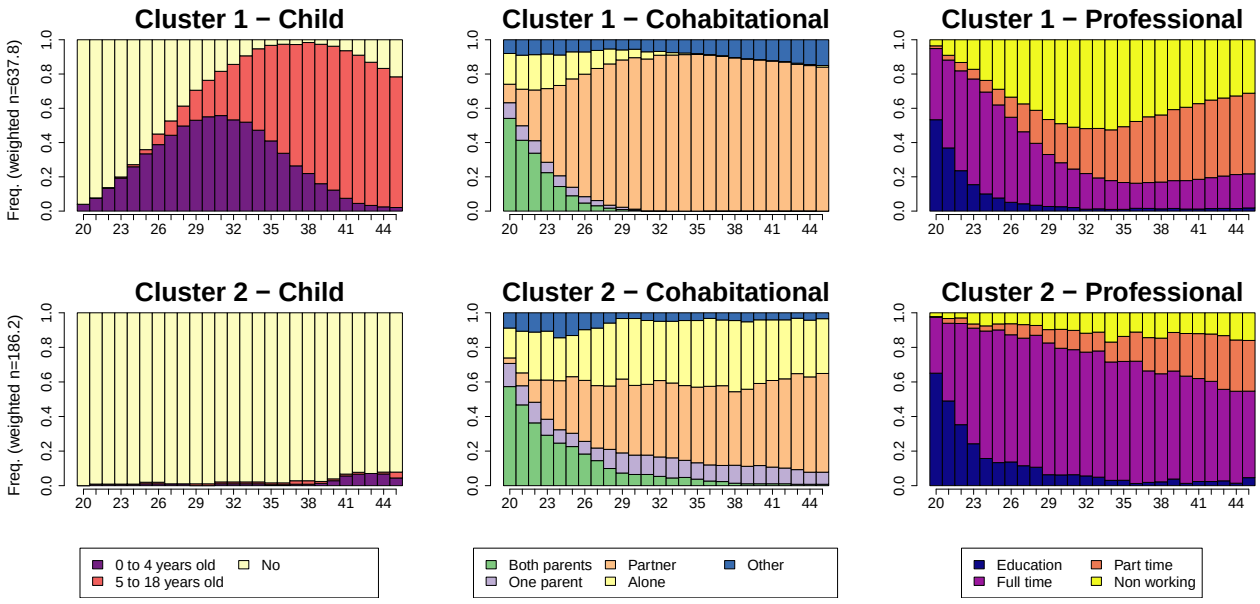



Figure 5.3: Women subset: Chronograms of the two-group typology of child, cohabitational and professional status channels obtained with MSA with standard optimal matching.

the household and remained in this situation the child growing. Women of the second and third clusters differ from the first one in terms of their professional status, they are respectively mainly working full-time and part-time. The third group is slightly less homogeneous because it also contains women that have short spells of non-working time during the child growth, but the ASWw by group is still 0.21. The main issue with this clustering is the fourth cluster that groups every other situation, meaning women that never had a child and women having a child with rarest life courses (e.g. very late child birth or other cohabitational status). This is explainable by the fact that combining the three channels produced an extended alphabet of 60 states (with 44 of them really observed in the dataset) and most of them are rare. The sequences that are composed of these states are dissimilar from each other sequences and are hence isolated. The idea of defining the substitution costs based on transition rates is to induce proximity between states that have transitions between each other. However, in practice, most substitution costs are close to 2. Concerning the thirteen groups clustering, it kept the first two clusters of the four-cluster solution, split the third cluster in three and the fourth one in eight. It therefore distinguishes between many situations, inducing very small groups. Therefore, none of the two clusterings is satisfactory. This example highlights one of the limitations of the EA approach.

Concerning the other dissimilarity measures, a clustering in five groups emerged when MSA was applied with substitution costs based on transition rates. However, it lagged behind, the

clustering in six groups that was obtained with substitution costs set at 1, especially in terms of heterogeneity and interpretability. Hamming-like dissimilarity measures applied with MSA created some clusterings of interest, whose groups were mainly characterised by small differences in terms of timing of the child birth. This could of interest if the researcher is really interested in differences of timing. Results obtained from the application of the other concerned dissimilarity measures with EA lead, as with optimal matching with substitution costs based on transition rates, to either clusterings with a large number of groups or clusterings with a large residual cluster that groups sociologically different situations.

**Child/Professional status** Regardless of the dissimilarity measure used, MSA leads to a small number of groups. Two clusterings stand out according to the criteria. The first one was obtained with standard optimal matching and is composed of four clusters, while the other one is a five-group clustering that was built with Hamming distance. Sharing otherwise similar characteristics, the clustering in five groups is preferable from the point of view of the interpretability, since it involves differences in terms of timing of the child birth. The first cluster was composed of women that mostly did not have a child, the second and third one involved women that stopped working at the child birth and differ on the timing of the child birth, women of the fourth group transitioned to part-time work at child birth and the last group of women stayed in full-time employment (Figure 5.7). The clustering takes into account the association between channels, summarises the individual domains efficiently, and has clusters, that are not too small, relatively well separated and homogeneous. However, it is more driven by the child channel, the R2 being, respectively 0.83 for the child and 0.74 the professional channel.

The use of EA leads to more detailed typologies. According to ASWw and HC, optimal matching with substitution costs based on transition rates leads to an eight groups clustering and optimal matching with unit substitution costs to a seven-group typology. The two clusterings share five almost identical clusters. Women that mostly work full-time before having a child late and switching to part-time work and women not having a child with unstable professional trajectories compose one cluster in the eight-group typology built with optimal matching with substitution costs based on transition rates. As the grouping of these two situations is clearly not desirable, the clustering in seven groups built with standard optimal matching is preferable. This seven-group clustering has no cluster that is completely heterogeneous (the minimum ASWw by group being 0.18), and it summarises the channels efficiently and equally. Moreover, it is interesting since it separates women that stopped working at child birth between the ones that stayed in this situation from the ones that started working part-time the child growing (Figure 5.8). Concerning the two other dissimilarity measures, the results were inconclusive for dynamic Hamming distance and a ten-group stand out with Hamming distance.

Even if it lags slightly behind in terms of statistical characteristics, it may be of interest, depending on the analysis, to have a typology that takes into account differences in the timing of the child birth. In the case of the joint analysis of child and professional trajectories, EA is more suitable.

**Child/Cohabitational status** Applying MSA, the two groups clustering clearly gives the best ASWw and HC values for every dissimilarity measure considered. This clustering, which is almost identical for every dissimilarity measure, separates women according to whether they had a child in their household at some point or not. However, as mentioned before, this clustering is not interesting from the point of view of the analysis. Among the other potentially selected clusterings, very few takes the association between the domains into account. The two clusterings that are the most sensible are the four- (Figure 5.9) and seven-group (Figure 5.10) clusterings built with optimal matching with substitution costs based on transition rates. Even if the seven-group clustering does not clearly take the association between the domains into account since only the ASWw is outside the 95% interval, it captures more situations. In particular, the separation between early, standard and late child birth is of interest.

Concerning EA, the application of optimal matching either with unit substitution costs or substitution costs based on transition rates both leads to an almost identical five-group solution. However, we observe the same behaviour as with the joint analysis of child, cohabitational and professional channel: the fourth cluster contains all the rarest situations. For example, this cluster merges together women that have a child after 40 years old with women that mainly lived with one parent (Figure 5.11). Moreover, it consists of one large group, containing around 70% of the sequences, and four small clusters. Another possibility was the seven-group clustering built with dynamic Hamming distance. It is close to the seven-group clustering selected with MSA, agreeing on the classification of 81% of the sequences. Therefore, both clusterings share similar statistical characteristics. However, the clustering built with EA is less robust in terms of interpretation. Indeed, women that have a child in their forties are grouped with women that do not have a child.

Obtaining clusterings that were similar with two different methods, namely MSA and EA, and two different dissimilarity measures, is a good sign towards the robustness of this clustering in seven groups.

## 5.5 Discussion

In this study, we compared two approaches that allow us to take into account simultaneously sequences from different domains. With the extended alphabet (EA), the multiple sequences are combined in a single one, while multichannel sequence analysis (MSA) is an extension of

optimal matching to multiple sequences. To compare these two approaches in a real context, we used data from the Swiss household panel. Four channels, namely child, cohabitational status, professional status, and health issues, were considered. Moreover, even if the focus was mainly on the comparison between these two approaches, the procedure that we applied may serve as a guide to realise a joint sequence analysis. In particular, we introduce several criteria, based on the last methodological advances, to compare joint clusterings.

For a joint analysis to be sensible, the channels should be associated. Therefore, as a first step, Cronbach's alpha and PCA are applicable to determine the level of (linear) association between the channels. Then, clustering tools should be chosen. It mainly involves three choices: the clustering algorithm, the dissimilarity measure and the method (EA vs. MSA). Ward's hierarchical algorithm is a standard choice with sequences. Then, the choice of the dissimilarity measure has a strong impact on the characteristics that will shape the clustering. Along this line, Studer and Ritschard (2016) give guidelines, in the case of a single channel, on how to choose the right measure for each analysis, especially in terms of sensibility to sequencing, duration and timing. In this study, we observed that their analysis extends smoothly to the case of multiple sequences. Hamming-like dissimilarity measures were more sensitive to differences in terms of timing, while OM-like measures were more sensitive to differences in duration, without being completely insensitive to differences in timing. The use of substitution costs based on transition rates is not very convincing. Although the idea is to create proximity between states, the substitution costs are generally each very close to each other. Moreover, this strategy is theoretically questionable. First, the transition between states is not always synonymous with sociological proximity. Second, the substitution costs may induce violations of metric properties, such as the triangle inequality. Since algorithms for computing the optimal matching distance, such as Needleman and Wunsch (1970), assume that the metric properties are satisfied by the costs, extended alphabet with substitution costs derived from the transition rates should not be used in this case (see also Elzinga and Studer (2015) for a thorough explanation of the importance of the triangle inequality). This issue is more salient when the alphabet is large and, therefore, when EA is used. Hence, we argue against the use of substitution costs based on transition rates.

The choice of the method is the key aspect of our study. Results show that neither of the two approaches is obviously superior. This is not surprising since there is generally no absolute truth when examining a typology based on real data. However, we observe differences in behaviour between these two methods and situations where one or the other method appears more appropriate. The main difference between the two methods is that EA considers each state to be different from the others, whereas with MSA, substitution costs are lower if there is a state in common. For example, considering the case of professional and child channels, the states 0 to 4 years old/Non working and 0 to 4 years old/Full time are considered as two

completely different states by EA, but the difference is less with MSA, since the difference is then only due to the professional channel. Therefore, EA seems useful in mainly three situations. First, when states of the combined alphabet cover different sociological realities, even if they share an individual state in common. Indeed, EA will have a tendency to group them well in different clusters, while this could be less clear with MSA, since there is a proximity between the states due to the common individual state. Then, if one is especially interested in framing rare situations, EA could be of interest. For example, with the joint analysis of child and cohabitational channels, the chosen clustering has two very small clusters of women that never had a child and who, respectively, worked mostly part-time or not at all. Finally, EA is more flexible regarding the dissimilarity measures available, when only optimal matching-like dissimilarities are available with MSA. In all other situations, MSA seems more adapted.

This research suffers from some limitations. By selecting only complete sequences without any missing data, we based our analyses on a relatively simple dataset, which could be considered a limitation even though our goal was only to compare MSA and EA. Then, we compared the method on a single dataset. Even if we studied different scenarios (four channels with various interrelations between them and separated analyses for men and women), we cannot claim to have captured all possible situations. Moreover, other choices could have been made regarding the clustering tools, possibly providing alternative conclusions. For instance, even if a large range of alternative dissimilarity measures are available, we focused on dissimilarity measures that are sensitive to differences in terms of timing and duration but not in terms of sequencing (e.g. the order of the states). Then, we used a hierarchical clustering with Ward's linkage, whereas many other clustering algorithms are available. We could have opted for a different linkage (e.g. a complete one), and partitioning around medoids would also have been possible. Finally, there is a wide range of cluster quality indices that have the ability to capture the additional characteristics of a clustering. In particular, the process we applied to determine whether a clustering takes into account the association depends on the cluster quality index chosen. Therefore, it may happen that a clustering quality index is better on the empirical data than on unstructured data, while another is not, depending on the characteristic of a clustering that is captured by the cluster quality indices.

Although we decided to keep only sequences without missing data to avoid an interaction between them and the object of our research, missing data are unavoidable in practice. It would thus be interesting to determine how missing data and the procedures to deal with them interact with the MSA and EA approaches. The two most commonly used strategies to deal with missing data in the case of sequences are to consider missing data as an additional state in the alphabet, or to impute them. With the first strategy, an extended alphabet could become even larger, especially if the missing data in each channel are replaced by a different additional state. For instance, in the case of the child, cohabitational status, and professional status

channels, the extended alphabet would have a size of 120 with missing data as an extra state in each individual channel, while it is 60 without missing data. Moreover, when states are missing in multiple channels, EA takes that into account since it is considered as a different state, while this does not affect MSA markedly since the missing data are substituted into each channel separately. Considering an imputation strategy such as the one proposed by Halpin (2016b), the impact on the results provided by both approaches is less clear. Nevertheless, if multiple imputation is used and if the typology is identified on the basis of a large dataset combining all replications of the multiple imputation instead of working independently on each replication, the size of the extended alphabet could increase greatly. More research is clearly necessary on this point.

To summarise, we found that although the results are sometimes close, the MSA and EA approaches are still two distinct methods. Although MSA is generally easier to use and applies to more situations, EA can sometimes identify original typologies. Hence, it should also be considered when multiple channels are analysed simultaneously. It could also be of interest to combine the two approaches by building an extended alphabet from some channels and then using MSA to combine it with other channels. In this way, it could be possible to control for the risk of a too large extended alphabet.

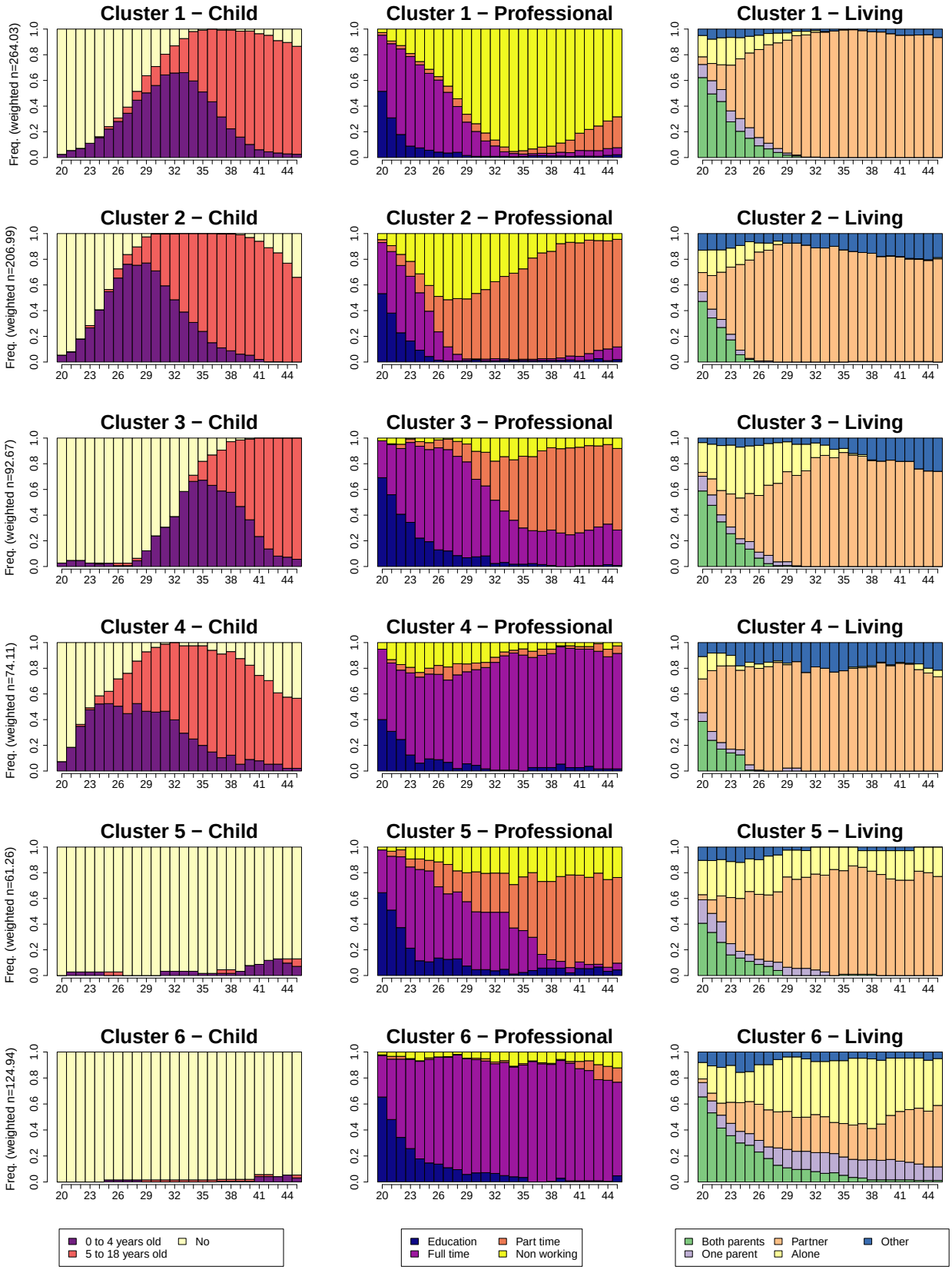



Figure 5.4: Women subset: Chronograms of the six-group typology of child, cohabitational and professional status channels obtained with MSA with standard optimal matching.

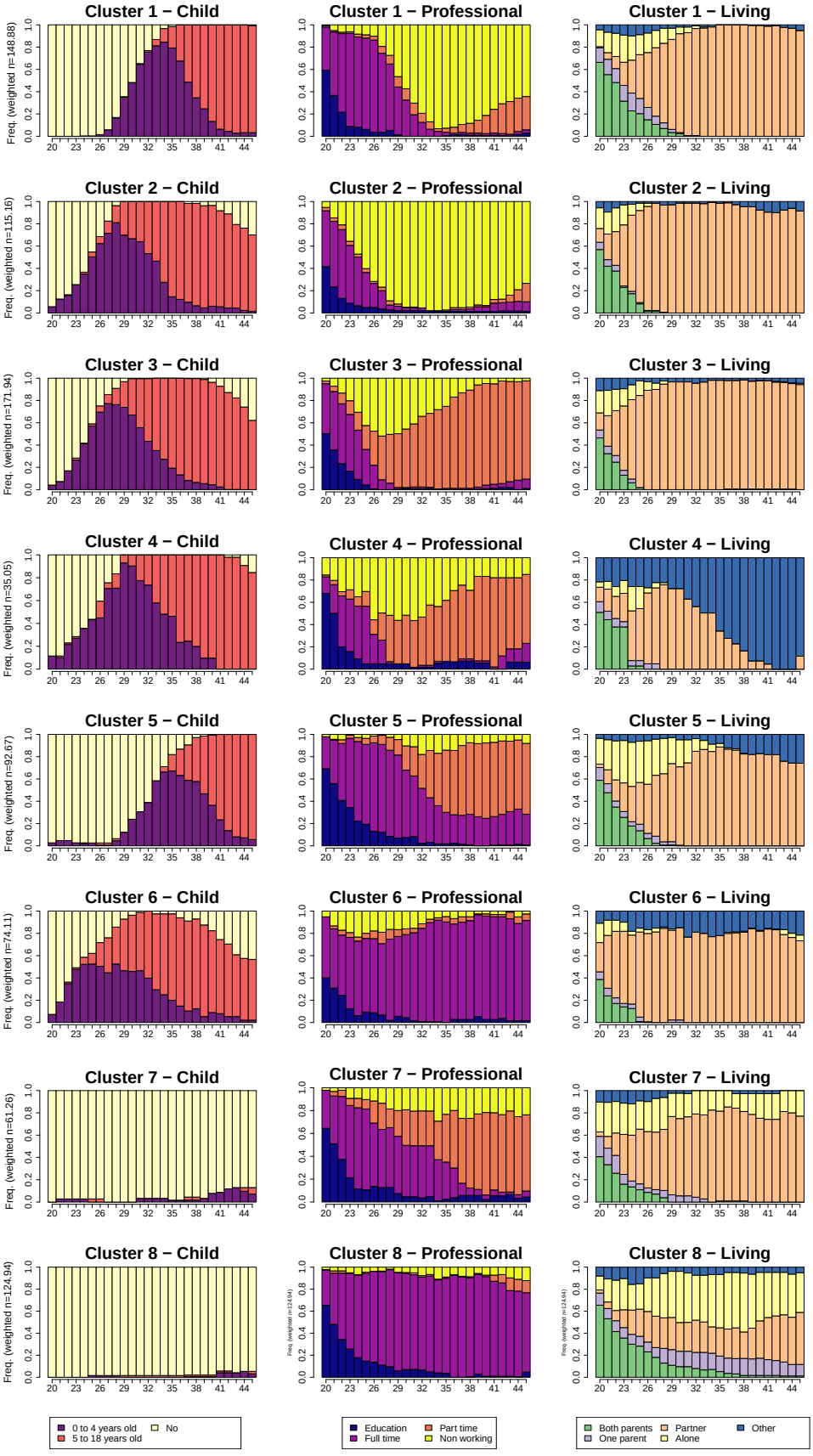



Figure 5.5: Women subset: Chronograms of the eight-group typology of child, cohabitational and professional status channels obtained with MSA with standard optimal matching.

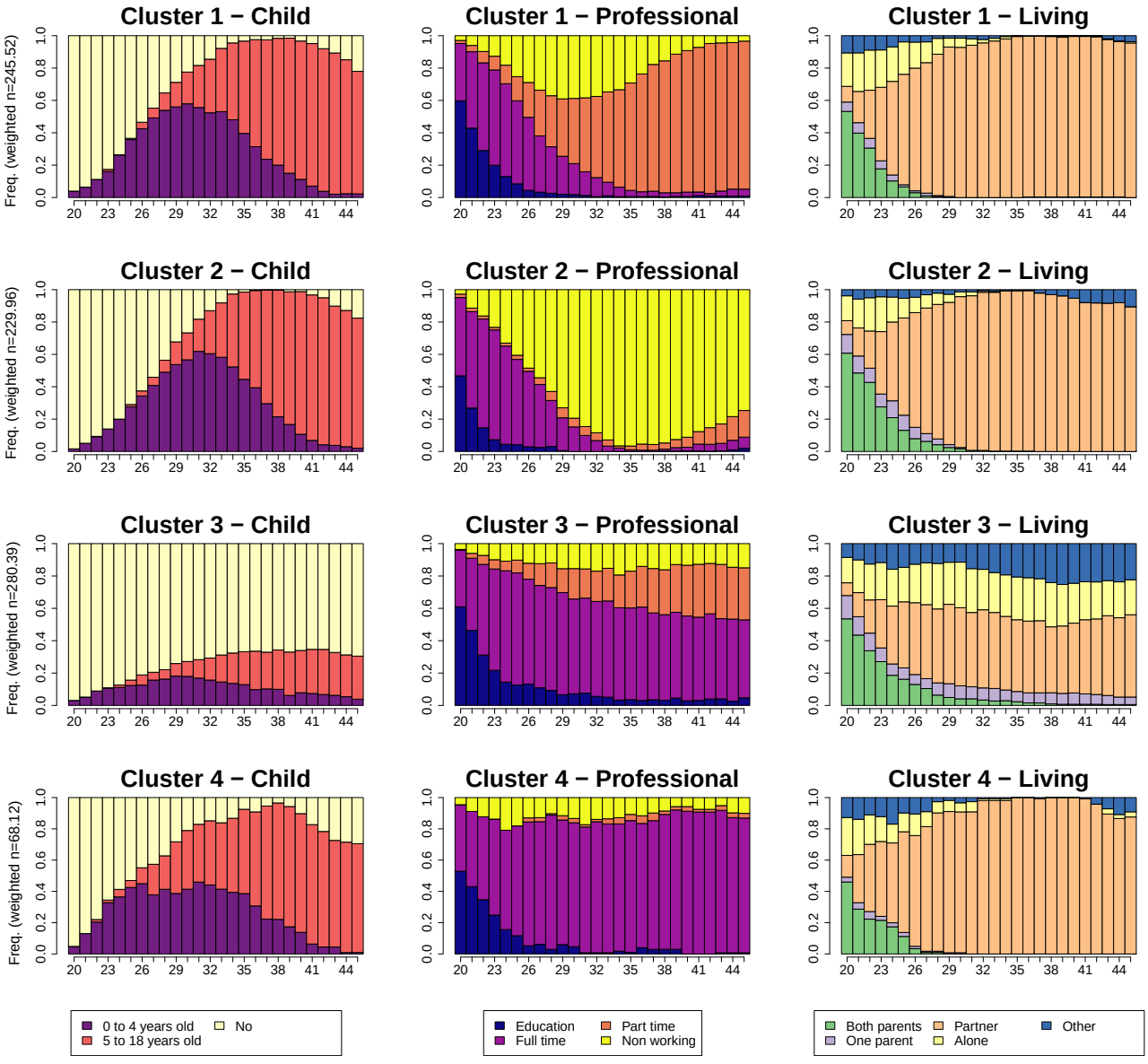



Figure 5.6: Women subset: Chronograms of four-group typology of child, cohabitational and professional status channels obtained with EA with substitution costs based on transition rates.

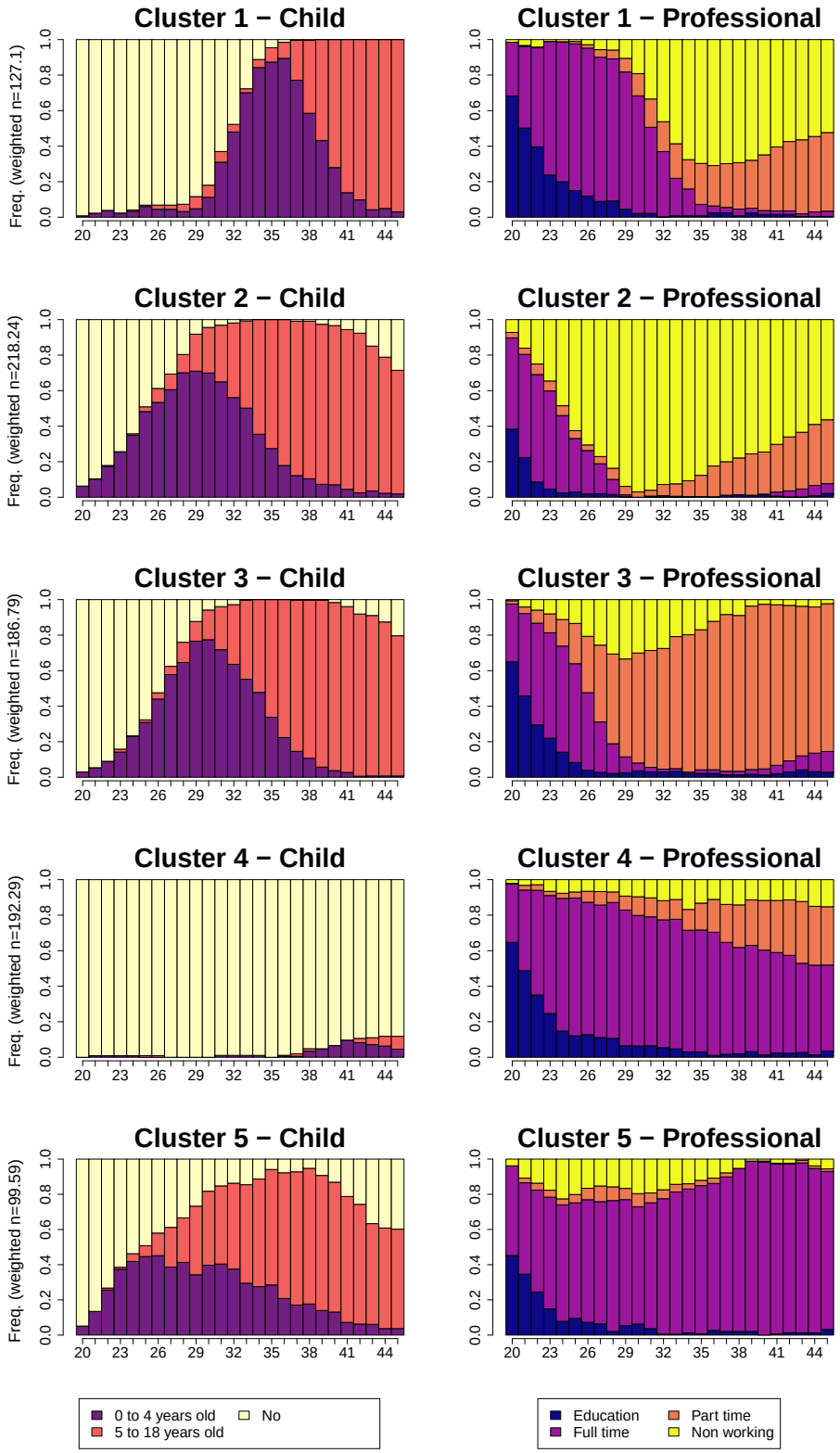



Figure 5.7: Women subset: Chronograms of the five-group typology of child and professional status channels obtained with MSA with Hamming distance.

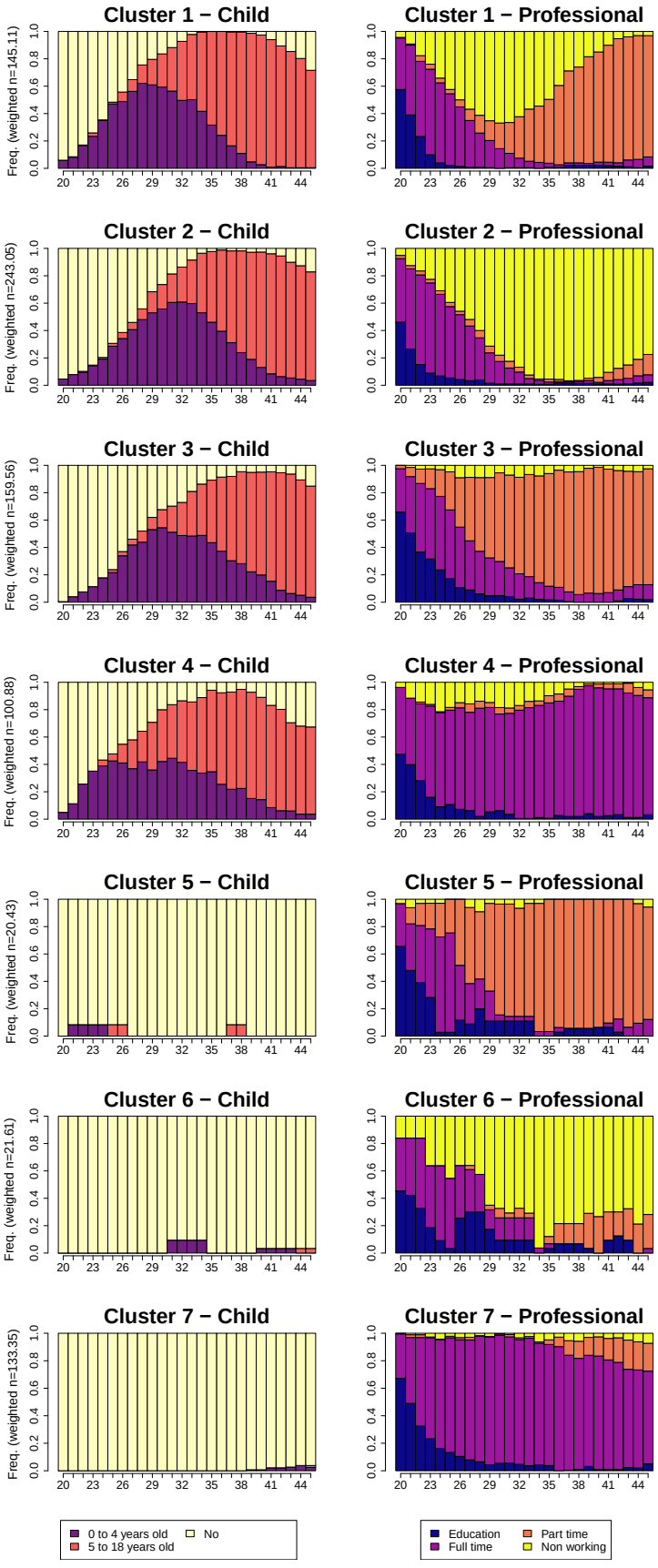



Figure 5.8: Women subset: Chronograms of the seven-group typology of child and professional status channels obtained with EA with standard optimal matching.

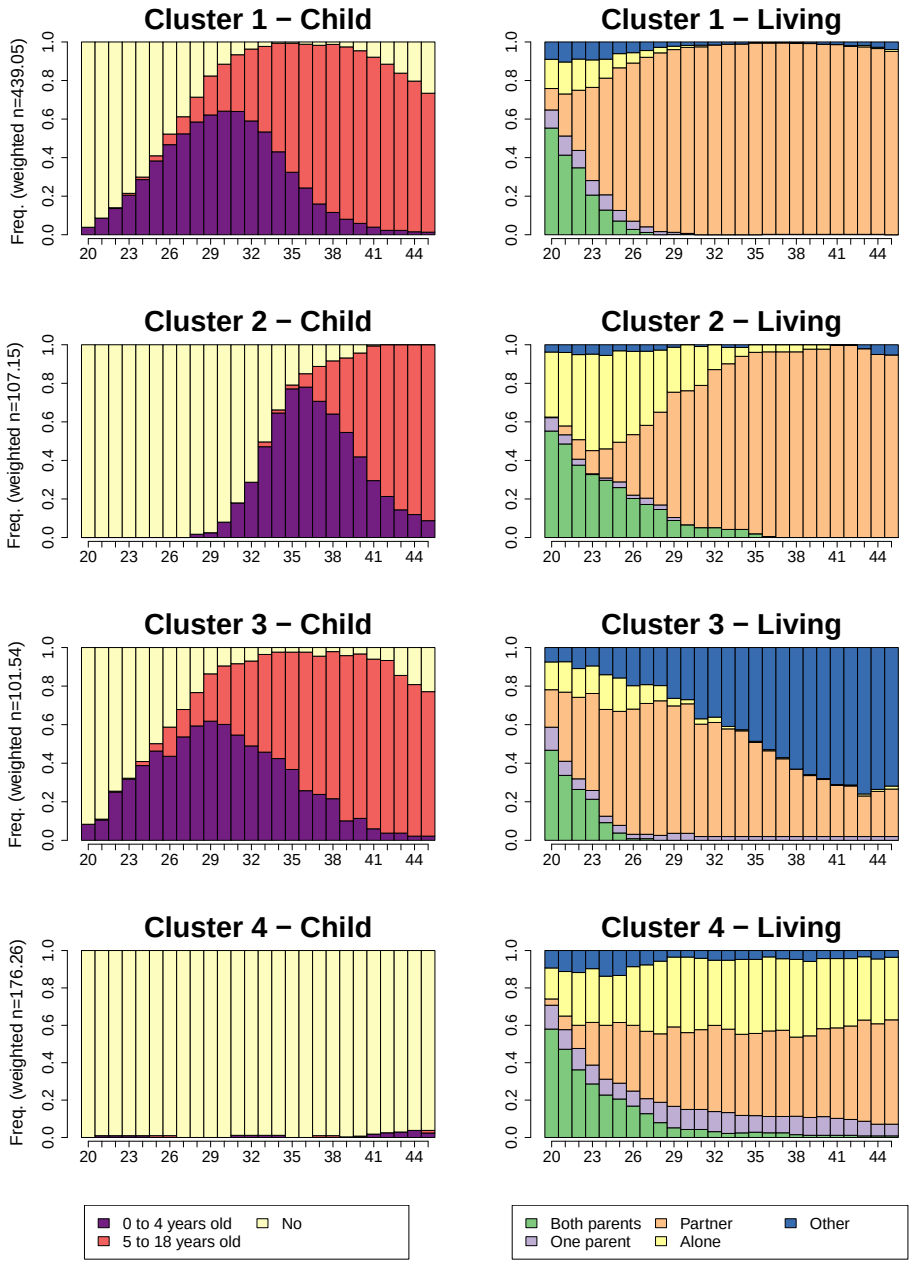



Figure 5.9: Women subset: Chronograms of the four-group typology of child and cohabitational status channels obtained with MSA with substitution costs based on transition rates.

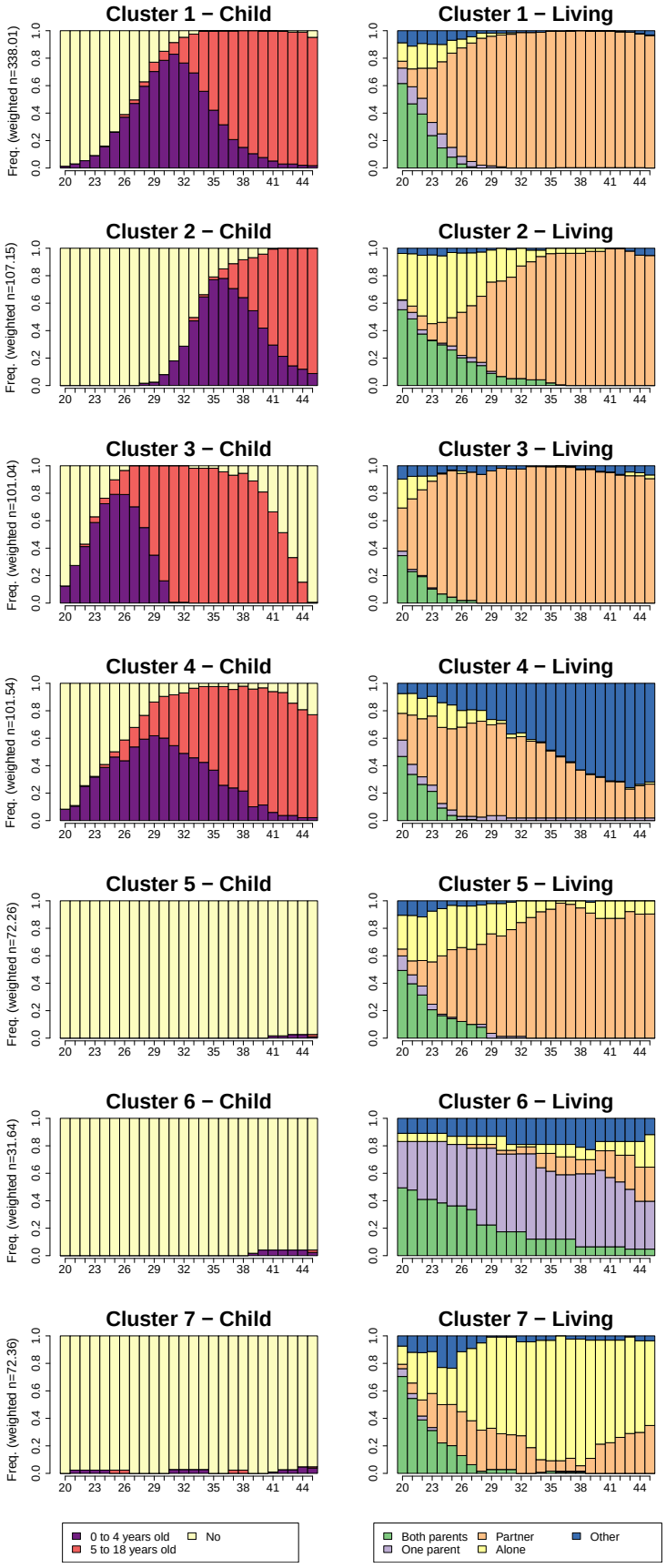



Figure 5.10: Women subset: Chronograms of the seven-group typology of child and cohabitational status channels obtained with MSA with substitution costs based on transition rates.

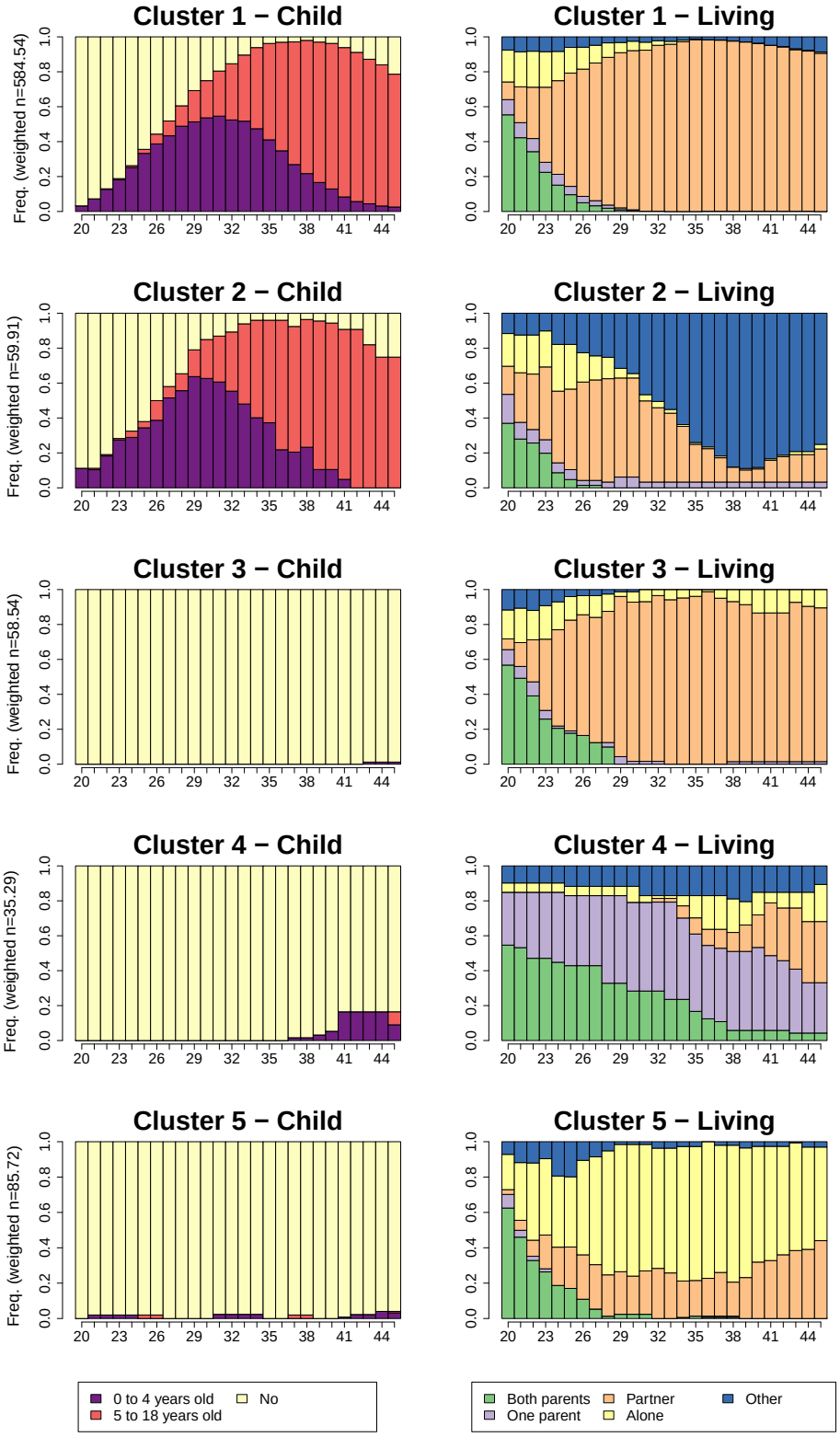



Figure 5.11: Women subset: Chronograms of the five-group typology of child and cohabitational status channels obtained with EA with substitution costs based on transition rates.

# Chapter 6

# Multichannel imputation

## 6.1 Introduction

In chapter 4, MICT and MICT-timing have emerged as the preferred multiple imputation methods for life course data and, more broadly, for categorical longitudinal data. One of the main limitations of both methods is that they are restricted to the imputation of single trajectories. However, in many cases, the focus is on analysing multiple trajectories jointly, such as when studying work and family domains. Moreover, even with a single channel, the imputation quality may improve when other related trajectories are considered. Therefore, the main goal of this chapter is to extend MICT and MICT-timing for the imputation of multichannel sequences. This new algorithm, which we call MICT-multichannel, was compared to other algorithms commonly used for dealing with missing data in a multichannel context. Furthermore, this comparison enabled the derivation of preliminary guidelines for imputing missing data in multichannel sequences, addressing the current lack of such guidelines.

The current approaches available for handling missing data in multichannel sequences do not provide complete satisfaction. The deletion of the trajectories that have missing values is generally unsuitable. First, a single missing value in one of the trajectories leads to the deletion of all the information for an individual, which is wasteful and impact the power of the statistical analysis. Then, when the missing data are not MCAR, which is the most common situation, deleting missing data often induces bias. Weighting may be used to reweight the fully observed trajectories potentially reducing bias. However, the sample size would still be smaller, and vulnerable individuals, such as ones with frequent changes in their professional trajectories are more likely to be missing, which cannot be fully corrected even with weighting. Likelihood and Bayesian methods are difficult to apply with categorical data, because in most applications, a Gaussian distribution is fitted even to categorical data (Honaker et al., 2011). This approach is not ideal because Gaussian distributions are suitable for continuous data that can take any value along a range. Categorical variables, on the other hand, represent distinct

categories generally without a natural ordering, making Gaussian distribution ill suited for modelling them accurately.

Regarding multiple imputation methods, fully conditional specification (FCS) can be easily extended to impute multiple trajectories by considering the repeated measurements of each trajectory as distinct variables. However, FCS may sometimes experience convergence issues (Kalaycioglu et al., 2016) and does not really consider the longitudinal nature of the data. Nevalainen et al. (2009) proposed two-fold fully conditional specification (two-fold FCS) to address these issues. Two-fold FCS runs through the variables at a given time point several times before moving on to the next time point. However, even if two-fold FCS better takes the longitudinal characteristic into account, it is still based on the functioning of FCS, which was built for cross-sectional data. As we have seen in the last chapter, FCS lies behind MICT and MICT-timing for the treatment of missing data in life course data. However, while MICT and MICT-timing have emerged as preferred methods for addressing missing data in longitudinal categorical data, they are not well suited for imputing multichannel sequences. Imputing multichannel sequences using these methods would involve treating each channel separately, thereby ignoring the association between the channels.

We extend MICT for the imputation of multichannel sequences and compare its performance with other commonly used algorithms for handling missing data in a multichannel context. We evaluate these algorithms under different scenarios to determine whether MICT-multichannel is always the most effective approach or if another algorithm is more suitable in some cases. In the previous chapter, we observed that MICT-multichannel performs particularly well when the trajectories exhibit minimal transitions. We investigate whether this observation extends to the multichannel case and the MICT-multichannel algorithm. Additionally, we explore the impact of the inter-channel association on the imputation process, as the importance of cross-sectional information varies depending on the level of association between channels. Since MICT-multichannel operates on the same principles as the MICT algorithm, which is designed for the imputation of longitudinal data, it is of interest to study its behaviour as the inter-channel association grows stronger and, hence, a good use of the cross-sectional information becomes crucial.

When comparing multiple imputation methods, a critical question is how to determine which method is superior. One approach is to ensure that the imputed datasets have similar characteristics to the complete dataset since datasets with similar characteristics tend to produce similar statistical results. However, given the impossibility of capturing all the characteristics of the complete dataset, it is also essential to measure the impact on statistical results. In this chapter, we take a comprehensive approach by examining both a wider perspective and a narrower point of view. The wider perspective focuses on the similarity of characteristics between imputed datasets and the complete dataset, while the narrower viewpoint examines

the impact on clustering results. By considering both perspectives, we gain a more thorough understanding of the performance of imputation methods, allowing us to make more informed decisions.

To evaluate the performance of imputation methods, we employed two simulation frameworks, both of which involve the simulation of missing data on complete datasets. These simulations were designed to mimic the types of missing data that can occur in multichannel sequences. The first framework provides a controlled environment to evaluate the performance of the imputation methods under different rates of transition and association between channels. Concretely, we created multichannel sequences by duplicating single channels and permuting varying percentages of sequences from the duplicated channel. Through these permutations, we were able to manipulate the level of association between channels. For this framework, we chose three datasets as a basis for duplication, based on their transition rates, controlling for the reliability of longitudinal predictors during the imputation process.

While the first simulation framework allowed us to identify the strengths and weaknesses of the MICT-multichannel algorithm and the algorithms it is compared with, it has limitations. By artificially creating multichannel trajectories, we approximate the relationships between channels that may be observed in reality, but this approximation may not be perfect. Specifically, clustering these artificially constructed multichannel sequences may not accurately represent the clustering of real-world multichannel data. To address these limitations, we employed a second framework based on a real dataset. This framework allowed us to test the performance of the imputation methods in a more realistic and applicable setting. In particular, we tested the impact of the imputation methods on clustering results. Moreover, we added the standard method applied in sequence analysis, namely considering a missing state as a state itself and considering a missing data as maximally distinct from any other state, including another missing state.

The remainder of this article is as follows. We first describe the MICT-multichannel imputation algorithm as well as the other algorithms that were compared. Then, the two simulation frameworks are introduced. Finally, the results are presented and a discussion ends the chapter.

## 6.2 Algorithms

This section introduces the imputation algorithms that were compared and the parametrisations that were applied. We considered four different imputation algorithms: MICT, MICT-multichannel, FCS and two-fold FCS. We applied the four imputation algorithms with a multinomial imputation model. Indeed, as pointed out in the previous chapter, the random forest model appears neither suitable for FCS nor MICT. Each imputation method was applied in a multiple-imputation way. Ten completed datasets were built each time these algorithms were

applied.

In addition to the described imputation methods, we also employed a complete case analysis as a baseline when comparing the structure of the imputed datasets. This strategy, as previously emphasised, is still commonly used in social sciences (Berchtold, 2019). Furthermore, for the comparison of clustering results, two ad hoc strategies tailored to sequence analysis were included. The first strategy treats missing data as a separate state, allowing for the inclusion of trajectories with missing data. However, this approach introduces unwanted similarities between trajectories with missing data in the same locations. The second strategy, which considers missing data as maximally different from any other state, is less commonly used. However, as discussed in the thesis introduction, it does not have the same drawbacks as the first strategy and could serve as an interesting alternative.

### MICT

The MICT imputation algorithm (Halpin, 2012, 2013, 2016b) was designed to specifically impute missing data in a longitudinal setting. It fills gaps of missing data recursively from the edges. The algorithm was already detailed twice in this thesis (subsection 2.7.4 and section 4.2), and therefore, we do not provide an extensive description here.

In the case of multichannel sequences, the algorithm is applied individually to each channel. However, this strategy is not expected to effectively recreate cross-sectional consistency. The purpose of applying the algorithm separately to each channel is primarily to investigate whether the potential improvement in cross-sectional consistency achieved by MICT-multichannel comes at the cost of longitudinal consistency.

Although the parametrisation with five previous and subsequent observations was shown the best in Chapter 4, we applied MICT with only one predictor in the past and future to limit computational burden. In most cases, both parametrisations showed similar results.

### MICT-multichannel

The MICT-multichannel imputation algorithm is an extension of the MICT algorithm to handle missing data in multichannel sequences. Its primary goal is to maintain the functionality of the MICT algorithm by recursively filling in missing data gaps from the edges while ensuring consistency across all channels. We first provide a general description of its idea, followed by a detailed breakdown of its steps, and conclude by providing an example to illustrate its implementation.

In the MICT-multichannel algorithm, to impute a missing value, in addition to previous and subsequent observations (ensuring the longitudinal consistency), at least the corresponding observed (or imputed) values from all the other channels, arising at the same time point to the one to impute are used as predictors (ensuring the cross-sectional consistency). In some cases, it

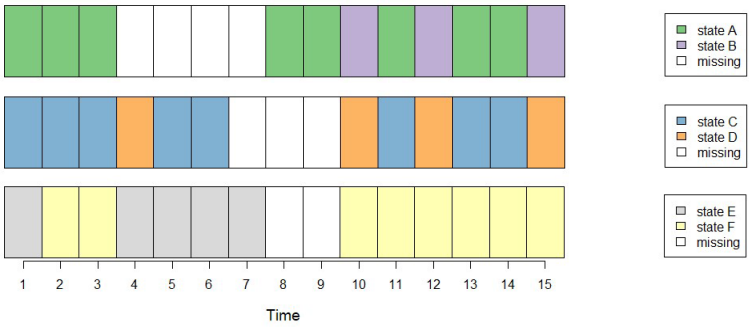
might be needed to add other time points from other channels (like the one just before or after). For instance, when studying both professional and family domains, it is important to consider the previous time point of the family domain when imputing the professional domain, as having a child in the previous year may impact the individual's professional status, particularly for women.

- Initialisation phase
  - The channels are ordered.
  - Each channel is imputed with the *MICT* (or the *MICT-timing*) algorithm.
- Iteration phase
  1. Going through the determined order, each channel is independently imputed with the *MICT* (or the *MICT-timing*) algorithm using as covariates all other channels.
  2. Step 1. is repeated a predefined number of iterations $i$.

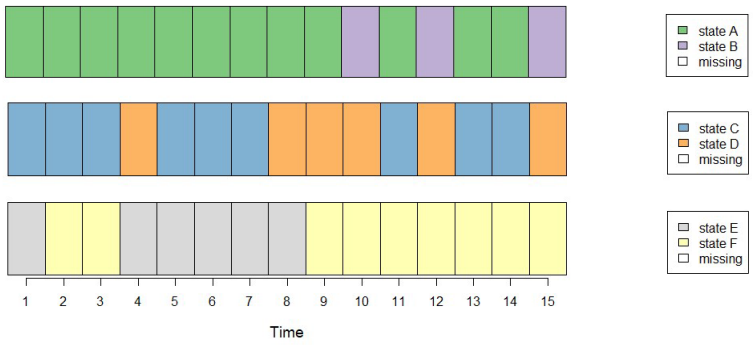
The whole process is repeated a predefined number of times $m$ to produce several completed datasets.

During the initialisation phase, imputing each channel separately may lead to imputations that do not make sense cross-sectionally, requiring several iterations to improve the imputations. Additionally, some channels may be easier to impute due to their stability or lower rate of missing data, resulting in fewer iterations needed when these channels are imputed first. To account for these factors, we allow the user to modify two key parameters: the number of iterations $i$ and the order in which the channels are imputed.

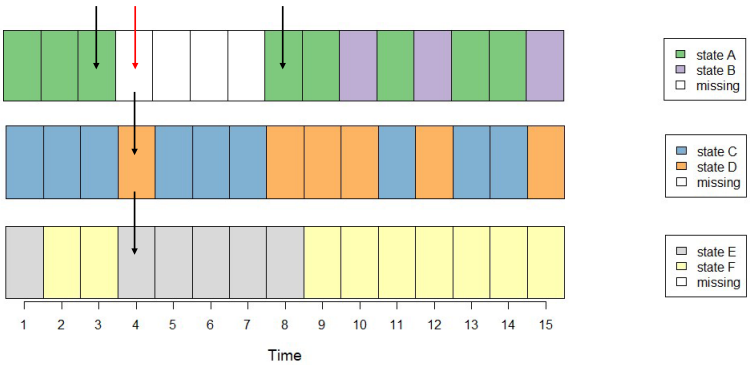
We illustrate the algorithm with an example. It consists of three channels. Each channel can take two states. We consider a minimal imputation model with one predictor in the past and one in the future from the same channel and predictors from the same time point on the other channels. We focus on one multichannel sequence:



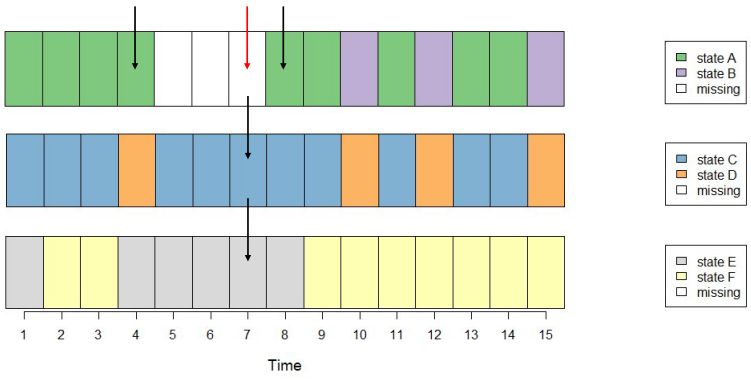
Each channel is first imputed separately with the MICT algorithm. This can lead to the following imputed dataset:



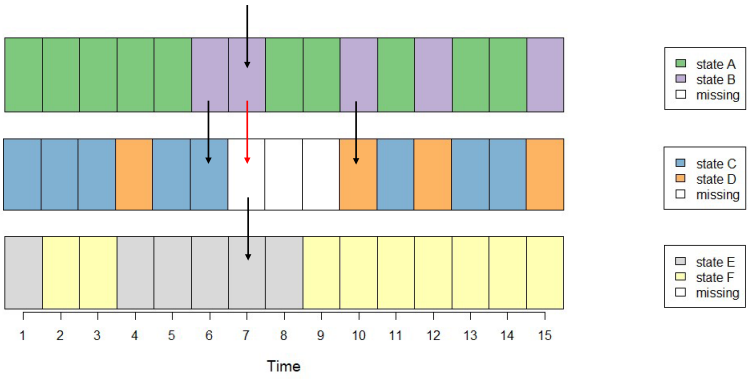

Then, the first channel is imputed again. The core functioning of the MICT algorithm is kept, which means the gaps are still filled recursively from the edges, but information from the other channels is used to keep the cross-sectional consistency. To impute the first missing data of the gap (pointed by a red arrow in the following figure), the previous and subsequent observations of the same channel and the observation on the other channels at the same time point are used (pointed by black arrows on the next Figure).




Then, the last missing value of the gap is imputed using the previously imputed value and the observations arising at the same time point in the two other channels (pointed by black arrows).




Similarly, the two remaining missing values of this gap are imputed. Once the first channel is imputed, the second channel is re-imputed, using the previous and subsequent observations of the same channel and observations at the same time on every other channel. For example, to impute the first missing data of the gap on the second channel (red arrow), the observations pointed out by the black arrows are used as predictors.




Then, the third channel is imputed similarly, ending an iteration. The process of imputing each channel again is repeated a predefined number of iterations.

As a summary, the MICT-multichannel imputation process keeps the core idea of the MICT algorithm that fills gaps of missing data recursively from the edges. It extends to multichannel sequences by integrating cross-sectional predictors during the imputation process.

Based on preliminary testing, it was observed that the algorithm performs well with a low number of iterations. Hence, in this research, we conducted tests using one to three iterations. In the second simulation setting, we tested each permutation of the channels. However, since the individual channels are nearly identical in the first simulation framework, changing the order does not provide significant variations.

Moreover, we base MICT-multichannel solely on MICT and not on MICT-timing. Incorporating MICT-timing would introduce additional complexity due to specific choices tailored

to it, such as determining the time frame radius or deciding which channels should apply MICT-timing versus MICT.

To strike a balance between computational burden and algorithm functionality, we employ a minimal imputation model. This model involves using one predictor from the same channel in both the past and future, along with the corresponding values from all other channels that arise at the same time point.

### FCS

FCS applies to longitudinal data by treating each measurement of each channel as a distinct variable. Detailed descriptions of this algorithm can be found in two sections of this thesis, namely subsection 2.7.4 and section 4.2. Therefore, similar to the MICT algorithm, we do not provide a detailed explanation here.

Generally, the observations from the same time point of all the other channels and some information from the same channels are used (Van Buuren, 2018). In Chapter 4, we observed that using one predictor in the past and future is generally the best parametrisation with general patterns of missing data. Therefore, one predictor in the past and future and information from all other channels at the same time are used to impute a missing value.

### Two-fold FCS

Two-fold FCS is a variation of FCS that involves iterating multiple times through the variables at a specific time point before moving on to the next one (Nevalainen et al., 2009). This algorithm aims to effectively capture the longitudinal nature of the data while addressing convergence challenges encountered by the standard FCS approach. It works as follows for the imputation of longitudinal categorical data:

1. Missing data are first imputed using the marginal distribution of each variable.
2. A distinct imputation model is defined for each variable using a multinomial regression or a random forest for categorical data.
3. The algorithm then runs through each time point.
   (a) For each variable, it fits the imputation model. Missing values are then imputed according to a random draw based on the probabilities predicted by the model.
   (b) The previous operation is repeated a predefined number of iterations $i$
4. Step 3 is repeated until it reaches a predefined number of iterations $j$
5. The values obtained in the last iteration are kept.

6. The process is executed several times if multiple imputations are required.

For its parametrisation, we followed the guidelines of Nevalainen et al. (2009), who suggested using as predictors all observations arising at the same time point and a limited amount of information from the same channel. Therefore, similarly to FCS, in addition to the states observed simultaneously on the other channels, we used one predictor both in the past and future. Moreover, we applied it with $i = 3$ and $j = 10$.

## 6.3 Simulation frameworks

We describe the two simulation frameworks. Firstly, we concentrate on the first simulation framework, where we introduce the samples, the procedures for generating missing data, and the criteria used for comparing the imputation algorithms. Then, we shift our attention to the second simulation framework, which shares the same missing data generation procedures and evaluation criteria as the first one. Therefore, we only describe the sample employed and the clustering-related comparison framework.

### 6.3.1 First simulation framework

The main objective of this simulation framework is to investigate how the strength of cross-sectional and longitudinal information affects the performance of MICT-multichannel relative to the other algorithms. Indeed, as demonstrated in Chapter 4, MICT is particularly well suited to impute trajectories with limited transitions.

To achieve this objective, the first simulation framework involves using a real dataset and creating two identical channels from it. Three real samples were carefully selected based on their transition rates. As the number of transitions in a trajectory increases, the reliability of states within that trajectory diminishes for imputation purposes. Next, a percentage of the sequences in one channel was permuted. The strength of association between the channels weakens as more sequences are swapped and the imputation methods benefit less from the information available in another channel. We explored three distinct percentages for sequence permutations: 20%, 50%, and 80%. In the first case, the association between channels is high, the second case represents an intermediate situation and in the last case, the association is weak. Finally, missing data were simulated on both channels.

We first describe how the samples were built. Then, the different processes of missing data generation are introduced. Finally, the different criteria to evaluate the quality of the imputations are described.

### Samples

Three datasets were selected as a base for the creation of multichannel sequences, namely the satisfaction with health status, the civil status and the professional status of women. Among these three datasets, satisfaction with health status exhibited the highest degree of variability (36.5% of transitions), while civil status demonstrated the lowest rate of transitions (3.6%). The professional status of women represents an intermediate situation with 10% of transitions.

The trajectories for satisfaction with health and civil status were derived from the prospective survey conducted by the Swiss household panel, and were detailed in Chapter 4. On the other hand, the professional status of women was constructed using the retrospective life-history calendar from the Swiss household panel, which was also described in Chapter 4. To diminish the computational burden, we selected professional trajectories of women instead of using the entire dataset.

For each of the three datasets and varying percentages of permuted sequences, we constructed 100 multichannel datasets.

### Missing data generation

The objective of the missing data generation process is to simulate patterns of missing data that closely resemble those observed really in multichannel sequences. We have identified three key variations in missing data patterns within multichannel sequences.

Firstly, the mechanism of missing data can vary, including Missing Completely at Random (MCAR), Missing at Random (MAR), or Missing Not at Random (MNAR). Then, missing data can occur at the same location across all channels or at different locations within each channel. The former case arises when an individual misses a wave of data collection, resulting in missing data in every variable or channel (unit-level missing data). The latter case typically occurs when individuals fail to respond to certain items in a questionnaire, leading to item-level missing data, or when individuals share more or less information about a variable compared to others in a retrospective life course calendar. Moreover, multichannel sequences may have varying degrees of missing data. This variability can be influenced by factors such as the life domain (e.g. health data being more prone to missing data) or the data collection methodology employed.

In the rest of this subsection, we first discuss the impact of these three parameters on the imputation process. Subsequently, we describe how we operationalised these parameters in our study. Finally, we analyse the characteristics of the datasets with generated missing data to gain insights into their specific properties.

In terms of the missing data mechanism, we specifically simulated scenarios involving MAR and MNAR missing data. We excluded MCAR as it is often considered an unrealistic assumption. MAR is commonly assumed when applying multiple imputation techniques. However,

since it is impossible to ascertain whether the missing data are MAR or MNAR, it is valuable to explore how the methods perform when mistakenly assuming MAR while the data are MNAR. Under the MAR mechanism, an imputation process can use observed information to make accurate predictions, resulting in generally unbiased imputations. Conversely, with an MNAR mechanism, missing data depend on unobserved information, leading to bias. Furthermore, we considered two different scenarios for the percentage of missing data. As the proportion of missing data increases, imputation becomes more challenging, and statistical analyses may exhibit increased bias.

Finally, either the same pattern of missing data was applied to each channel or missing data was applied independently to each channel. From the point of view of the treatment of missing data, the differentiation between these two situations is interesting because, in the first case, no information from other channels is available during the imputation process. In contrast, in the second scenario, one or more observations from other channels may be present.

We now focus on the implementation of the three key parameters. With the MAR generation process, a predefined percentage of sequences on which we simulate missing data is first selected. Then, the probability of starting a missing spell depends on the previous state; some predefined states have a higher probability of starting such a missing spell. If the previous state is missing, the probability of missing is high (0.66) to create gaps. The MNAR generation process differs from the MAR in that the probability of missing depends on the current state instead of the previous one. Therefore, the probability of missing depends on information that is not observed any more. We ensured that a given sequence has at most 75% of its time points that are missing.

Then, we considered two different scenarios for the percentage of missing data. For both cases, the probability of missing is 0.06 if it is the first observation of the trajectory, 0.66 if the previous observation is missing, and 0.20 for some predefined states, respectively, the previous one for the MAR process and the current one for the MNAR process. However, it differs both on the percentage of sequences of the dataset on which we simulated missing data, which are, respectively, 60% for the "low" percentage scenario and 80% for the "high" percentage scenario, and on the probability of triggering a missing spell for the other states, which are, respectively 0.03 and 0.05.

Finally, we either generated the same pattern on each channel or generated missing data for each channel individually. For the first, we identified some combination of states between the channels that have a larger probability of triggering missing values (for the MAR process) or being missing (for MNAR process). Therefore, the probability of missing also depends on values in other channels. For the latter, we applied the missing data generation process separately to each channel.

Summarising, the characteristics considered are:

- MAR vs. MNAR missing data

- Percentage of missing data (low vs. high)
- Same pattern for each channel vs. Different pattern for each channel

Table 6.1 provides the average percentages of complete sequences, the overall average percentages of missing data, the average mean length gap and the average percentages of gaps of length one.

Since the probability of pursuing a gap is the same in all scenarios, the percentage of gaps of length one and the average gap length are homogeneous. Regarding the percentage of complete sequence and missing data, on the one hand, the results are almost identical between the MAR and MNAR mechanisms. On the other hand, differences appear between missing data rates, patterns and datasets. First, the percentage of missing data is higher, and the percentage of complete sequences is lower when the scenario with a "high" rate of missing data is applied instead of the "low" one. Then, the percentage of complete sequences is lower with different patterns of missing data than with the same ones, but this is not necessarily the case for the percentage of missing data. The preselection of the sequences subject to missing data was made independently on each channel with different missing data patterns, inducing more multichannel sequences with at least one missing value. Finally, the percentage of complete sequences and missing data depends on the dataset and, more specifically, on the distribution of the states with a higher probability of triggering a missing gap. For example, the Education and Non-working states, which are the states with a higher probability of triggering a missing gap, are widespread in the professional trajectories of women.

| dataset | pattern | rate of missing | mechanism | % complete sequences | % missing data | mean length gap | % gaps of length 1 |
|---|---|---|---|---|---|---|---|
| Civil status | same | low | MAR | 64.6 | 7.1 | 2.6 | 37.7 |
| | | | MNAR | 64.4 | 7.2 | 2.6 | 37.6 |
| | | high | MAR | 41.8 | 12.2 | 2.6 | 37.5 |
| | | | MNAR | 41.6 | 12.3 | 2.6 | 37.5 |
| | different | low | MAR | 35.6 | 10.7 | 2.6 | 37.7 |
| | | | MNAR | 35.6 | 10.6 | 2.7 | 37.1 |
| | | high | MAR | 14.5 | 16.3 | 2.7 | 37.1 |
| | | | MNAR | 14.5 | 16.2 | 2.7 | 37.1 |
| Health satisfaction | same | low | MAR | 58.6 | 8.4 | 2.6 | 37.6 |
| | | | MNAR | 58.3 | 8.5 | 2.6 | 37.6 |
| | | high | MAR | 36.9 | 13.7 | 2.6 | 37.5 |
| | | | MNAR | 36.7 | 13.8 | 2.6 | 37.4 |
| | different | low | MAR | 17.8 | 18.4 | 2.6 | 37.4 |
| | | | MNAR | 17.9 | 18.3 | 2.6 | 37.3 |
| | | high | MAR | 5.0 | 24.9 | 2.6 | 37.2 |
| | | | MNAR | 5.1 | 24.9 | 2.6 | 37.2 |
| Professional | same | low | MAR | 42.2 | 16.8 | 2.7 | 36.7 |
| | | | MNAR | 42.2 | 16.9 | 2.7 | 36.6 |
| | | high | MAR | 22.1 | 23.1 | 2.7 | 36.6 |
| | | | MNAR | 22.1 | 23.3 | 2.7 | 36.7 |
| | different | low | MAR | 24.5 | 11.4 | 2.7 | 36.2 |
| | | | MNAR | 25.3 | 11.2 | 2.7 | 36.1 |
| | | high | MAR | 8.0 | 17.1 | 2.7 | 36.1 |
| | | | MNAR | 8.4 | 16.8 | 2.7 | 36.1 |
| Real case | same | low | MAR | 48.4 | 11.5 | 2.7 | 36.4 |
| | | | MNAR | 48.5 | 11.7 | 2.7 | 36.4 |
| | | high | MAR | 27.3 | 17.2 | 2.7 | 36.3 |
| | | | MNAR | 27.2 | 17.4 | 2.7 | 36.3 |
| | different | low | MAR | 13.4 | 9.7 | 2.8 | 35.5 |
| | | | MNAR | 13.6 | 8.8 | 2.8 | 35.5 |
| | | high | MAR | 2.6 | 14.2 | 2.8 | 35.6 |
| | | | MNAR | 2.6 | 14.2 | 2.8 | 35.7 |

Table 6.1: Average percentage of complete sequences, missing data by dataset, mean length of the gaps of missing data and percentage of gaps of length 1 by dataset and scenario (pattern x rate of missing x mechanism).

### Evaluation criteria

Studying multichannel trajectories within the framework of the life course theory necessitates a focus on three key characteristics: longitudinal characteristics, local association between trajectories, and global association. Studer and Ritschard (2016) screened three longitudinal characteristics of interest when studying a sociological process: the duration, the sequencing and the timing of the process. In addition, the local association between trajectories is crucial in understanding the immediate interconnectedness and dependencies between different life domains or individuals. For instance, research has shown that couples can directly influence each other's health trajectories (Kiecolt-Glaser and Wilson, 2017), highlighting the importance of considering local associations when examining multichannel trajectories. Finally, the global association captures the broader interconnectedness and potential long-term effects between different life domains or individuals. An illustrative example of this is the influence of employment trajectories on later health outcomes (see e.g. Devillanova et al., 2019).

By considering these three characteristics, we can gain a comprehensive understanding of multichannel trajectories. Therefore, an effective imputation process should aim to create completed datasets that preserve these characteristics. We next provide a detailed explanation of how we measured these three characteristics in the context of this research.

**Longitudinal consistency** To assess the data's consistency over time, we used three criteria introduced in the previous chapter. These criteria relate to the states' timing, duration, and sequencing. A suitable imputation method should result in sequences similar to the original in terms of these criteria. With the application of multiple imputation methods, ten completed datasets were generated. As a result, these criteria were computed for each replication, and the mean value was calculated. Therefore, for each criterion, a value of zero means that the treated dataset possesses characteristic regarding this criterion identical to the original dataset. The higher the value, the more distorted this characteristic is. Even if the raw values of these criteria are difficult to interpret, we can compare the values obtained between the algorithms to see the relative gains (or losses) between the imputation methods.

**Local association** This criterion was computed for each pair of channels. It is based on Cramer's V. This measure quantifies the association between two categorical variables.

At each time point, we calculated the absolute difference in terms of Cramer's V between the imputed and original datasets. These absolute differences were then averaged. The resulting criterion ranges from 0 to 1, with values closer to 0 indicating better agreement between the imputed and original datasets.

The distribution of the mean Cramer's V in the complete duplicated datasets is illustrated through boxplots in Figure 6.1. The values are almost identical between the different datasets.

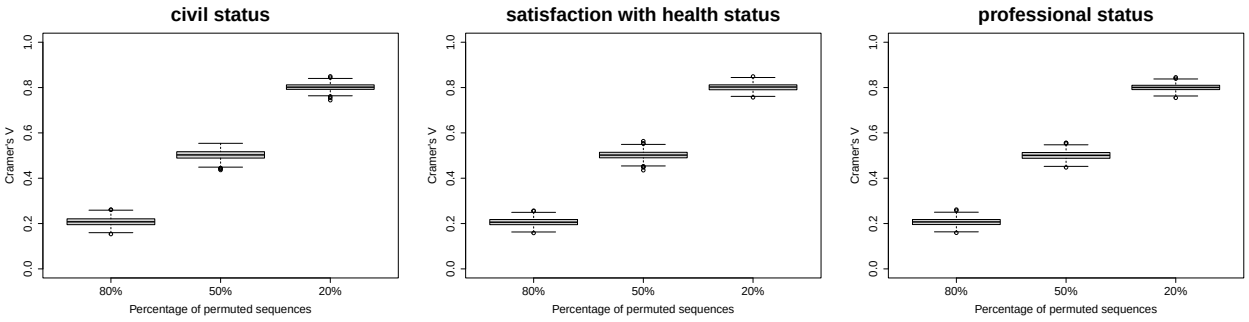



Figure 6.1: Boxplots, by dataset, of the mean Cramer's V for the three percentages of permuted sequences (20%, 50% and 80%). Each boxplot is based on 800 values, which correspond to the 100 generated datasets of the eight scenarios.

The median is around 0.2, 0.5 and 0.8 when 80%, 50% and 20% of the trajectories of the duplicated dataset are permuted. For example, Cramer's V ranges between 0.13 and 0.31, 0.43 and 0.59, and 0.75 and 0.88 for the civil status. Therefore, as expected, we have three different scenarios based on different degrees of associations between the channels.

**Global association** As for the local association, the global association criterion was computed for each pair of channels. It is based on Cronbach's $\alpha$, which measures the global association between channels (Piccarreta, 2017). Cronbach's $\alpha$ works the following way:

- A dissimilarity measure is chosen.
- For each domain, the dissimilarity is computed for each pair of sequences.
- The pairwise dissimilarities are normalised by domain.
- The Cronbach's $\alpha$ is computed as

$$\alpha = \frac{C}{C-1}\left(1 - \frac{\sum_{i=1}^{C} var(\vec{d_i})}{var(\sum_{i=1}^{C} \vec{d_i})}\right), \tag{6.1}$$

  where $\vec{d_i}$ is the vector of the pairwise dissimilarities computed on the $i^{th}$ channel and $C$ is the number of channels.

Even though Cronbach's $\alpha$ can theoretically take negative values, Cronbach's $\alpha$ typically ranges from 0 to 1. A higher value indicates a stronger association between the channels. The computation of Cronbach's $\alpha$ depends on the choice of dissimilarity measure, and in this research, we use the standard optimal matching measure.

The criterion is defined as the difference between Cronbach's $\alpha$ calculated on the imputed dataset and the original dataset. The criterion ranges from -1 to 1, with values closer to zero

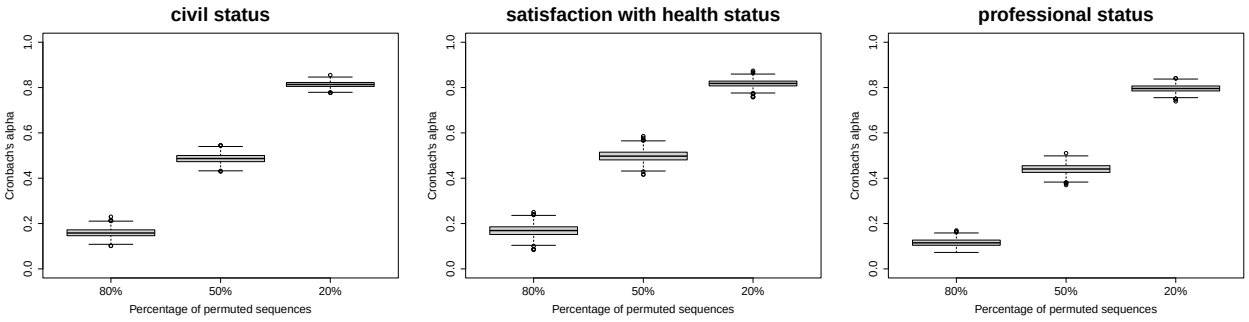



Figure 6.2: Boxplots of the Cronbach's $\alpha$ for the three percentages of permuted sequences (20%, 50% and 80%) Each boxplot is based on 800 values, which correspond to the 100 generated datasets of the eight scenarios.

indicating better agreement. A negative value suggests an underestimation of $\alpha$ on the imputed dataset, while a positive value indicates an overestimation.

Figure 6.2 shows the values computed on the complete duplicated datasets. Cronbach's alpha takes smaller values with the professional status trajectories than the two other datasets. The median values are 0.12, 0.44 and 0.8 on the professional status, while they are 0.16, 0.49 and 0.81 on the civil status and 0.17, 0.50 and 0.82 on the satisfaction with health status.

### 6.3.2 Second simulation framework

The second simulation framework serves two purposes. Firstly, it facilitates the incorporation of a more realistic association between channels compared to the first framework. Secondly, it enables the examination of the effects of missing data handling methods on clustering results. For the latter point, we compared both the imputation methods and two standard strategies tailored to sequence analysis, namely considering a missing state as a state itself or considering a missing state as maximally different from any other states, including another missing state.

We considered one of the most typical cases in joint analysis of multichannel trajectories, specifically the analysis of professional and family domains. To conduct our analysis, we used a dataset which was derived from the retrospective life history calendar employed during the 2013 wave of the Swiss household panel. The sample consists of women who provided responses pertaining to professional activities, living arrangements, and family events, without any missing data between the ages of 20 and 45. This results in a sample size of size 1260. The dataset comprises three channels: child, cohabitational, and professional status.

In the context of joint clustering, we focused solely on the child and professional status domains, as the primary interest lies in exploring the association between these trajectories rather than cohabitational status. However, the inclusion of cohabitational status during the imputation process could be beneficial and potentially enhanced the quality of the imputations, despite not being directly involved in the clustering analysis.

The missing data generation processes and the criteria employed to assess the structure of the imputed dataset obtained through multiple imputations are identical to those in the first simulation framework. Concerning the criteria, Cronbach's alpha is 0.53 between child and cohabitational status, 0.33 between child and professional status and 0.22 between cohabitational and professional status. The mean Cramer's V is 0.36 between child and cohabitational status, 0.34 between child and cohabitational status and 0.23 between cohabitational and professional status.

Only the framework used to study the impact of different methods for handling missing data on the joint clustering of professional and child status remains to be described.

### Impact on clustering results

With multichannel sequences, the goal is often to extract a typology through clustering. We focus here on the clustering of the child and professional status. We applied a hierarchical clustering with Ward's linkage and optimal matching with unit cost to measure the pairwise dissimilarities between sequences. This is the standard procedure in sequence analysis. Moreover, both MSA and EA were considered. We simplified the situation by assuming that the number of groups is known beforehand.

We first describe potential clusterings obtained with the framework developed to compare joint clusterings in Chapter 5. Then, we detail how clustering work with multiple imputation methods. Finally, we describe the criteria applied to assess the quality of imputation methods regarding clustering results.

To identify potential clusterings, we applied the framework developed in Chapter 5. With the extended alphabet, we observe an extremum in terms of ASWw and HC with a clustering in five groups. Moreover, a local extremum appears with a seven groups clustering. The clustering into five groups takes the association between the domains into account since both ASWw and HC values are outside the 95% interval built with unstructured data. For the clustering in seven groups, only ASWw is outside this interval. The five-group clustering has no completely heterogeneous group, while one group of the seven-group clustering has an ASWw by group smaller than 0.1. The five-group clustering is more driven by the child domain, while the $R^2$ by group is balanced for the seven-group clustering. Even if it is worse in terms of statistical characteristics, the seven-group clustering allows having a more detailed clustering. For example, this clustering distinguishes between women who stay unemployed, the child growing, and those who switch to part-time employment. Therefore, we considered these two clusterings in this analysis.

Concerning MSA, an extremum is attained for ASWw with a four-group clustering, while a global minimum is attained for HC with the seven-group clustering. In both cases, we do not observe an extremum for the other cluster quality index, but we do observe an inflexion

point. The seven-group clustering has both ASWw and HC values outside the 95% interval, while it is the case only for ASWw in the case of the four-group clustering. Both clusterings are more driven by the child domain. In both cases, no group has an ASWw smaller than 0.1. The smaller group of the seven-group clustering contains 8% of the weighted sequences, which is not too small. The seven-group clustering has the advantage to distinguish between an early, standard and late child's arrival in the household. We considered both clusterings for the analysis.

With multiple imputation methods, we apply the strategy of Halpin (2012) to build a clustering after missing data were imputed through multiple imputation. This method involves stacking all the completed datasets and realise the clustering on this stacked dataset. Subsequently, a multichannel sequence is assigned to the group that contains the majority of its corresponding imputations. In the event of a tie, the sequence is randomly assigned to one of the tied groups.

To evaluate the similarity between the original dataset clusterings and those obtained after handling missing data, we used the Adjusted Rand Index (ARI) (Rand, 1971; Hubert and Arabie, 1985). This measure is widely used for comparing two clusterings (Santos and Embrechts, 2009; Warrens and van der Hoef, 2022). It is based on the Rand index, which is defined as the fraction of object pairs that are classified in the same way in both clusterings. This includes cases where object pairs are assigned to the same group in both clusterings, as well as cases where they are assigned to different groups. ARI corrects the Rand index for agreement obtained by chance (Albatineh et al., 2006).

To summarise, we have identified two clusterings built with MSA and two with EA on the original dataset, that we tried to reproduce after handling missing data with the different methods. The quality of the clusterings was measured through ARI.

## 6.4 Results

We split the presentation of the results between the two simulation frameworks. In both frameworks, we examine the characteristics of the completed datasets generated through multiple imputation algorithms and the complete dataset obtained with complete case analysis (CCA). To assess the performance of these methods, we used criteria related to longitudinal consistency, local and global association. Furthermore, in the second simulation framework, we assessed the influence on the joint clustering of professional and child domains.

In each application of the imputation algorithms, ten completed datasets were constructed. Consequently, the criteria were computed for each replication, and the resulting values were aggregated. Regarding CCA, the computation of the criteria was performed with the dataset obtained by removing each multichannel trajectory that contained at least one missing value.

The results between the MAR and MNAR mechanisms were similar. Therefore, to simplify the discussion, we only present results obtained with the MAR mechanism.

### 6.4.1 First simulation framework

We first detail the MICT-multichannel imputation algorithm's parametrisation. The algorithm requires the user to specify two parameters: the number of iterations and the order of the channels. Then, we present the comparative analysis, evaluating MICT-multichannel optimal parametrisation against other considered methods, including CCA, separate application of MICT on each channel, FCS, and two-fold FCS.

#### Parametrisation of MICT-multichannel

The results of the MICT-multichannel algorithm do not improve with an increase in the number of iterations for the civil status trajectories, either in terms of Cramer's V (Figure F.1), Cronbach's $\alpha$ (Figure F.2), or longitudinal characteristics (Figures F.3 and F.4).

The picture is different for the satisfaction with the health and professional status dataset. On the one hand, the number of iterations does not change the results when the same missing data patterns were applied to each channel. On the other hand, the local association is better when increasing the number of iterations from one to two, as shown in Figures 6.3 and 6.5. The lower the percentage of permuted sequences, the greater the differences. These differences are less marked for professional status than satisfaction with health status trajectories. For the latter, these gains come at the expense of Cronbach's $\alpha$, which is slightly worse with two iterations than with one, as shown in Figure 6.4. However, Cronbach's alpha does not change for the professional trajectories (Figure F.7), and the longitudinal characteristics are not impacted for either dataset, as seen in Figures F.5, F.6, F.8 and F.9.

Therefore, two iterations is the most suitable parametrisation.

#### Comparison between the methods

We proceed with the comparative analysis between the recently identified optimal parametrisation for the MICT-multichannel algorithm and the other approaches for handling missing data. The primary objective is to assess the performance of the MICT-multichannel algorithm relative to these methods, with a specific focus on the impact of the strength of both longitudinal and cross-sectional information.

The results for the local association are displayed in Figures 6.6, 6.10 and 6.14 for civil status, health satisfaction status and professional status, respectively. Figures 6.9, 6.13 and 6.17 show the results for Cronbach's $\alpha$. Boxplots for the three longitudinal criteria are shown in Figures 6.7 and 6.8 for the civil status (the first one corresponds to the same patterns of

missing data and the second one to different patterns), Figures 6.11 and 6.12 for the satisfaction with health status and Figures 6.15 and 6.16 for the professional status.

Applying the MICT algorithm separately to each channel underestimates Cramer's V, and hence the local association between channels. The size of the bias depends on the magnitude of the local and global association, the rate of transition in the dataset and the percentage of missing data: the larger the local association, the higher the percentage of missing data and the more subject to transitions the channels are, the larger the bias. Since MICT-multichannel often presents a bias close to zero for Cramer's V, the gain obtained from its application over the standard MICT algorithm is also more prominent in these three situations. Those behaviours are not surprising. First, if a dataset is not subject to many transitions, previous and subsequent observations are good predictors and values from the other channels add little information. Then, the lower the Cramer's V, the less information is added by considering the other channel. Finally, more missing data implies more uncertainty and, potentially, bias. As for Cramer's V, MICT leads to an underestimation of Cronbach's $\alpha$. Conversely, except for the civil status, MICT-multichannel generally generates completed datasets where Cronbach's $\alpha$ is overestimated.

In terms of longitudinal characteristics, MICT and MICT-multichannel show an almost identical performance on the civil status trajectories. MICT is better than MICT-multichannel regarding the imputation of professional status. The differences are more marked for the sequencing characteristic. For satisfaction with health status, MICT-multichannel is slightly better than MICT when missing data were generated separately on each channel. In addition, MICT-multichannel is slightly better on the sequencing but worse on the duration when the missing data are generated simultaneously on each channel. However, in general, the differences between the two algorithms regarding the longitudinal criteria are marginal.

In most cases, CCA is worse than MICT-multichannel regarding Cramer's V. The bias is more variable when different patterns of missing data were applied to each channel due to the significantly reduced number of complete remaining multichannel sequences in this case. However, there are scenarios, such as the same pattern based on a low rate of missing data, where CCA is best, showing less bias than MICT-multichannel. CCA tends to overestimate Cronbach's $\alpha$ when the same pattern of missing data was applied to each channel. In contrast, the median bias is closer to zero when different patterns were applied. In some scenarios, the median bias is lower for CCA than for MICT-multichannel. However, the range of values is clearly wider with CCA. Regarding the longitudinal characteristics, CCA is, as was already pointed out during the analysis of the last chapter, more impacted in terms of timing than duration and sequencing. Concerning the timing criteria, the results are clearly worse than all the imputation methods. However, for the duration and the sequencing, it shows better results than FCS and two-fold FCS, but not MICT-multichannel, in most scenarios related to the civil

status.

Results between FCS and two-fold FCS are very close, especially regarding the local and global associations. Concerning the longitudinal characteristics, there is a tendency for two-fold FCS to lag behind FCS in terms of duration. Both algorithms lead to overestimated Cronbach's $\alpha$. Contrary to MICT-multichannel, it is also the case for the civil status trajectories.

Both FCS and two-fold FCS are outperformed by MICT-multichannel regarding the imputation of the civil status dataset, in every scenario and for every criterion considered. In particular, regarding the longitudinal criteria, the differences are more pronounced for the duration and the sequencing than the timing. On the professional status trajectories, the longitudinal criteria are, except for the duration, close between MICT-multichannel and the two FCS algorithms, the last two being even better in some cases. However, in most cases, the biases of Cramer's V and Cronbach's $\alpha$ are smaller with MICT-multichannel. With satisfaction with health status, MICT-multichannel is better regarding Cronbach's alpha, and the results are similar for the longitudinal criteria. Concerning Cramer's V, MICT-multichannel is better when the same pattern is applied to each channel, while FCS and two-fold FCS are better when different patterns are applied. Note that when different patterns were generated with a high percentage of missing data, the variance of the bias is smaller with MICT-multichannel.

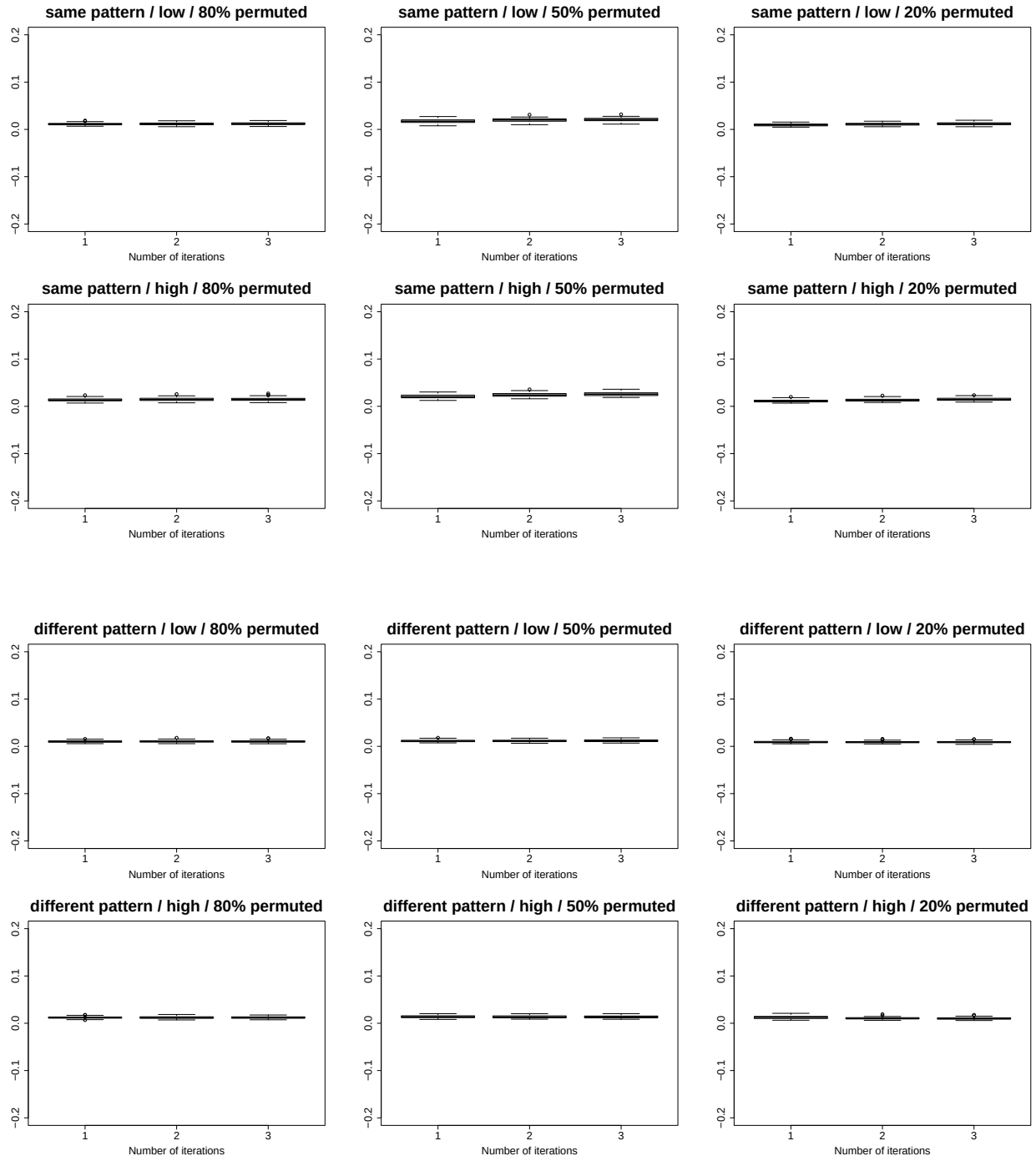



Figure 6.3: MAR mechanism - Boxplots of the criteria relative to the local association, , obtained from handling missing data on the satisfaction with health status dataset with 1 to 3 iterations of the *MICT-multichannel* algorithm. Each subplot corresponds to a scenario of missing data generation and is labelled as "type of pattern / rate of missing data / % of sequences from the duplicated dataset permuted".

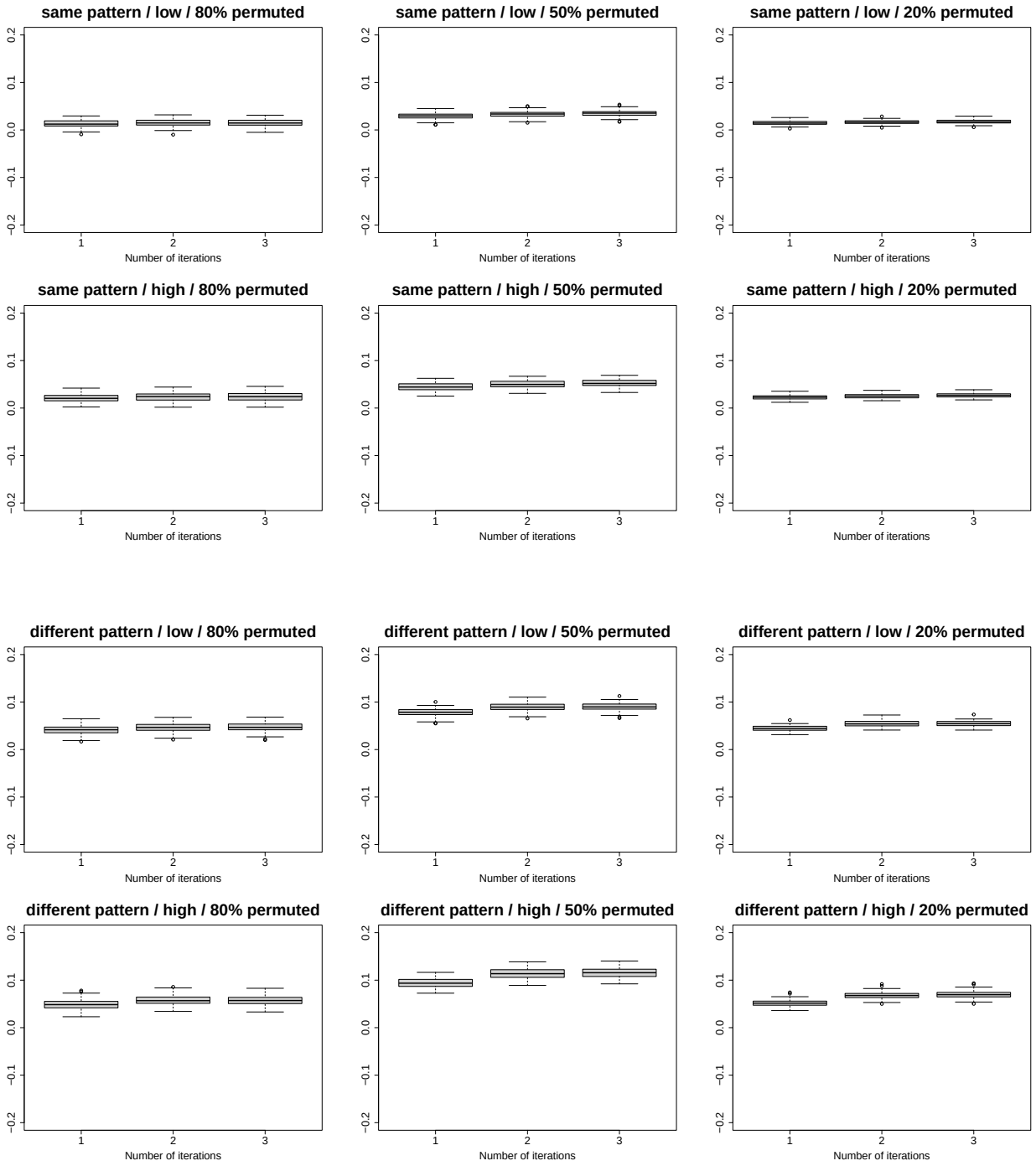



Figure 6.4: MAR mechanism - Boxplots of the Cronbach's $\alpha$ bias, obtained from handling missing data on the satisfaction with health status dataset with 1 to 3 iterations of the *MICT-multichannel* algorithm. Each subplot corresponds to a scenario of missing data generation and is labelled as "type of pattern / rate of missing data / % of sequences from the duplicated dataset permuted".

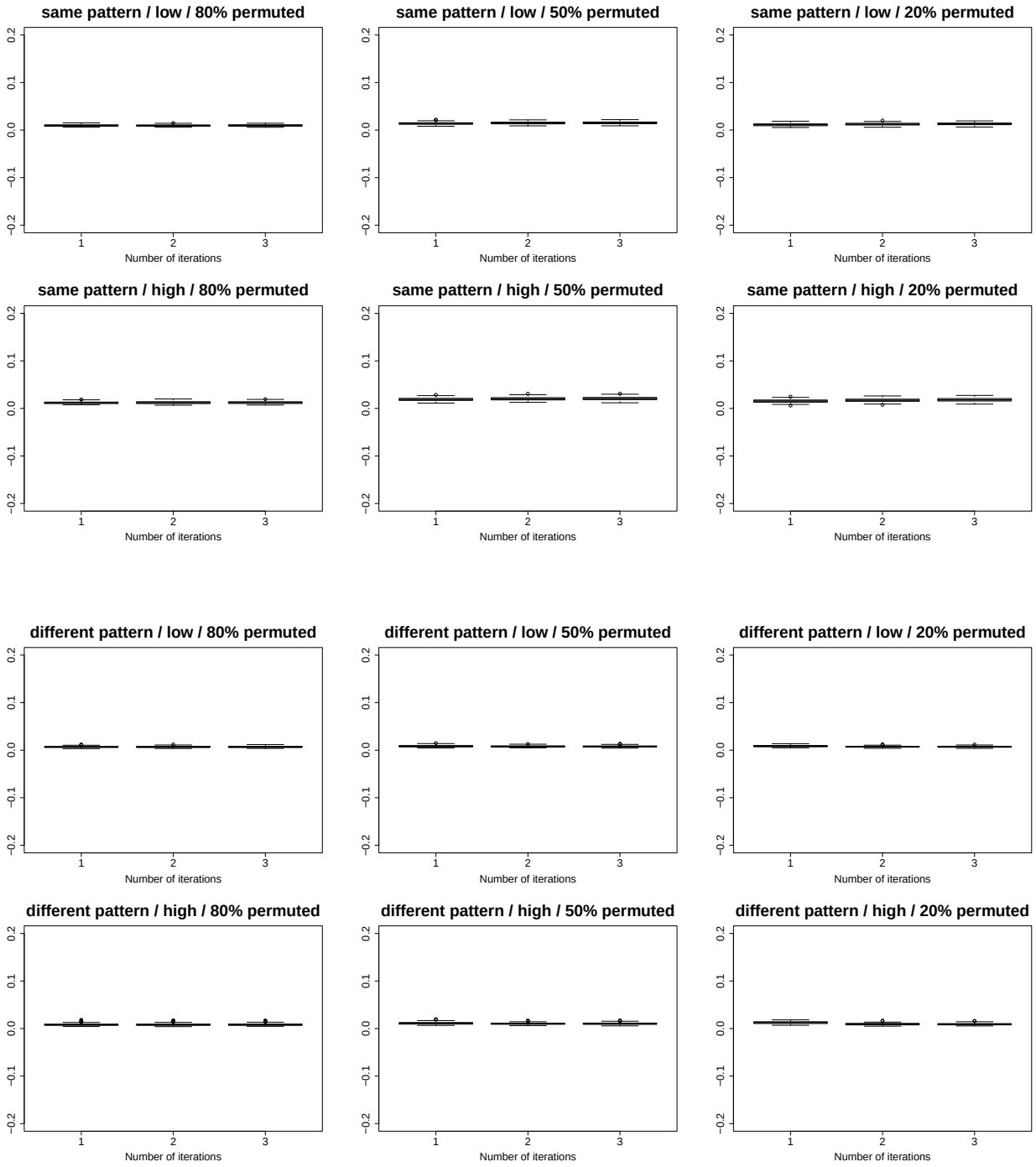



Figure 6.5: MAR mechanism - Boxplots of the criteria relative to the local association, , obtained from handling missing data on the professional status dataset with 1 to 3 iterations of the *MICT-multichannel* algorithm. Each subplot corresponds to a scenario of missing data generation and is labelled as "type of pattern / rate of missing data / % of sequences from the duplicated dataset permuted".

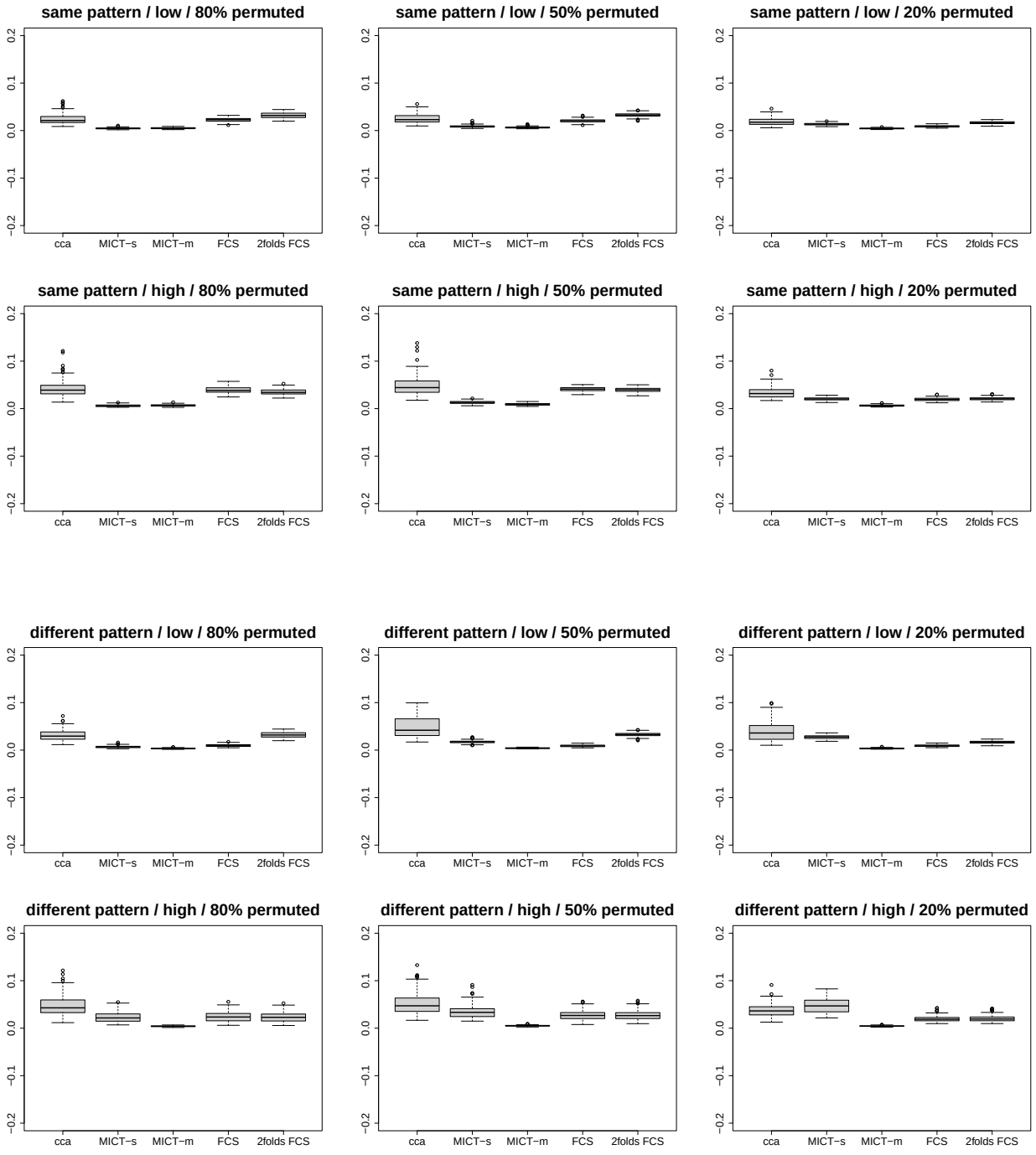



Figure 6.6: MAR mechanism - Boxplots of the criteria relative to the local association, , obtained from handling missing data on the civil status with the five considered methods, namely *CCA*, *MICT* applied to each channel separately (labelled as "MICT-s"), *MICT-multichannel* (labelled as "MICT-m"), *FCS* and *two-fold FCS* (labelled as "2folds FCS"). Each subplot corresponds to a scenario of missing data generation and is labelled as "type of pattern / rate of missing data / % of sequences from the duplicated dataset permuted".

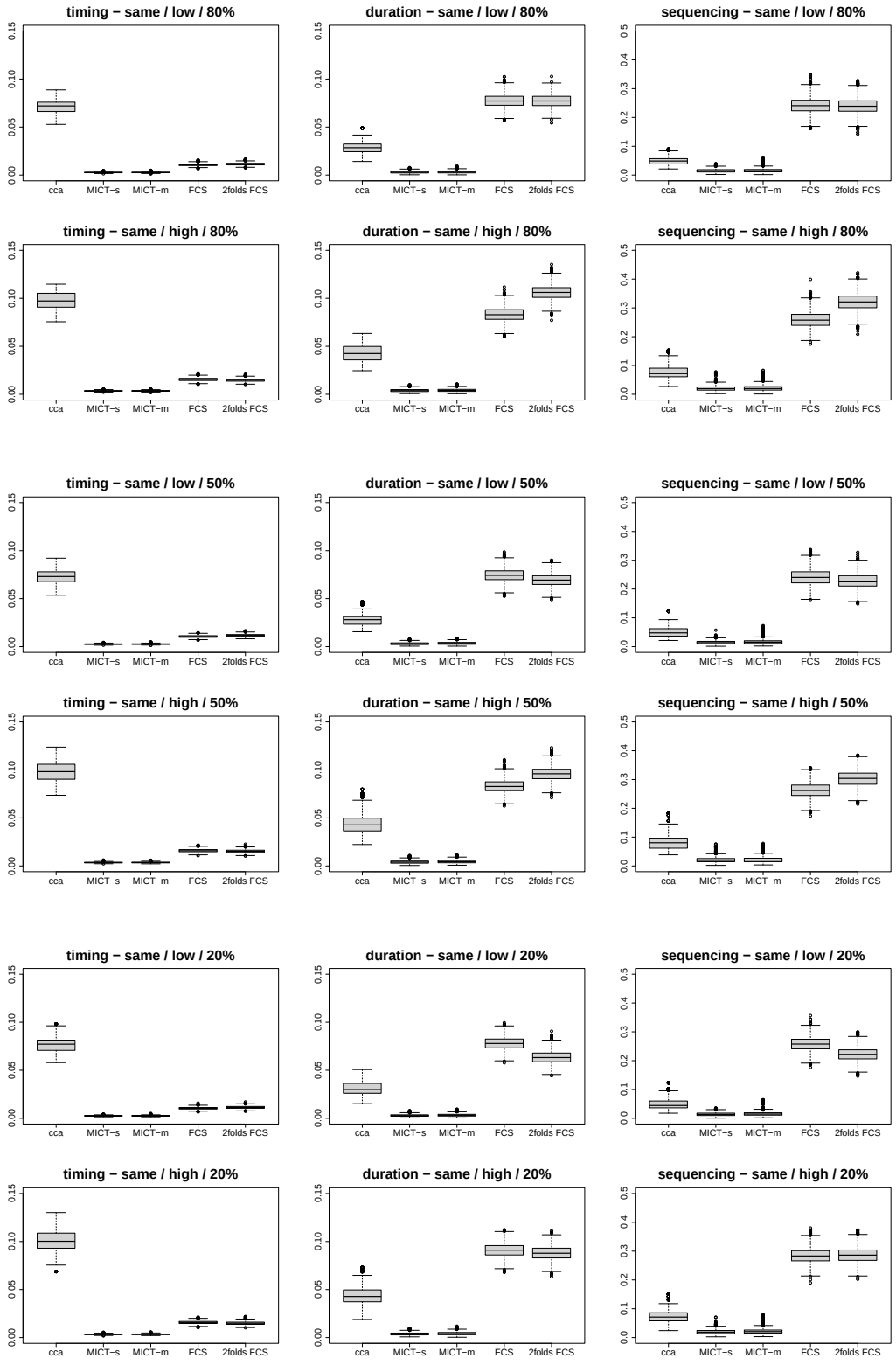



Figure 6.7: MAR mechanism - Boxplots of the longitudinal characteristics bias, obtained from handling missing data on the civil status dataset with the five considered methods, namely *CCA*, *MICT* applied to each channel separately (labelled as "MICT-s"), *MICT-multichannel* (labelled as "MICT-m"), *FCS* and *two-fold FCS* (labelled as "2folds FCS"). Each row of subplots corresponds to a scenario of missing data generation with a same pattern of missing values and is labelled as "longitudinal characteristic - type of pattern / rate of missing data / % of sequences from the duplicated dataset permuted".

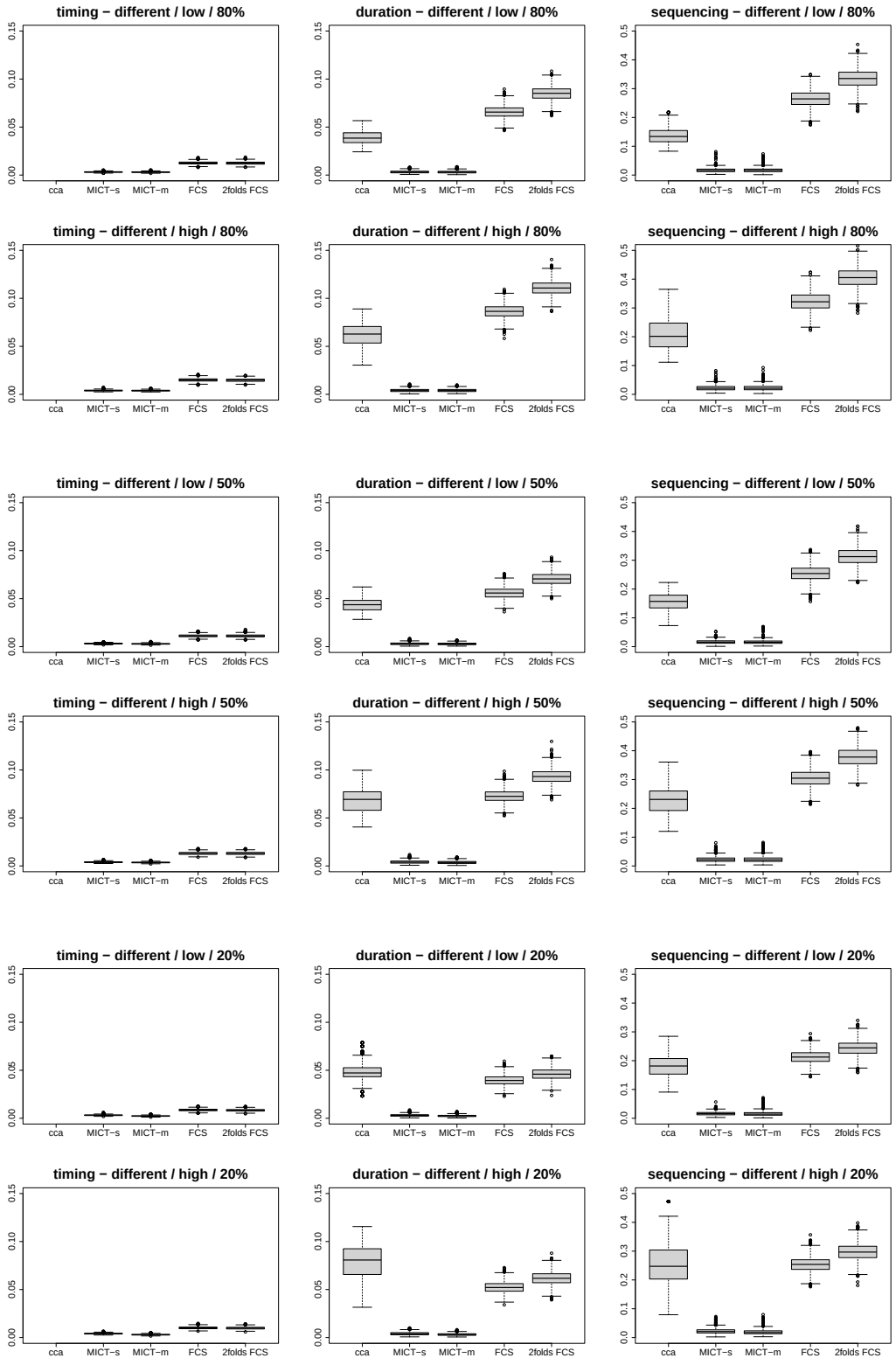



Figure 6.8: MAR mechanism - Boxplots of the longitudinal characteristics bias, obtained from handling missing data on the civil status dataset with the five considered methods, namely *CCA*, *MICT* applied to each channel separately (labelled as "MICT-s"), *MICT-multichannel* (labelled as "MICT-m"), *FCS* and *two-fold FCS* (labelled as "2folds FCS"). Each row of subplots corresponds to a scenario of missing data generation with a different pattern of missing values and is labelled as "longitudinal characteristic - type of pattern / rate of missing data / % of sequences from the duplicated dataset permuted".

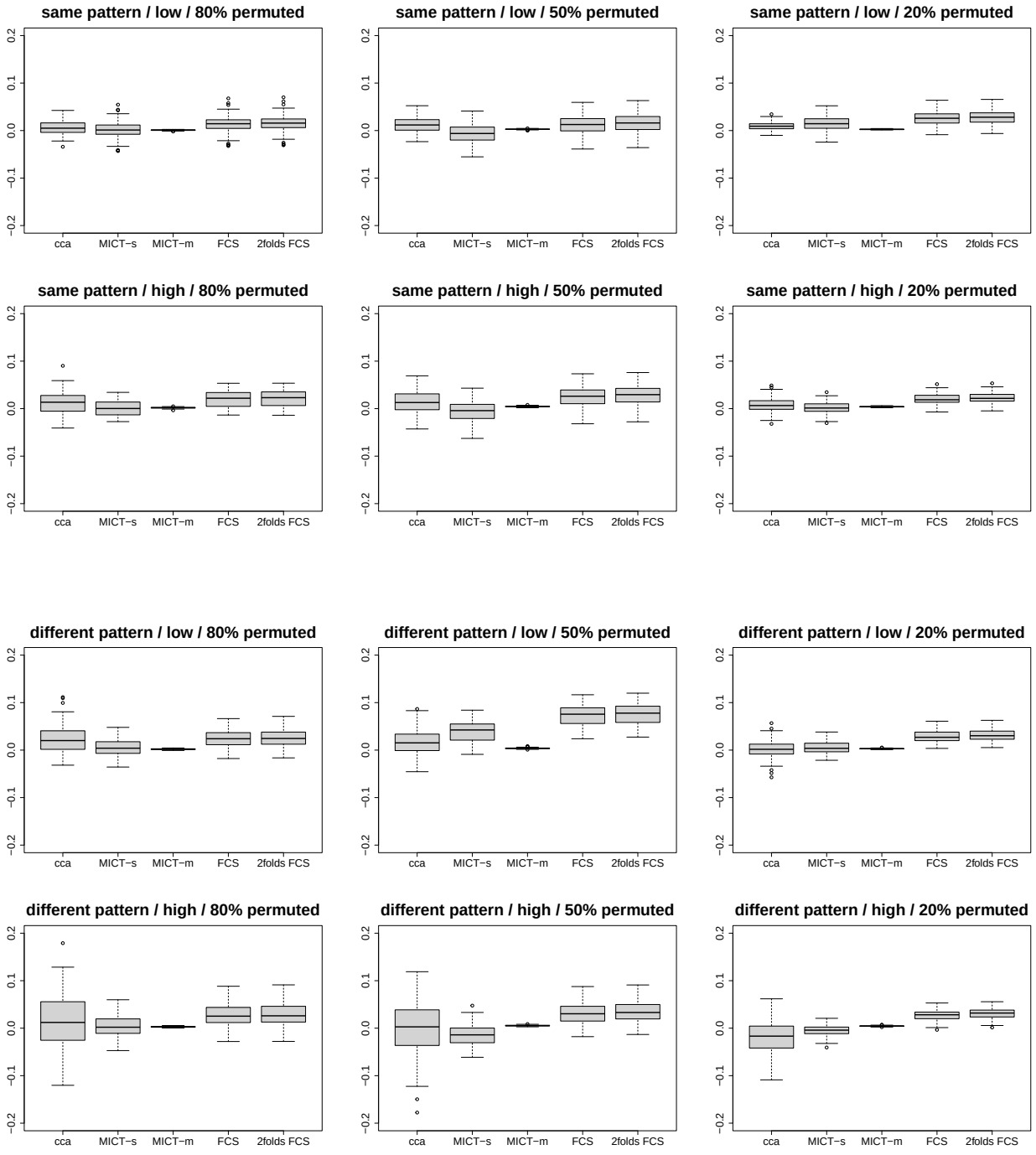



Figure 6.9: MAR mechanism - Boxplots of the Cronbach's $\alpha$ bias, obtained from handling missing data on the civil status dataset with the five considered methods, namely *CCA*, *MICT* applied to each channel separately (labelled as "MICT-s"), *MICT-multichannel* (labelled as "MICT-m"), *FCS* and *two-fold FCS* (labelled as "2folds FCS"). Each subplot corresponds to a scenario of missing data generation. It is labelled as "type of pattern / rate of missing data / % of sequences from the duplicated dataset permuted".

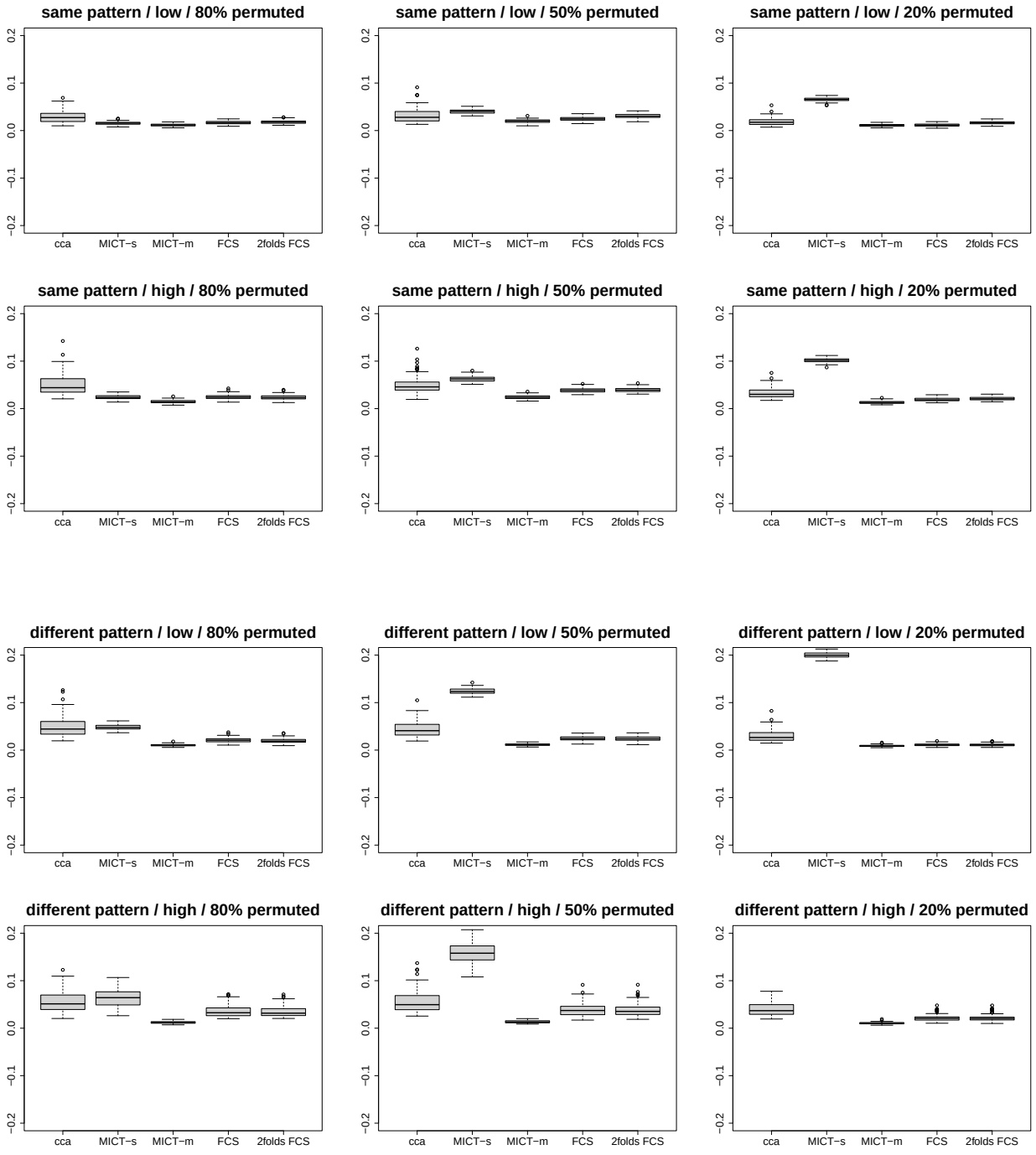



Figure 6.10: MAR mechanism - Boxplots of the criteria relative to the local association, , obtained from handling missing data on the satisfaction with health status with the five considered methods, namely *CCA*, *MICT* applied to each channel separately (labelled as "MICT-s"), *MICT-multichannel* (labelled as "MICT-m"), *FCS* and *two-fold FCS* (labelled as "2folds FCS"). Each subplot corresponds to a scenario of missing data generation and is labelled as "type of pattern / rate of missing data / % of sequences from the duplicated dataset permuted".

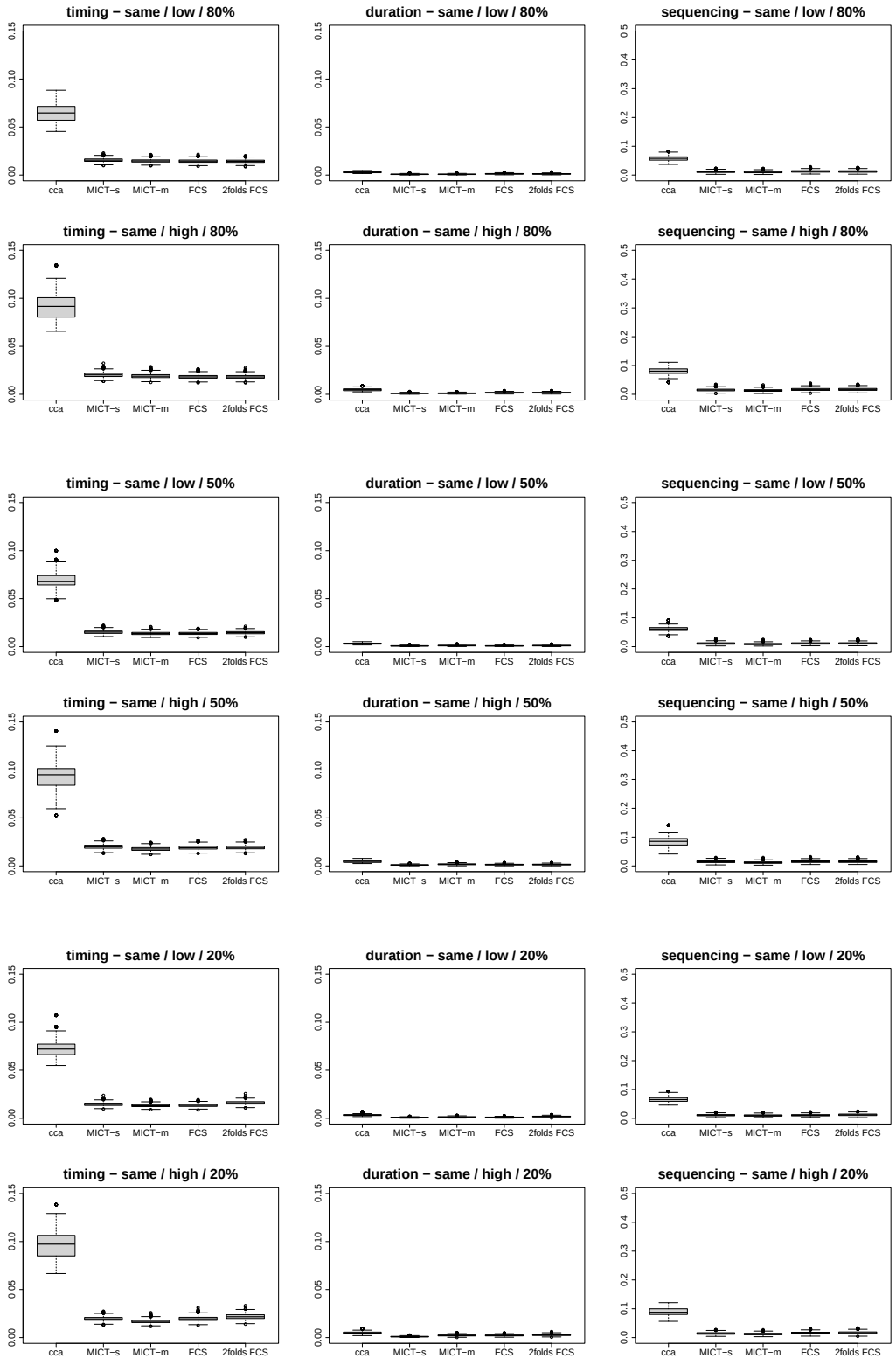



Figure 6.11: MAR mechanism - Boxplots of the longitudinal characteristics bias, obtained from handling missing data on the satisfaction with health status dataset with the five considered methods, namely *CCA*, *MICT* applied to each channel separately (labelled as "MICT-s"), *MICT-multichannel* (labelled as "MICT-m"), *FCS* and *two-fold FCS* (labelled as "2folds FCS"). Each row of subplots corresponds to a scenario of missing data generation with a same pattern of missing values and is labelled as "longitudinal characteristic - type of pattern / rate of missing data / % of sequences from the duplicated dataset permuted".

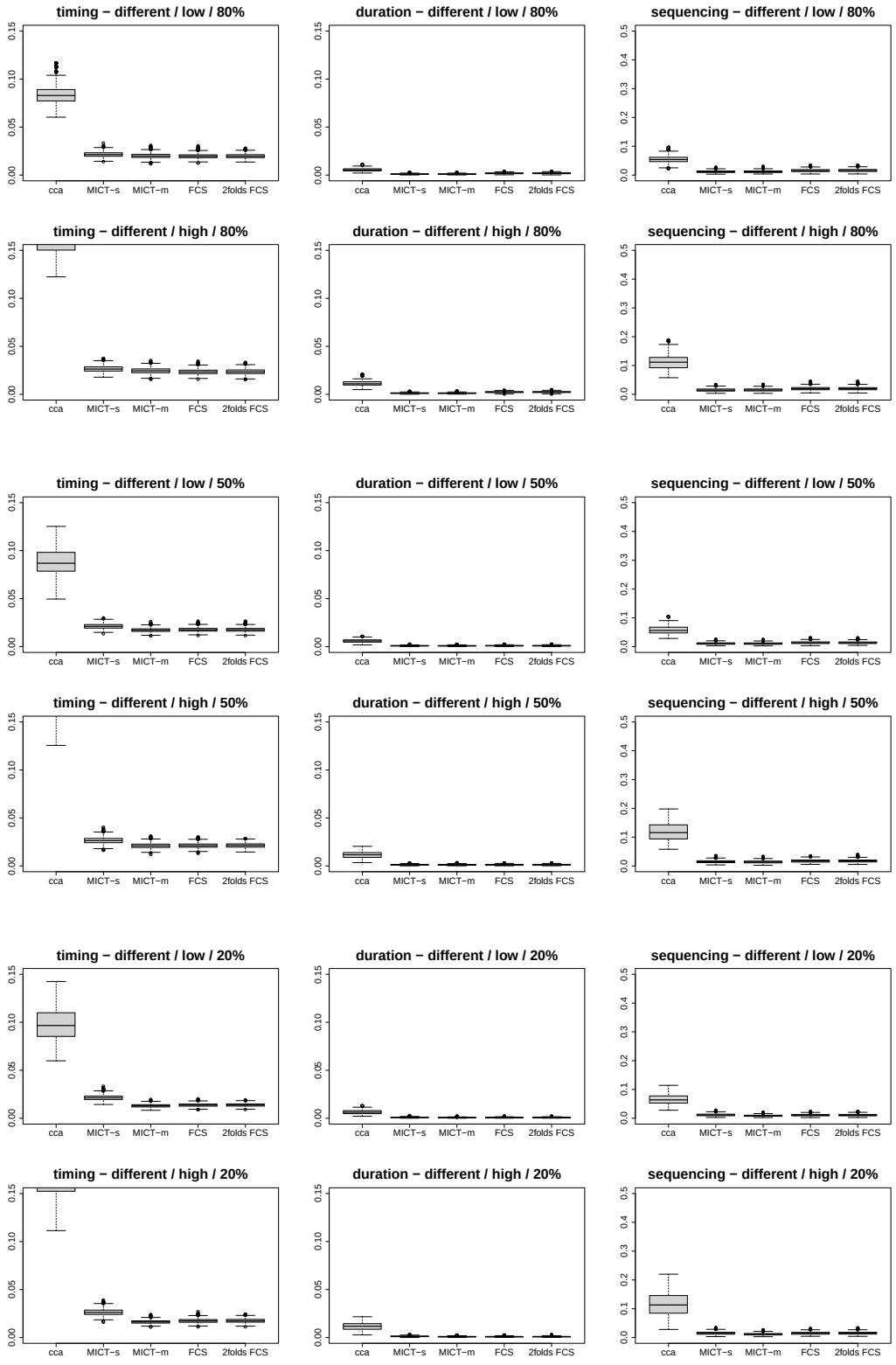



Figure 6.12: MAR mechanism - Boxplots of the longitudinal characteristics bias, obtained from handling missing data on the satisfaction with health status dataset with the five considered methods, namely *CCA*, *MICT* applied to each channel separately (labelled as "MICT-s"), *MICT-multichannel* (labelled as "MICT-m"), *FCS* and *two-fold FCS* (labelled as "2folds FCS"). Each row of subplots corresponds to a scenario of missing data generation with a different pattern of missing values and is labelled as "longitudinal characteristic - type of pattern / rate of missing data / % of sequences from the duplicated dataset permuted".

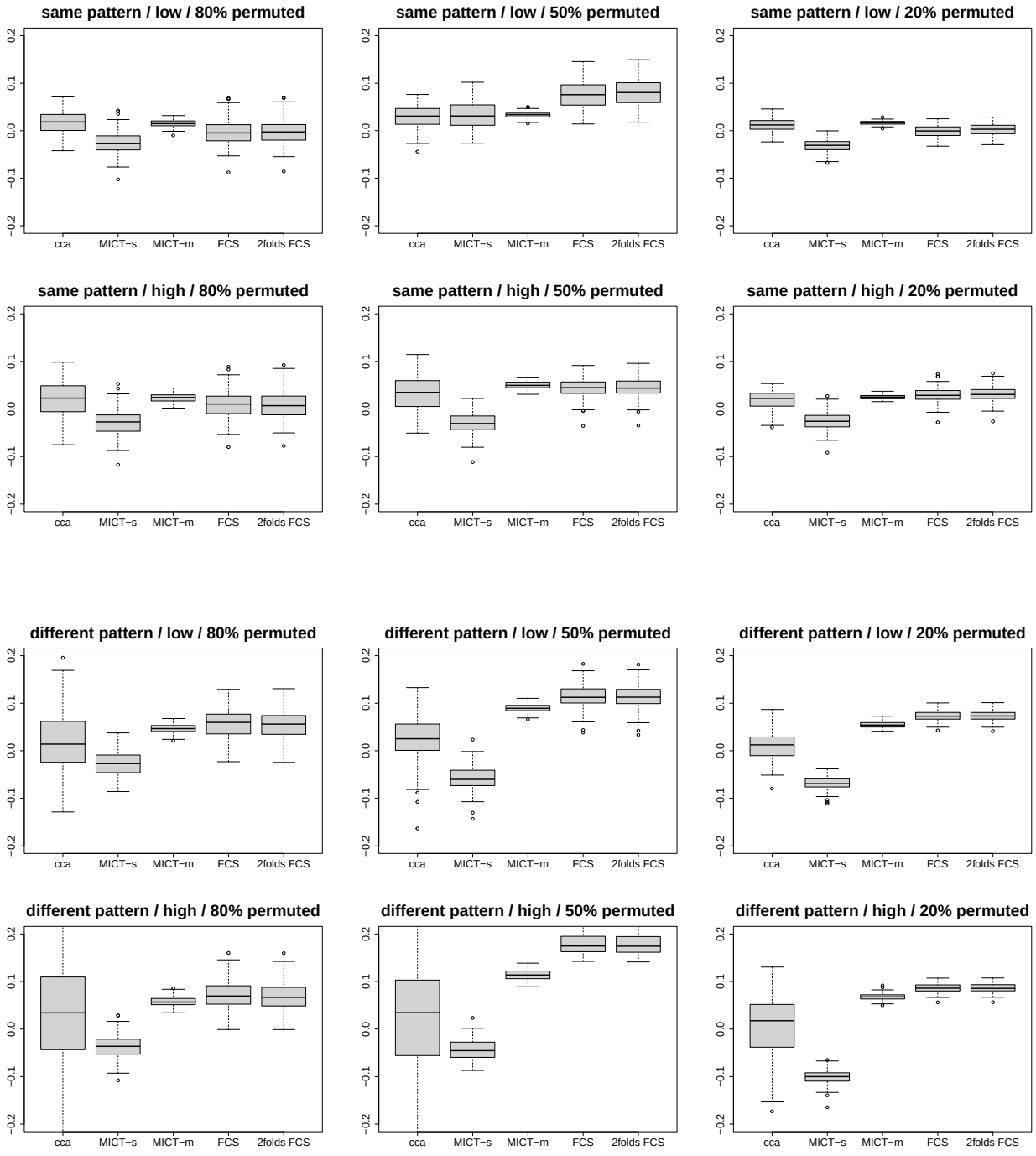



Figure 6.13: MAR mechanism - Boxplots of the Cronbach's $\alpha$ bias, obtained from handling missing data on the satisfaction with health status dataset with the five considered methods, namely *CCA*, *MICT* applied to each channel separately (labelled as "MICT-s"), *MICT-multichannel* (labelled as "MICT-m"), *FCS* and *two-fold FCS* (labelled as "2folds FCS"). Each subplot corresponds to a scenario of missing data generation. It is labelled as "type of pattern / rate of missing data / % of sequences from the duplicated dataset permuted".

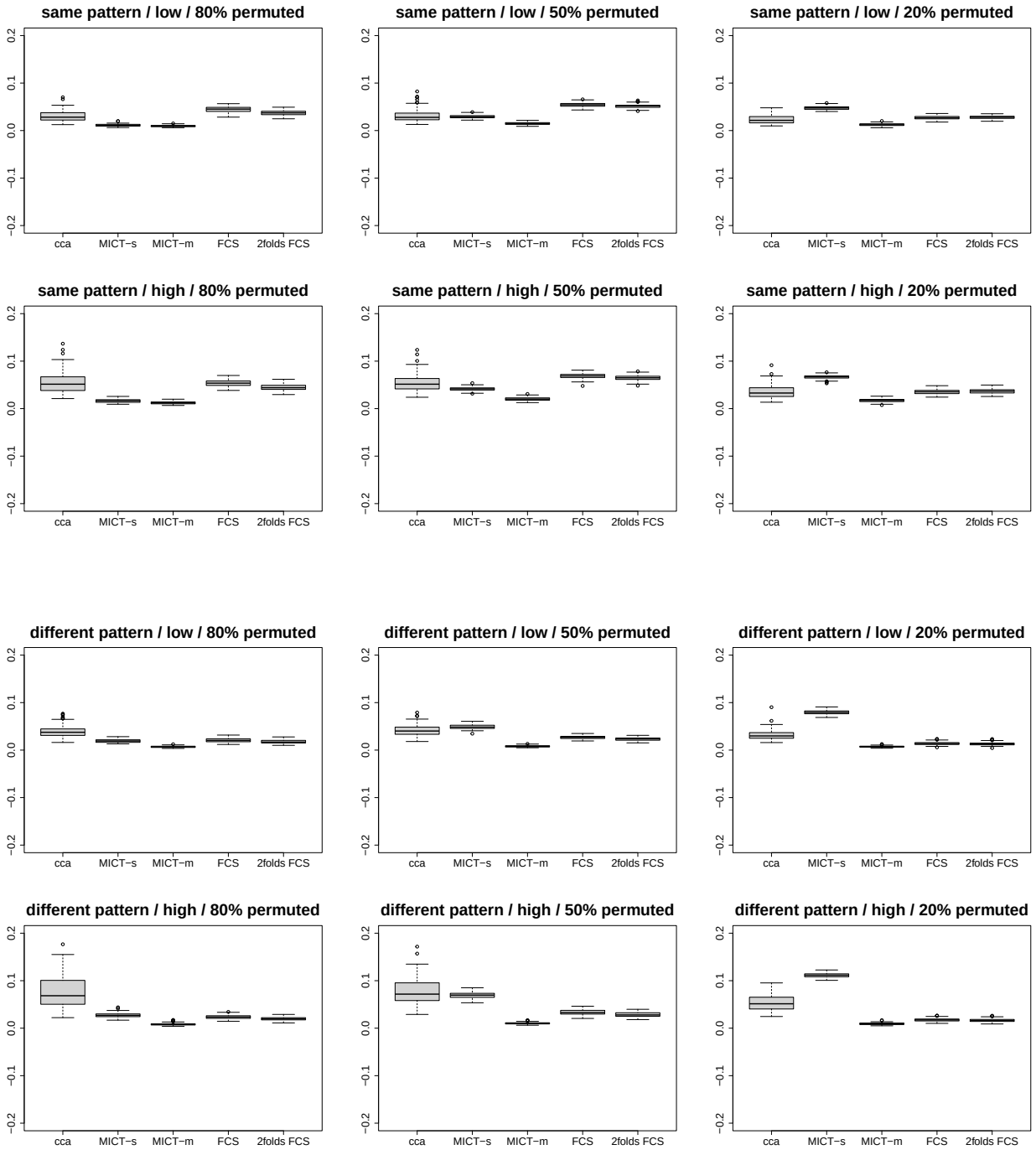



Figure 6.14: MAR mechanism - Boxplots of the criteria relative to the local association, , obtained from handling missing data on the professional status with the five considered methods, namely *CCA*, *MICT* applied to each channel separately (labelled as "MICT-s"), *MICT-multichannel* (labelled as "MICT-m"), *FCS* and *two-fold FCS* (labelled as "2folds FCS"). Each subplot corresponds to a scenario of missing data generation and is labelled as "type of pattern / rate of missing data / % of sequences from the duplicated dataset permuted".

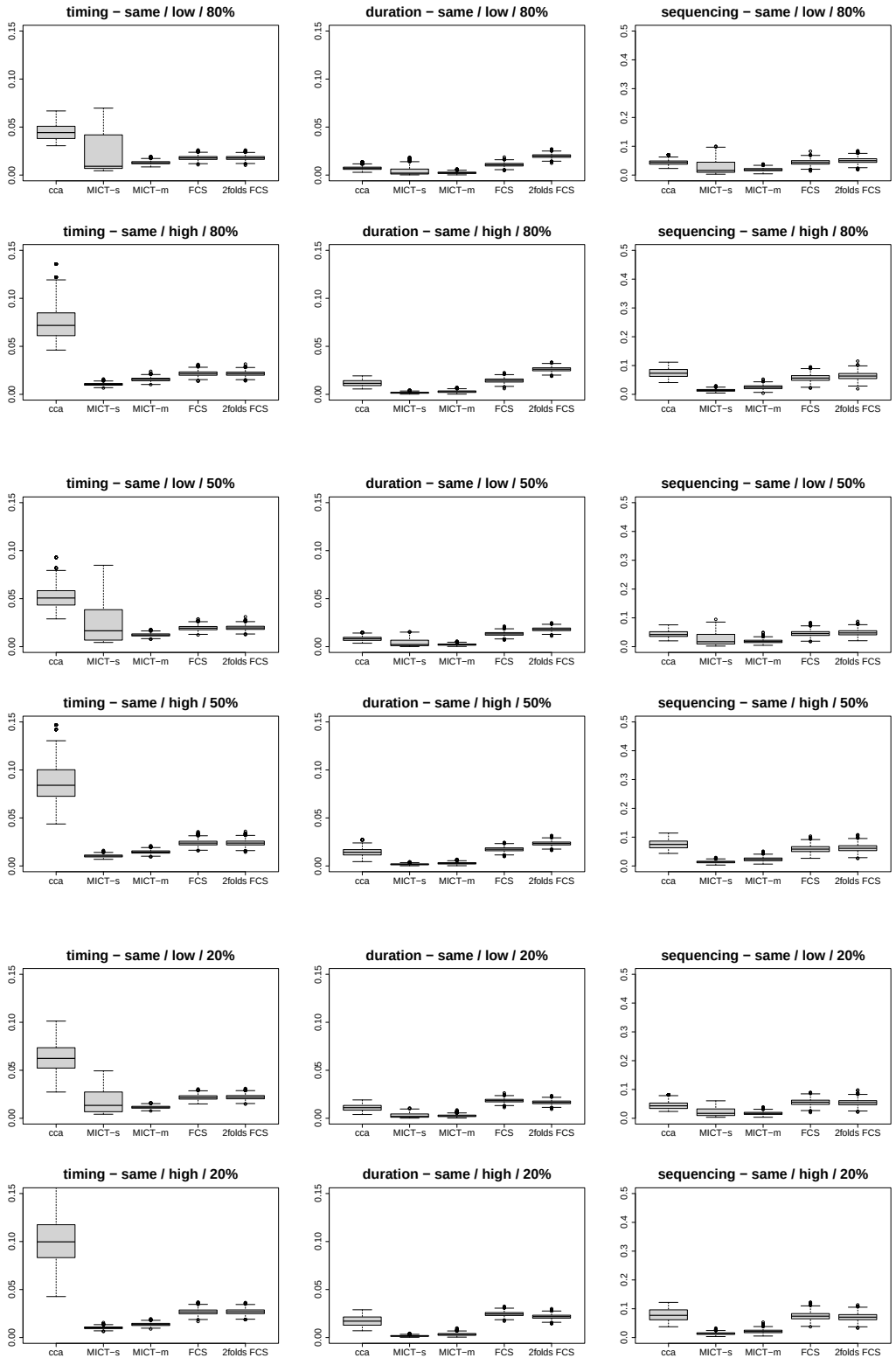



Figure 6.15: MAR mechanism - Boxplots of the longitudinal characteristics bias, obtained from handling missing data on the professional status dataset with the five considered methods, namely *CCA*, *MICT* applied to each channel separately (labelled as "MICT-s"), *MICT-multichannel* (labelled as "MICT-m"), *FCS* and *two-fold FCS* (labelled as "2folds FCS"). Each row of subplots corresponds to a scenario of missing data generation with a same pattern of missing values and is labelled as "longitudinal characteristic - type of pattern / rate of missing data / % of sequences from the duplicated dataset permuted".

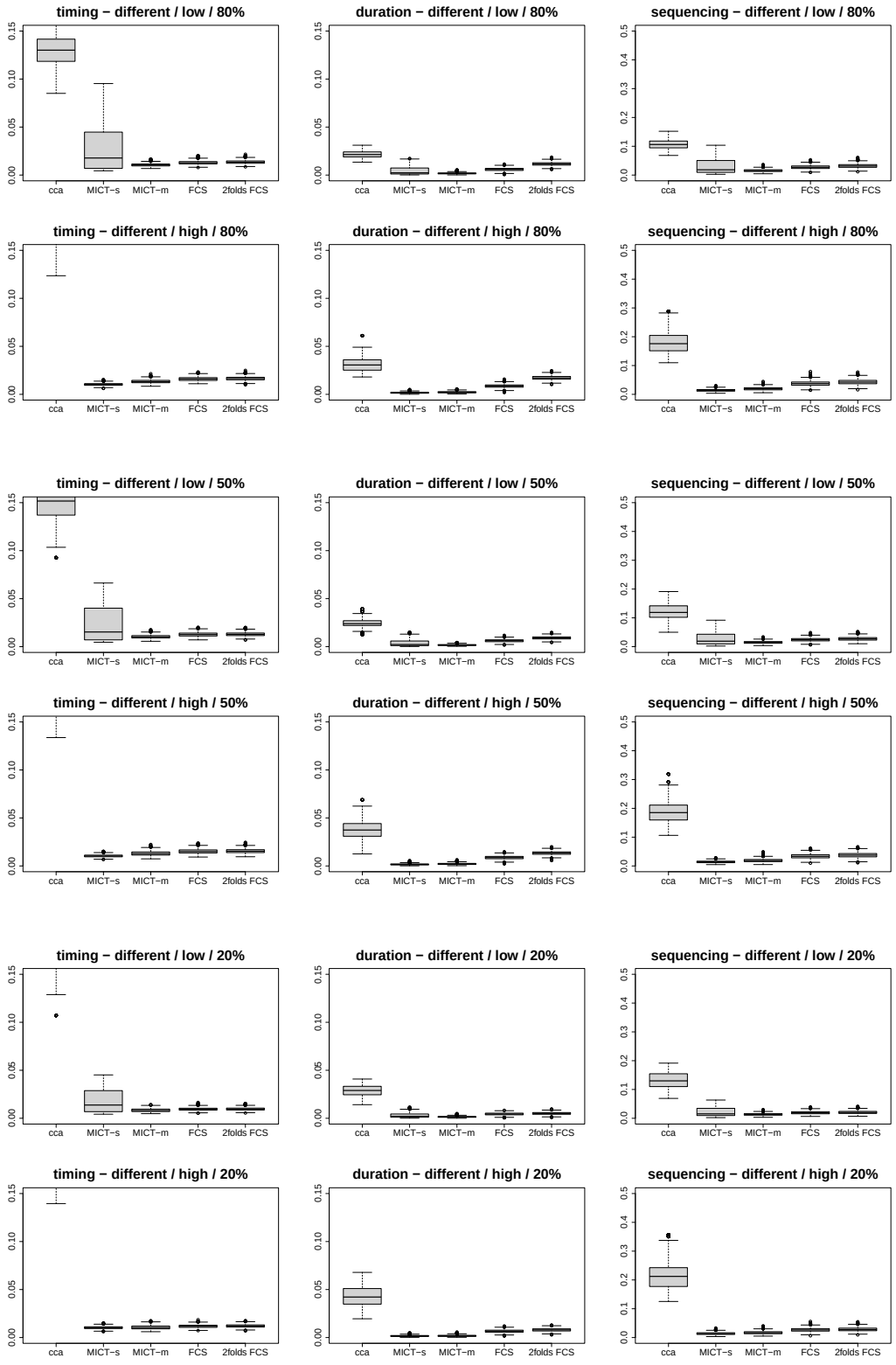



Figure 6.16: MAR mechanism - Boxplots of the longitudinal characteristics bias, obtained from handling missing data on the professional status dataset with the five considered methods, namely *CCA*, *MICT* applied to each channel separately (labelled as "MICT-s"), *MICT-multichannel* (labelled as "MICT-m"), *FCS* and *two-fold FCS* (labelled as "2folds FCS"). Each row of subplots corresponds to a scenario of missing data generation with a different pattern of missing values and is labelled as "longitudinal characteristic - type of pattern / rate of missing data / % of sequences from the duplicated dataset permuted".

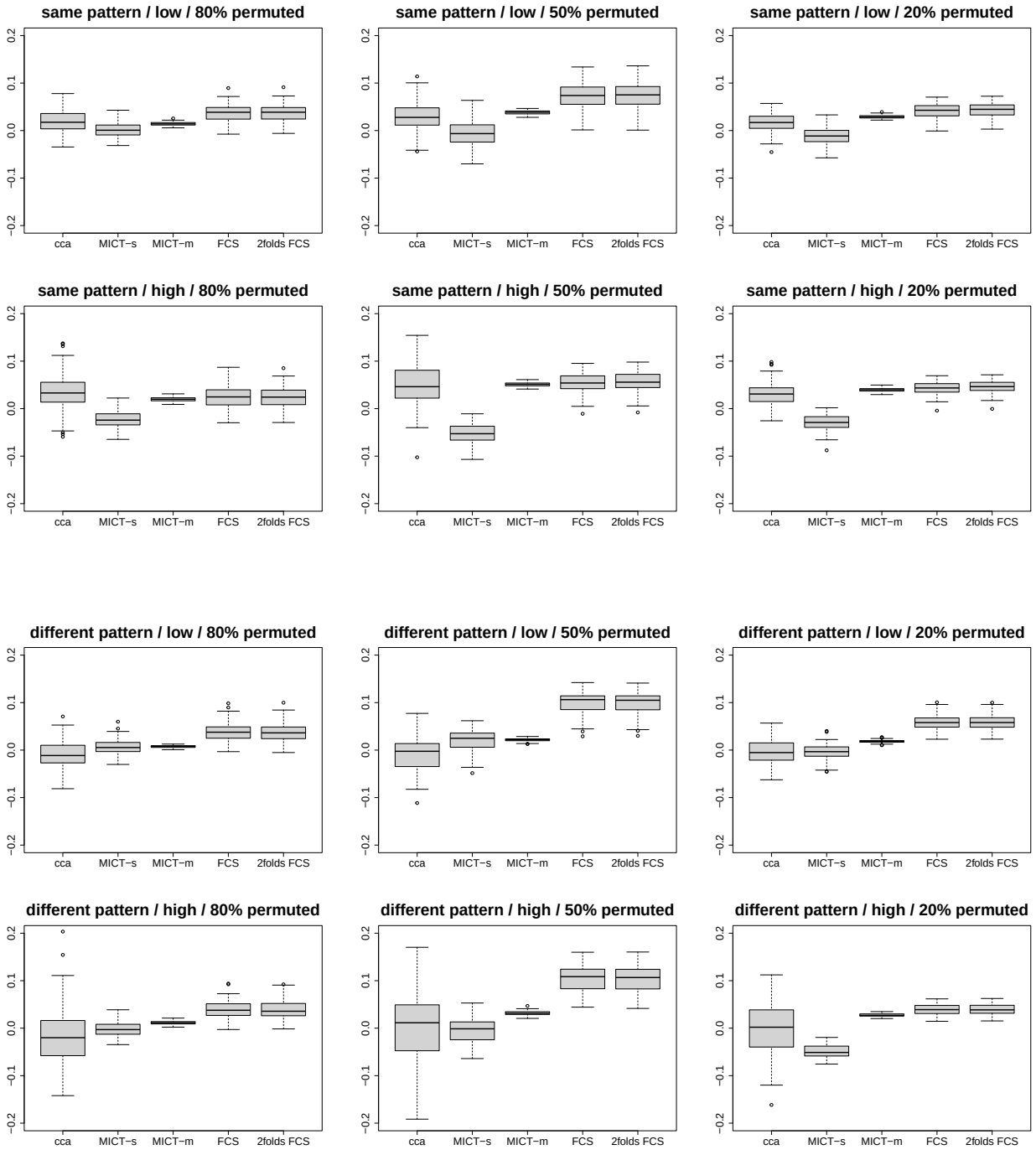



Figure 6.17: MAR mechanism - Boxplots of the Cronbach's $\alpha$ bias, obtained from handling missing data on the professional status dataset with the five considered methods, namely *CCA*, *MICT* applied to each channel separately (labelled as "MICT-s"), *MICT-multichannel* (labelled as "MICT-m"), *FCS* and *two-fold FCS* (labelled as "2folds FCS"). Each subplot corresponds to a scenario of missing data generation. It is labelled as "type of pattern / rate of missing data / % of sequences from the duplicated dataset permuted".

### 6.4.2 Second simulation framework

In accordance with the structure of the first simulation framework, we first present the optimal parametrisation of the MICT-multichannel algorithm. Subsequently, we proceed with the comparative analysis, contrasting the performance of MICT-multichannel with other methods used for handling missing data. In this second simulation framework, we further expand our analysis by exploring the influence of the imputation methods and two ad hoc techniques designed specifically for sequence analysis, on the joint clustering of professional and child status.

#### Parametrisation of MICT-multichannel

Neither the number of iterations (Figures F.10, F.11, F.12 and F.13), nor the order of channels (Figures F.14, F.17, F.15 and F.16) impact the MICT-multichannel algorithm's performance.

Therefore, we chose the results obtained with two iterations, for consistency with the choice made on the first simulation framework, and an ordering of the channels that is child, cohabitational and finally professional status.

#### Comparison between the methods

The results are shown in Figures 6.18 regarding the bias of Cramer's V, in Figure 6.19 for Cronbach's $\alpha$ and Figures 6.20 and 6.21 for the longitudinal criteria.

A slight negative bias was induced, both for Cramer's V and Cronbach's $\alpha$, when MICT is applied separately to each channel. Its magnitude is, as on the framework based on duplicated datasets, higher with higher rates of missing data. However, it remains globally small since, in every situation, it is smaller than 0.03. MICT-multichannel gives better results regarding Cramer's V and Cronbach's $\alpha$ than MICT applied to each channel separately. The biases for Cramer's V and Cronbach's $\alpha$ are close to zero in every scenario and channel. The median is close to zero, and the variance is small. Most of the results regarding the longitudinal characteristics are almost identical between the two algorithms. The only exception is the child domain sequencing when the same patterns of missing data were applied to each channel, where the results are slightly better with MICT.

CCA presents a small positive median bias for Cramer's V and Cronbach's $\alpha$ for each of the four scenarios and generally induces higher bias results. For example, considering the same pattern of missing data, keeping only the completely observed multichannel trajectories lead to a bias larger than 0.05 for Cramer's V. In contrast, with MICT-multichannel, FCS and two-fold FCS, the worse bias is smaller than 0.02. The conclusions regarding the longitudinal characteristics are the same as with the duplicated datasets; CCA is worse on every criterion, and the differences are especially marked for the timing characteristic.

Comparing FCS and two-fold FCS, the results are usually similar. However, in every scenario

except for different patterns applied with a high rate of missing data, two-fold FCS is better than FCS for retrieving the duration characteristic of the cohabitational and professional trajectories.

On the one hand, FCS, two-fold FCS and MICT-multichannel show close results regarding Cramer's V, Cronbach's $\alpha$ and timing, with MICT-multichannel being better, however. On the other hand, MICT-multichannel clearly outperforms the two other algorithms regarding every channel's duration and sequencing characteristics. These differences are more marked when the same missing data patterns were applied on each channel. For example, the distortion of the sequencing characteristic of the child channel is more than five times higher when FCS and two-fold FCS are used to impute datasets with the same patterns of missing data on each channel.

### Clustering results

Using the framework established in the previous chapter, we have identified two joint clusterings constructed through MSA: a four-group clustering and a seven-group clustering. Additionally, we have derived two potential clusterings through EA: a five-group clustering and another comprising seven groups. In this subsection, we compare the quality of the joint clusterings of professional and child status for women, either built from the imputed datasets or obtained with the two ad hoc methods specific to sequence analysis. ARI is used as a measure of performance. Boxplots of the ARI are shown in Figure 6.22 for MSA and in Figure 6.23 for EA.

In terms of ARI, we observe similar outcomes across the imputation methods, indicating comparable performance. In contrast, the two ad hoc methods generally exhibit inferior performance compared to the imputation methods. Indeed, for the strategy involving the consideration of a missing state as an additional state, the results are close to the imputation methods when distinct patterns of missing data were applied to each channel with a low rate of missing data. However, as the rate of missing data increases, the disparity between the imputation methods and this strategy widens. Notably, it is clearly outperformed when the same pattern of missing data is simulated across all channels. Regarding the other ad hoc method, meaning considering a missing state maximally different from any other state (including a missing state), it performs worse than imputation methods when different patterns are applied to each channel. When the same pattern is applied to each channel, the conclusions differ between MSA and EA. While the performance of this ad hoc method is close to the one of imputation methods when EA was used to combine the information, it is worse when MSA was applied.

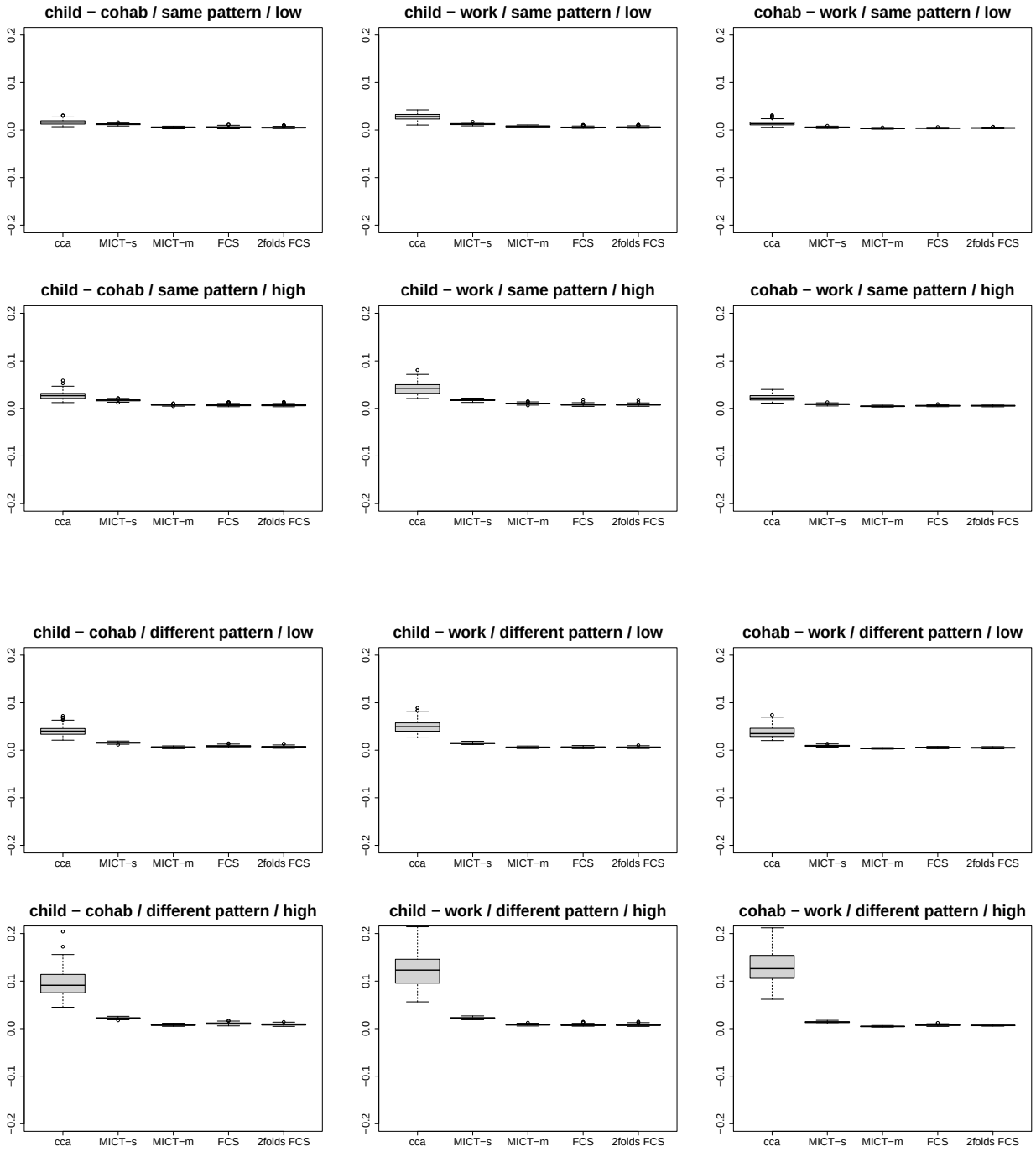



Figure 6.18: MAR mechanism - Boxplots of the bias of the mean Cramer's V, for each pair of channels, obtained with the five considered methods, namely *CCA*, *MICT* applied to each channel separately (labelled as "MICT-s"), *MICT-multichannel* (labelled as "MICT-m"), *FCS* and *two-fold FCS* (labelled as "2folds FCS"). Each subplot corresponds to a scenario of missing data generation. It is labelled as "channels considered / type of pattern / rate of missing data".

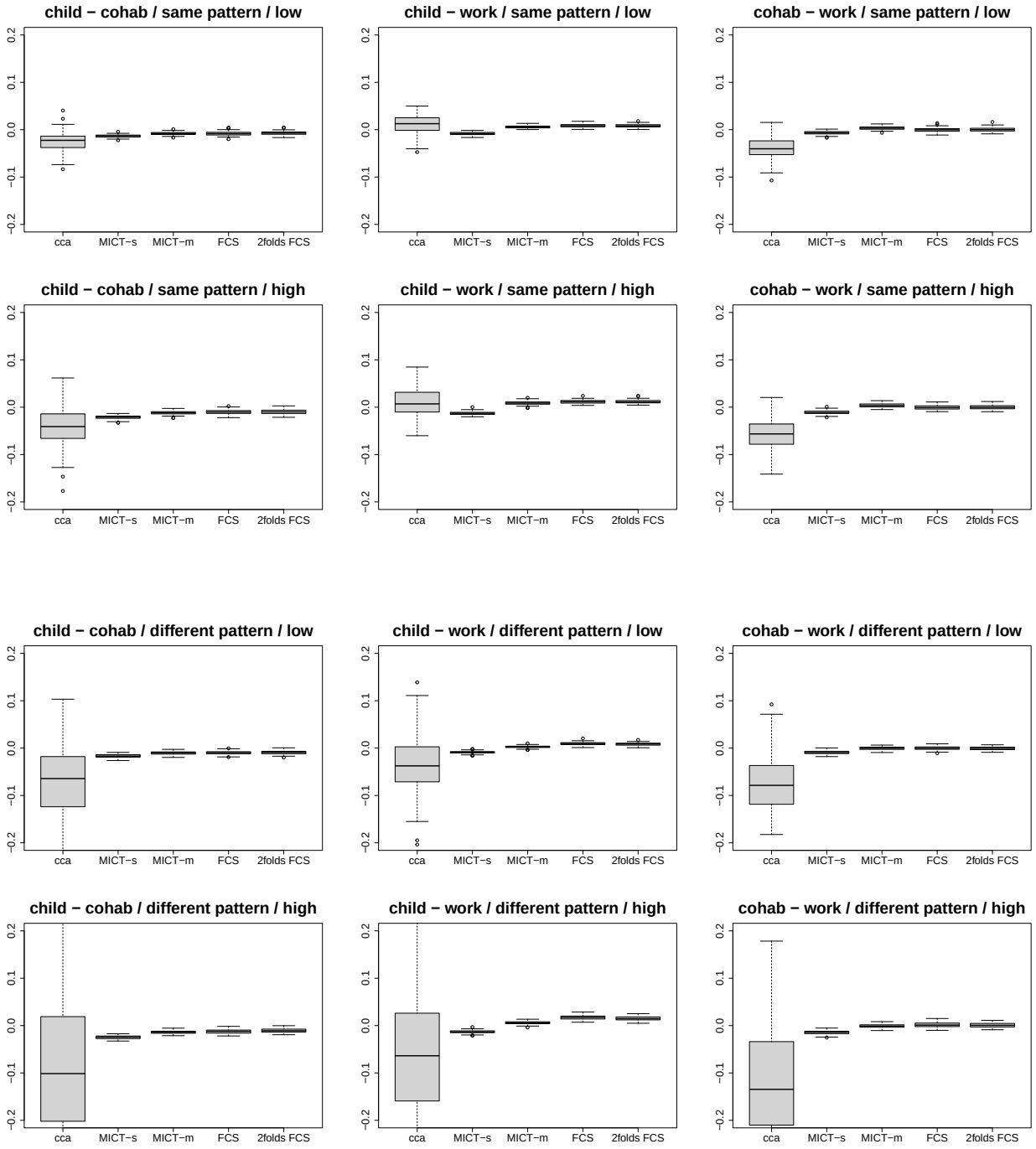



Figure 6.19: MAR mechanism - Boxplots of the bias of the Cronbach's $\alpha$ computed with each pair of channels, obtained with the five considered methods, namely *CCA*, *MICT* applied to each channel separately (labelled as "MICT-s"), *MICT-multichannel* (labelled as "MICT-m"), *FCS* and *two-fold FCS* (labelled as "2folds FCS"). Each subplot corresponds to a scenario of missing data generation. It is labelled as "channels considered / type of pattern / rate of missing data".

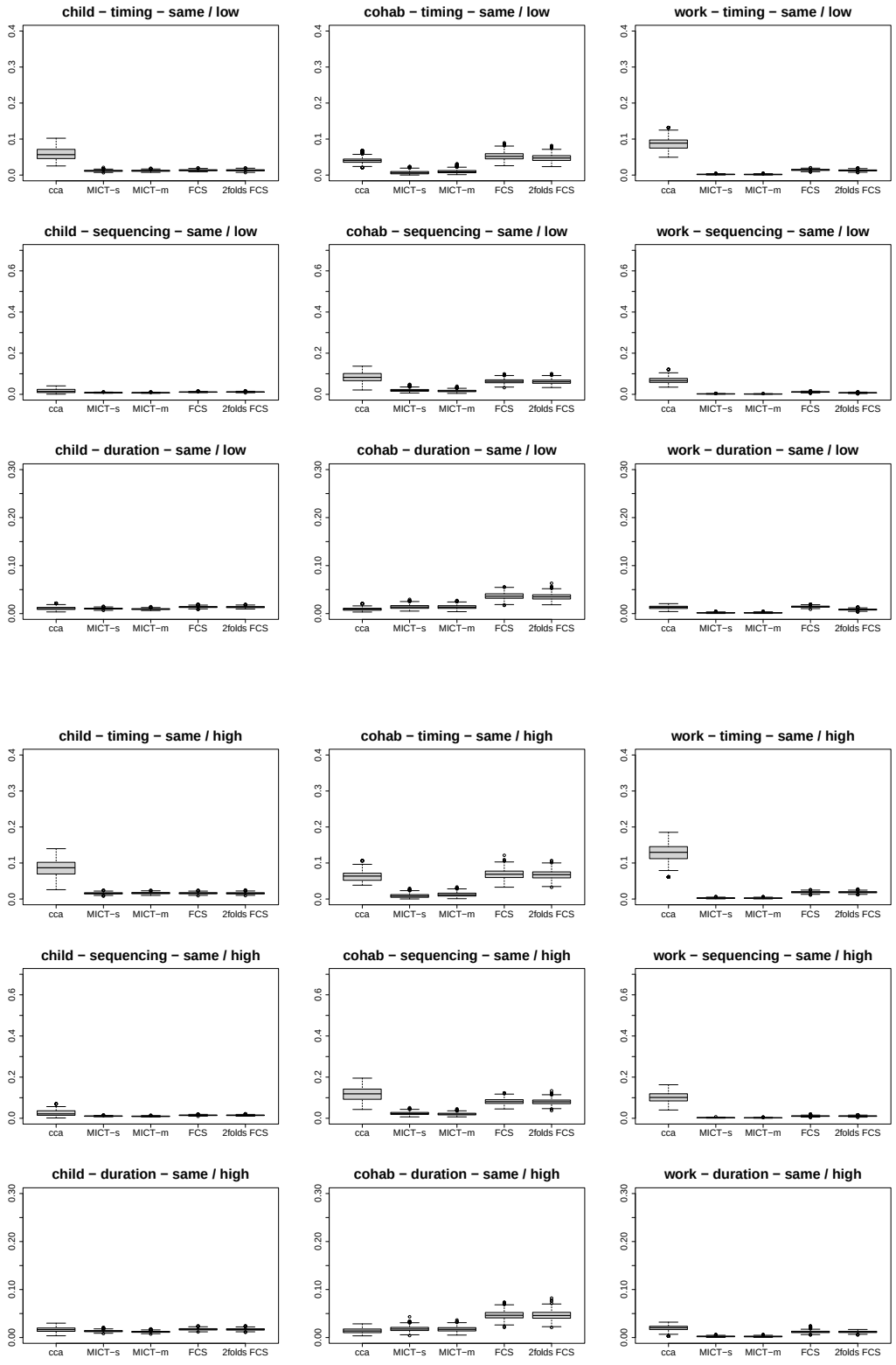



Figure 6.20: MAR mechanism - Boxplots of the bias on the longitudinal criteria, obtained with the five considered methods, namely *CCA*, *MICT* applied to each channel separately (labelled as "MICT-s"), *MICT-multichannel* (labelled as "MICT-m"), *FCS* and *two-fold FCS* (labelled as "2folds FCS"). Each subplot corresponds to a scenario of missing data generation with same patterns on each channel. It is labelled as "channel considered - criterion - type of pattern / rate of missing data".

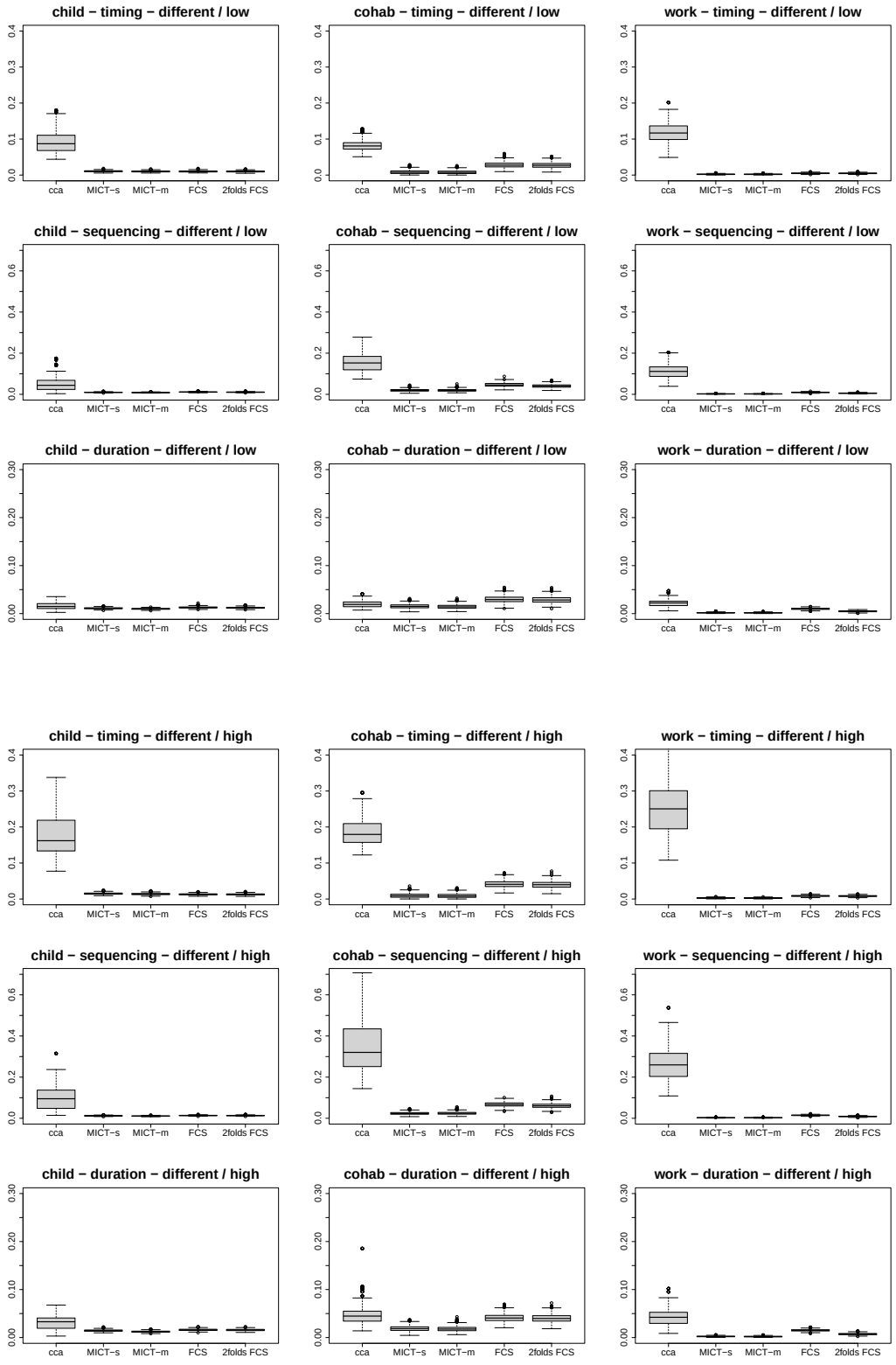



Figure 6.21: MAR mechanism - Boxplots of the bias on the longitudinal criteria, obtained with the five considered methods, namely *CCA*, *MICT* applied to each channel separately (labelled as "MICT-s"), *MICT-multichannel* (labelled as "MICT-m"), *FCS* and *two-fold FCS* (labelled as "2folds FCS"). Each subplot corresponds to a scenario of missing data generation with different pattern on each channel. It is labelled as "channel considered - criterion - type of pattern / rate of missing data".

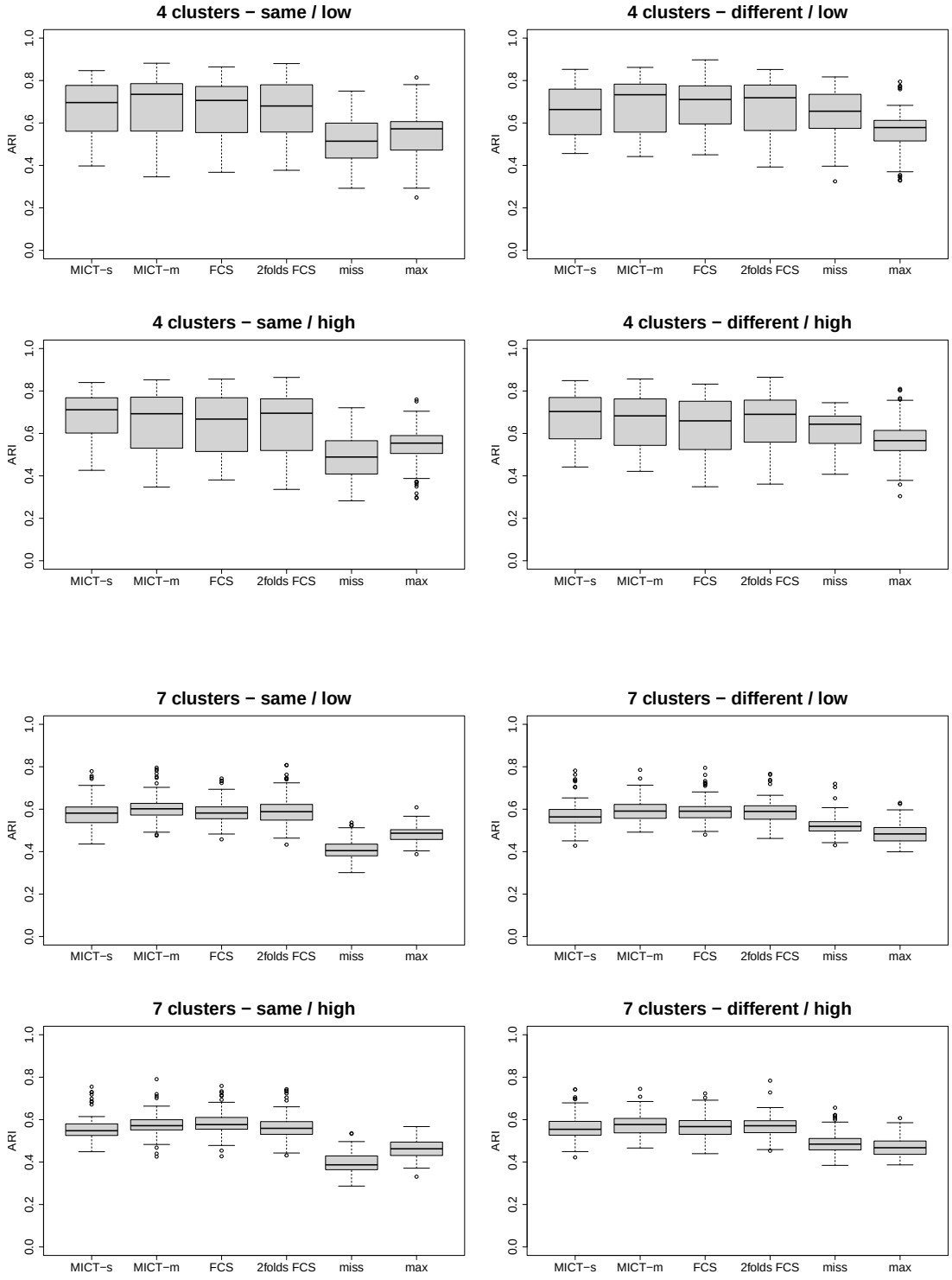



Figure 6.22: MSA - Boxplots of the ARI. The imputation methods considered are *MICT* applied to each channel separately (labelled as "MICT-s"), *MICT-multichannel* (labelled as "MICT-m"), *FCS* and *two-fold FCS* (labelled as "2folds FCS"). In addition, the two ad-hoc methods linked to sequence analysis are labelled as "miss" for the method that consider a missing state as another state and "max" for the method that consider a missing state as maximally different from any other state, including a missing state itself. Each subplot is labelled as "clustering - type of pattern / percentage of missing data". Each boxplot is built with 100 values, corresponding to the 100 datasets with missing data.

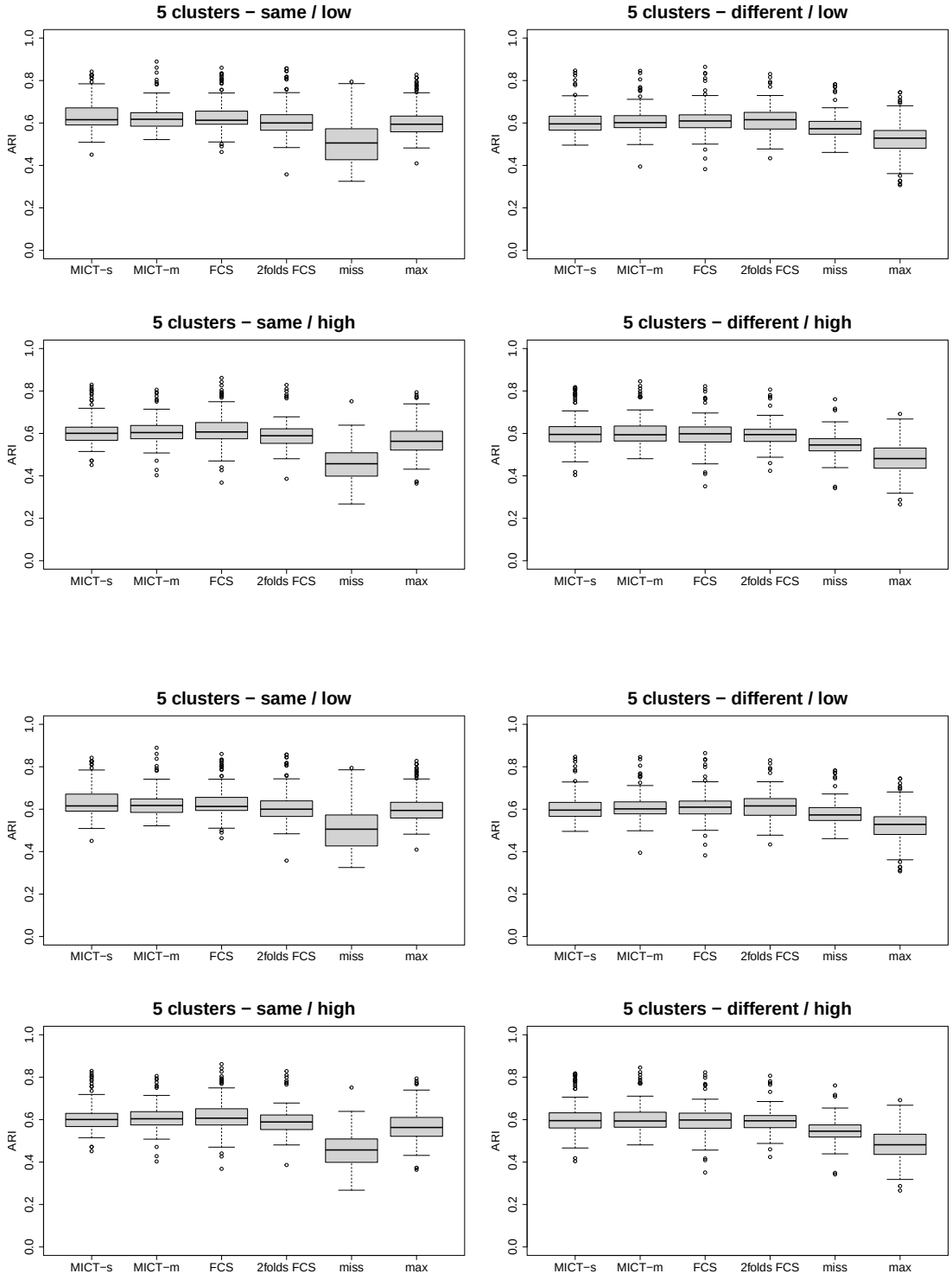



Figure 6.23: EA - Boxplots of the ARI. The imputation methods considered are *MICT* applied to each channel separately (labelled as "MICT-s"), *MICT-multichannel* (labelled as "MICT-m"), *FCS* and *two-fold FCS* (labelled as "2folds FCS"). In addition, the two ad-hoc methods linked to sequence analysis are labelled as "miss" for the method that consider a missing state as another state and "max" for the method that consider a missing state as maximally different from any other state, including a missing state itself. Each subplot is labelled as "clustering - type of pattern / percentage of missing data". Each boxplot is built with 100 values, corresponding to the 100 datasets with missing data.

## 6.5 Discussion

In this chapter, we developed an extension of the MICT imputation algorithm to the case of multichannel sequences. The proposed algorithm, which we call MICT-multichannel, keeps the core functioning of the MICT algorithm, which is to fill gaps of missing data recursively from their edges, while ensuring a cross-sectional consistency by taking into account information from the other channels. In addition, we intended to provide preliminary guidelines on the treatment of missing data in multichannel sequences.

MICT-multichannel fulfils its goal. Indeed, it brings an added value to the MICT algorithm by ensuring a cross-sectional consistency between the channels without sacrificing the longitudinal characteristics. It is even better in terms of the timing criterion. Moreover, it emerged as the preferred method when it comes to handling missing data in multichannel sequences. Indeed, it is globally better than the other tested imputation methods, complete case analysis and ad hoc methods tailored to sequence analysis.

We conducted a comparative analysis of the MICT-multichannel algorithm against other methods for handling missing data. Our comparison was based on two distinct frameworks, both focusing on simulating patterns of missing data that closely resemble real-world scenarios within complete multichannel trajectories. The first framework aimed to assess the algorithm's performance in relation to longitudinal stability and channel association. It involved the creation of multichannel datasets by duplicating a single channel and permuting a percentage of sequences within the duplicated channel. This framework has the advantage to offer a controlled setting for the evaluation of the methods but has the limitation of generating multichannel trajectories with either perfectly associated channels or nearly completely disassociated ones. To address this limitation, the second simulation framework used real multichannel trajectories, providing a more realistic representation. Moreover, this framework allows comparing the impact of methods to handle missing data on clustering results. By employing both frameworks, we ensured an evaluation that encompasses a range of scenarios, from varying association and rates of transition to more realistic data settings.

MICT-multichannel appears as the best imputation method with multichannel sequences. Its performance is especially better than FCS and two-fold FCS in terms of cross-sectional and longitudinal characteristics on multichannel sequences subject to few transitions. Since we built MICT-multichannel to keep the core functioning of MICT, which is to fill gaps of missing data recursively from the edges, it is especially suited to recreate long spells of states, which are typical of trajectories with few transitions. On the other hand, in this case, FCS and two-fold FCS tend to create too many transitions. However, even if on datasets with a high transition rate MICT-multichannel is not always the best, it is close to the best performing one. On datasets with many transitions, the longitudinal logic is less salient, so the strength of MICT filling the gaps from the edges is reduced. Regarding clustering, it shows similar performance

as the other multiple imputation algorithms.

CCA is not suitable for handling missing data in multichannel sequences. In most scenarios and on every criterion, CCA lead to worse results than MICT-multichannel. Moreover, we randomly chose the channels on which we simulated missing data. However, specific life courses, such as the ones with many transitions or vulnerable situations, are more prone to missing. Therefore, the performance of CCA may be even worse in such scenarios.

Concerning FCS and two-fold FCS, their performance was often close in the scenarios we have considered. However, in the second simulation framework, two-fold FCS was sometimes better than FCS in terms of duration and sequencing. Since the second simulation framework was based on three channels, while the duplicated dataset framework was based on only two channels, increasing the number of channels may make two-fold FCS more suitable. Except for channels subject to many transitions, both FCS and two-fold FCS lag behind MICT-multichannel.

Both ad hoc methods for handling missing data in sequence analysis proved to be inadequate. Treating a missing state as an additional state was particularly ineffective compared to imputation methods, especially when the same pattern of missing data was applied to each channel. This approach unintentionally induced undesired similarity among multichannel sequences that have missing data in the same locations. Moreover, even when considering different patterns, this method fell behind imputation methods. An explanation is that the substitution cost between a missing state and any observed state is fixed at one in this method. Consequently, if the value that is missing is the same as an observed state on another channel, the method incorrectly assigns a substitution cost of one instead of the correct value of zero. If they impute the value correctly, imputation methods solve this issue. On the other hand, the problem would become even worse if an imputation method wrongly imputes a value. In our analysis, each imputation method produced imputations of sufficient quality to surpass the performance of this ad hoc method.

The other ad hoc method, namely considering a missing state as maximally different from any other state, avoids the issue of generating unwanted similarities. However, in the case of two missing values that "hide" the same values, it wrongly assumes that they are different. If an imputation method correctly impute these values, it overcomes this issue. It is what appears to happen in our analysis since this ad hoc method lag behind imputation methods in most scenarios. The only case where it presented close performance with the imputation methods is the case of the same patterns of missing data applied to each channel and a combination of information between channels through an extended alphabet. The reason is that in the case of an extended alphabet, imputation methods need to impute accurately each channel to overcome the ad hoc method.

**Guidelines**

Although further research is needed, particularly regarding the parametrisation of the MICT-multichannel algorithm, we can derive preliminary guidelines to choose a method to handle missing data in multichannel sequences.

- Our results demonstrated that complete case analysis and the two ad hoc methods tailored to sequence analysis, which involve considering missing data as an additional state and treating a missing state as maximally distinct from any other state (including a missing state), should not be the default choices.
- In our simulations, MICT-multichannel emerged as the preferred algorithm among the various imputation methods evaluated.
- The MICT-multichannel algorithm provides the flexibility to adjust two parameters: the order of the channels and the number of iterations. While the order of imputation of the channels did not yield significant differences according to the analysis, it is recommended to prioritise channels that are easier to impute—those with greater stability or less missing data. This approach minimises the introduction of erroneous imputed values during the initial phase, which can be challenging to rectify later on. Regarding the number of iterations, an application with two iterations is the recommended choice.
- As for the predictors, the minimal model should incorporate one observation from the past and future, along with all values from the same time point as the missing value being imputed. If there is reason to suspect that other time points or covariates influence the imputation or missing data, it is recommended to include them in the imputation process as well.
- When analysing single trajectories, it is crucial to consider the information from other linked channels if available, as they can significantly improve the quality of imputations.

This chapter introduced an extension of the MICT algorithm dedicated to the imputation of multichannel sequences. The method is dedicated to the imputation of multichannel sequences. However, this work suffers from some limitations and opens the door for further research. First, even if we have considered several scenarios based on duplicated individual trajectories to control the association and a real dataset, we did not fully capture the diversity of situations related to missing data in multichannel sequences. In particular, the framework based on duplicated channels creates either completely associated or mostly disassociated channels. In reality, a larger range of situations appear. This scenario already allowed us to draw interesting conclusions regarding the algorithms, and the second simulation framework strengthens the

idea that the results are not specific to the duplicated datasets. However, we observed that Cronbach's $\alpha$ tends to get overestimated on imputed datasets in comparison with the original ones. Further investigation is needed to determine if imputation methods really tend to increase the relation between channels or if this behaviour is due to the form of the simulation framework or even Cronbach's $\alpha$. The latter is a recently developed tool and we may lack hindsight on how it behaves and what exactly it measures. Specifically, the measurement is contingent upon the dissimilarity measure chosen, and the interaction between the measure itself and the choice of dissimilarity measure remains unclear. Nevertheless, Cronbach's $\alpha$ emerges as the most reliable tool presently accessible for quantifying the overall association. Moreover, regarding the clustering framework, we have considered only one dataset, one way to compute the pairwise dissimilarity measures and one clustering algorithm. Moreover, we applied a simplified setting where the number of clusters is known beforehand. However, our analysis allows giving a first insight on this issue.

Then, not all interrogations regarding the parametrisations of the algorithm have been answered. In this study, we did not vary the number of predictors in the imputation models. We considered only one observation from the past and future, and the observed values at the same time point to the missing data to impute. In particular, we saw in the Chapter 4 that MICT worked better with five predictors both in the past and future, particularly on volatile trajectories. We can expect that it is still the case with the MICT-multichannel algorithm. However, adding too many predictors may hurt the performance of the multinomial model used for the imputation. Moreover, even considering the number of iterations, or the order of the channels, further investigations are needed. For example, multichannel sequences with a high amount of missing data may need more than two iterations, and the order may become crucial with many channels.

Finally, we only considered information from the multichannel trajectories in the imputation process. However, as emphasised in the previous chapter, relevant covariates should be included in the imputation process. Along the same line, we could also add information about survey metadata in the imputation process, such as the mode of collection in a survey. For example, the Swiss household panel, from which we constructed our datasets, is a mixed-mode data survey and, as mentioned in the introduction, the mode of data collection influences the presence of missing data. Incorporating these covariates would likely enhance the imputation quality, without altering the algorithms ranking.

To conclude, MICT-multichannel is a promising method for handling missing data in multichannel sequences. It outperforms in many situations other imputation algorithms. The results are notably better regarding trajectories subject to few transitions, which is a common feature of life course data.

# Chapter 7

# Conclusion

This thesis mainly focuses on handling missing data in life course data. The treatment of missing data is a complex issue that can significantly impact statistical results and conclusions, making it one of the most crucial methodological advancements in this field (Piccarreta and Studer, 2019; Liao et al., 2022). The thesis aims to address the challenge of missing data in life course analysis by proposing innovative methods and approaches to improve the treatment of missing data in longitudinal categorical data. In this concluding chapter, we provide a summary of each chapter (with the exception of the introduction and the state-of-the-art chapters), highlighting the main contributions and the reasons behind their necessity. Furthermore, we discuss the developments made to the *seqimpute* package and suggest potential directions for further development.

## 7.1 Multiple imputation in longitudinal datasets

The structure of longitudinal surveys, as well as the variables derived from them, are inherently complex. Dealing with missing data from such surveys requires special considerations. In particular, we often encounter logical missing values that arise from questions that are irrelevant or nonsensical to certain individuals. Additionally, the presence of categorical variables, which are common in sociological surveys, further complicates the process of treating missing data, especially when employing multiple imputation techniques. However, existing studies have only provided partial guidance on addressing these challenges.

To address these gaps, we have developed in Chapter 3 a comprehensive framework for handling missing data in the context of longitudinal surveys. It encompasses two steps. First, the distribution of missing data is studied. Then, a multiple imputation process is applied, which takes into account the main characteristics of a longitudinal dataset. In particular, our framework addresses two significant challenges: logical missing and categorical data. Conventional imputation algorithms, such as FCS, often fail to consider the issue of logical missing values,

potentially leading to unrealistic imputations and distorted relationships between variables. Although post-imputation corrections can partially mitigate these issues, they do not provide a comprehensive solution. Furthermore, categorical data present additional complexities compared to continuous variables. The performance of logistic and multinomial models is linked to the number of variables included, and the presence of rare categories introduces challenges at various levels. Along this process, we devised a sequence of questions to determine which values are really missing, to set either retrievable or deducible values and, among the remaining missing values, to determine the ones that should be imputed.

We have successfully applied this framework to a subsample of the LIVES-FORS Cohort survey, serving two purposes. Firstly, this application provides empirical evidence to validate and illustrate the effectiveness of our imputation process. Secondly, this application sheds light on the broader challenges associated with managing missing data in longitudinal survey datasets. It highlights the importance of acknowledging that while our framework serves as a foundation for missing data treatment, each dataset possesses unique characteristics that must be considered. The researcher's understanding and familiarity with the dataset become crucial in navigating these specificities and making informed decisions.

Summarising, the main contributions of this chapter are:

- A missing data treatment process specifically designed to address the key challenges encountered in longitudinal surveys
- The treatment of missing data in a concrete case with our method

## 7.2 Imputation of life course data

Life course methodology lacks guidelines regarding the treatment of missing data.While previous comparisons of imputation methods in longitudinal data do exist, they did not focus on categorical data, and hence, the methods specifically dedicated to it, such as MICT and VLMC were never considered. Therefore, in Chapter 4, we aimed to fill this gap by making such a comparison and providing guidelines for effectively treating missing data in life courses.

Additionally, we have introduced two extensions to the MICT algorithm, including random forest imputation models and an algorithm called MICT-timing, which is better suited for non-stationary processes. Random forest has theoretically appealing properties, such as the capacity to incorporate many predictors or non-linear effects. MICT-timing only uses configurations that are temporally close to the missing data to impute instead of all configurations, regardless of their position in the trajectory, as with the MICT algorithm.

In order to conduct this comprehensive comparison, we have devised a simulation framework. Six datasets representative of life course research were chosen. On these datasets, we simulated

missing data according to three scenarios, mimicking real patterns: a MAR process that mimics the missing data that occurs in longitudinal data, an attrition process simulating individuals leaving a study and a MAR process on randomly drawn subsamples. We then evaluated the imputation methods based on criteria relevant to life course research, including timing, duration and sequencing of a process.

We demonstrated that keeping only trajectories without missing data is an unsuitable strategy. Among the imputation algorithms considered, MICT applied with a multinomial imputation is the best. The random forest model, which has theoretically appealing properties, such as the capacity to incorporate many predictors or non-linear effects, did not prove suitable for the imputation of longitudinal categorical data. Despite its ability to impute individual values well, it tends to compromise longitudinal consistency. On the other hand, the MICT-timing variant was found to be a suitable alternative to MICT when facing non-stationary processes.

Although our framework focuses primarily on life course studies, its conclusions are more broadly applicable to the analysis of longitudinal categorical data. Notably, our comparisons included datasets that deviated from traditional life course datasets, such as those examining civil status or satisfaction with health status. Moreover, while only the MAR mechanism was simulated, the conclusions should extend to the MCAR case.

Summarising, the main contributions of this chapter are:

- The MICT-timing imputation algorithm
- Random forest as imputation model for MICT and MICT-timing
- Guidelines regarding the imputation of multichannel sequences
- A simulation framework to compare the performance of imputation method on single trajectories

## 7.3 Comparison of MSA and EA in a real case

When focusing on the joint study of several trajectories, two methods are mainly available: multichannel sequence analysis (MSA) and the extended alphabet (EA). MSA extends optimal matching to the case of multichannel sequences, while with EA, the alphabets for each channel are combined in an extended alphabet. The methods were compared through a simulation by Gauthier et al. (2010) and informal guidelines exist (e.g. Piccarreta (2017)), but no extensive comparison was ever made, in particular using real data. The goal of Chapter 5 was therefore to compare empirically these methods, via cluster analysis, providing guidelines on their use. For this aim, we have introduced a framework based on the last methodological advances to compare joint typologies.

To make this comparison, we used data from the Swiss household panel. We constructed four channels - health, child, cohabitational and professional status - to examine different degrees of association. Since at least professional and child trajectories proved different between men and women, we realised the analysis separately by sex. We identified key characteristics of interest when a clustering of multichannel sequences is performed. These characteristics were measured using the last methodological advances in sequence analysis. The clusterings obtained with both methods were compared based on this framework.

Our findings suggest that neither MSA nor EA is inherently superior. MSA is easier to use and applicable in a broader range of situations, but there are specific situations where EA may be more appropriate. These include cases where extended states represent different sociological realities, when rare events are of interest and when a dissimilarity not based on OM is used.

Summarising, the main contributions of this chapter are:

- A comparison between multichannel sequence analysis and the extended alphabet on a real case
- A framework to compare multichannel typologies

## 7.4 Multichannel imputations

The analysis made in Chapter 4 has shown that the MICT algorithm, or in some instances MICT-timing, is the preferred method when treating missing values in life course data. However, both methods lack an extension specifically designed for imputing multichannel sequences. Moreover, existing methods to treat missing data were not fully satisfying in this case. Thus, the primary contribution of Chapter 6 is the development of such an extension, called MICT-multichannel. Additionally, our objective was to provide preliminary guidelines for effectively imputing multichannel sequences.

We have devised two simulation frameworks both to test the best parametrisation of the MICT-multichannel algorithm and compare it to existing methods. Both frameworks are based on the simulation of patterns of missing data on complete datasets as realistic as possible. The first simulation framework, based on the duplication and permutation of single trajectories, allows controlling for the level of longitudinal and cross-sectional information. However, the association between channels is approximated with this framework and may not be perfect, hence clustering these multichannel sequences may be unrealistic. Therefore, we introduced a second simulation framework, that is based on a real multichannel dataset. With this setting, we were able to compare the impact of the methods to treat missing data on clustering results.

The MICT-multichannel algorithm's functioning keeps the MICT algorithm's core functioning, ensuring longitudinal consistency by recursively filling the gaps of missing data from

the edges while inducing cross-sectional consistency between the channels. In addition to the parameters relative to the MICT algorithm, two parameters are modifiable by the user: the number of iterations and the order of the channels. While the order of the channels does not impact the quality of the imputation, two iterations seem the best choice.

This chapter allowed us to provide preliminary guidelines regarding the treatment of missing data in multichannel sequences. First, even when a single channel is the goal of the analysis, using other associated channels, and impute them as multichannel channels increase the quality of the imputation. Then, consistently with the previous chapter, complete case analysis should not be the default choice. Finally, our results showed that MICT-multichannel generally outperforms the other considered methods, particularly on trajectories with low transition rates.

Summarising, the main contributions of this chapter are:

- The MICT-multichannel imputation algorithm
- Preliminary guidelines regarding the imputation of multichannel sequences
- Simulation frameworks for the comparison of imputation methods

## 7.5 Developments of seqimpute

Throughout this thesis, several tools were developed to address the challenges of imputing missing data in longitudinal categorical data. A key objective was to make these tools accessible to researchers, and to achieve this, we have integrated them into the *seqimpute* package within the *R* statistical software.

Previously, the MICT imputation algorithm had been implemented in this package. However, within the context of this thesis, significant improvements were made to enhance the speed of imputation using multinomial models, and the random forest algorithm was introduced as an additional option for the imputation process. These changes have been incorporated into version 1.8 of the package. Furthermore, in version 1.9, both the MICT-timing and MICT-multichannel algorithms have been made available. At the time of writing, this latest version is only available on *R-Forge*, a web platform for developing *R* packages, but it will soon be made available on *CRAN*.

## 7.6 Further developments

This thesis addressed several issues but also opened up many doors. In the conclusion of each of the four main chapters, potential developments were already discussed. To end this thesis, we focus on the required developments. More particularly, we discuss what we frame as the

four most needed developments, namely the inclusion of covariates in the imputation processes, the study of minimum information needed to impute, the link between treatment of missing data and the statistical analysis, and the writing of a user manual for the *seqimpute* package.

The inclusion of covariates was considered neither for single, nor multichannel trajectories. However, as highlighted throughout this thesis, there are notable differences in the professional transitions experienced by men and women. Therefore, incorporating sex as a covariate in the imputation process would likely enhance the quality of the results. Additionally, it is generally recommended to include all variables used in the statistical analysis within the imputation models to mitigate potential bias (see e.g. Van Buuren, 2018). Consequently, exploring the impact of covariate inclusion on the different imputation procedures becomes an important consideration. By extending the information included in the imputation models through the addition of covariates, the use of random forest as an imputation method may become more suitable. Random forest excels in handling large pools of predictors and effectively manages complex relationships between variables.

When considering the implementation of multiple imputation in life course data, an important question arises regarding the minimum amount of information required for its application. If a trajectory consists of only a few observed values, there is a risk that imputation might do more harm than simply excluding that trajectory. To illustrate, let's consider an extreme scenario where we only have information about an individual's professional situation at age 18, with missing data between the ages of 19 and 40. In such cases, imputing values would likely introduce substantial variability, thereby increasing the variability of estimators computed from this sample. Moreover, if the missing data follows a MNAR mechanism, the imputation may introduce additional bias. In light of this, a potential solution proposed by Seaman et al. (2012) is to combine multiple imputation with weighting methods. Specifically, trajectories with a percentage of missing data below a predefined threshold would be imputed, while those with a higher proportion of missing data would be excluded. Finally, the imputed trajectories would be reweighted to account for the excluded ones.

Further developments are directly associated with the impact of imputation on statistical analyses. Regarding the derivation of statistical outcomes with multiple imputation, we have seen that the strategy of Halpin (2012) that consists in stacking all imputed datasets together before proceeding with the clustering, appears as a promising strategy. However, we lack a comparison with other strategies, such as consensus clustering (Basagaña et al., 2013) or the *MultiCons* approach, which consists in building a unique clustering from several imputed datasets based on their common clustering patterns (Al-Najdi et al., 2016). Moreover, in life course analysis, the built typology is often used as a dependent or independent variable in a subsequent regression analysis. Therefore, we have an uncertainty induced by missing data both on the typology itself and the further regression analysis. The question is how to deal

with it. In addition, there is an inherent uncertainty when clustering and one can ask how it interacts with the uncertainty coming from missing data. Furthermore, there is a pressing need for additional research to thoroughly examine the influence of multiple imputation methods, as well as other approaches to address missing data, on statistical results. In Chapter 6, we initiated an investigation into this matter by comparing the effects of various methods on clustering outcomes. However, further exploration is warranted. Initially, we conducted our analysis in a simplified context where the number of groups was predetermined. However, in most practical applications, this is not the case. Moreover, comparing different methods for handling missing data in such analyses would contribute to extending the knowledge of the most appropriate methods for each specific context.

Finally, a crucial aspect of introducing new methodological tools and developments is ensuring their user-friendliness. In the context of this thesis, the methods developed were made accessible through the *seqimpute* package. However, to enhance their usability, the creation of a comprehensive user manual is of utmost importance. This user manual should include clear examples demonstrating the application of the developed methods, thereby simplifying their usage and promoting their wider adoption.

# Bibliography


Abbott, A. (1983). Sequences of social events: concepts and methods for the analysis of order in social processes. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 16(4):129–147.

Abbott, A. (1995). Sequence analysis: New methods for old ideas. *Annual Review of Sociology*, 21(1):93–113.

Abbott, A. and Forrest, J. (1986). Optimal matching methods for historical sequences. *The Journal of Interdisciplinary History*, 16(3):471–494.

Abbott, A. and Tsay, A. (2000). Sequence analysis and optimal matching methods in sociology: Review and prospect. *Sociological Methods & Research*, 29(1):3–33.

Aeby, G., Gauthier, J.-A., and Widmer, E. D. (2019). Beyond the nuclear family: Personal networks in light of work-family trajectories. *Advances in Life Course Research*, 39:51–60.

Aisenbrey, S. and Fasang, A. (2017). The interplay of work and family trajectories over the life course: Germany and the United States in comparison. *American Journal of Sociology*, 122(5):1448–1484.

Al-Najdi, A., Pasquier, N., and Precioso, F. (2016). Frequent closed patterns based multiple consensus clustering. In *Artificial Intelligence and Soft Computing: 15th International Conference, ICAISC 2016, Zakopane, Poland, June 12-16, 2016, Proceedings, Part II 15*, pages 14–26. Springer.

Albatineh, A. N., Niewiadomska-Bugaj, M., and Mihalko, D. (2006). On similarity indices and correction for chance agreement. *Journal of Classification*, 23(2):301–313.

Allison, P. D. (2001). *Missing data*. SAGE.

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., and Perona, I. (2013). An extensive comparative study of cluster validity indices. *Pattern recognition*, 46(1):243–256.

Arora, V. S., Karanikolos, M., Clair, A., Reeves, A., Stuckler, D., and McKee, M. (2015). Data resource profile: the European Union statistics on income and living conditions (eu-silc). *International journal of epidemiology*, 44(2):451–461.

Arpino, B., Gumà, J., and Julià, A. (2018). Early-life conditions and health at older ages: The mediating role of educational attainment, family and employment trajectories. *PloS one*, 13(4):e0195320.

Aßmann, C., Würbach, A., Goßmann, S., Geissler, F., and Bela, A. (2017). Nonparametric multiple imputation for questionnaires with individual skip patterns and constraints: The case of income imputation in the national educational panel study. *Sociological Methods & Research*, 46(4):864–897.

Bartlett, J. W., Harel, O., and Carpenter, J. R. (2015). Asymptotically unbiased estimation of exposure odds ratios in complete records logistic regression. *American journal of epidemiology*, 182(8):730–736.

Basagaña, X., Barrera-Gómez, J., Benet, M., Antó, J. M., and Garcia-Aymerich, J. (2013). A framework for multiple imputation in cluster analysis. *American journal of epidemiology*, 177(7):718–725.

Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *The annals of mathematical statistics*, 37(6):1554–1563.

Berchtold, A. (1999). The double chain Markov model. *Communications in Statistics-Theory and Methods*, 28(11):2569–2589.

Berchtold, A. (2019). Treatment and reporting of item-level missing data in social science research. *International Journal of Social Research Methodology*, 22(5):431–439.

Berchtold, A., Guinchard, A., Emery, K., and Taher, K. (2022). seqimpute: Imputation of missing data in sequence analysis. R package version 1.8.

Berchtold, A. and Raftery, A. (2002). The mixture transition distribution model for high-order Markov chains and non-Gaussian time series. *Statistical Science*, 17(3):328–356.

Berger, J. O. and Bernardo, J. M. (1992). On the development of the reference prior method. *Bayesian statistics*, 4(4):35–60.

Bernardi, L., Huinink, J., and Settersten Jr, R. A. (2019). The life course cube: A tool for studying lives. *Advances in Life Course Research*, 41:100258.

Biau, G. and Scornet, E. (2016). A random forest guided tour. *Test*, 25(2):197–227.

Billari, F. C. and Piccarreta, R. (2005). Analyzing demographic life courses through sequence analysis. *Mathematical population studies*, 12(2):81–106.

Bodner, T. E. (2008). What improves with increased missing data imputations? *Structural equation modeling: a multidisciplinary journal*, 15(4):651–675.

Bradley, P. E. (2019). Methodology for the sequence analysis of building stocks. *Building Research & Information*, 47(2):141–155.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Brémaud, P. (2013). *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*, volume 31. Springer.

Brum-Bastos, V. S., Long, J. A., and Demšar, U. (2018). Weather effects on human mobility: a study using multi-channel sequence analysis. *Computers, Environment and Urban Systems*, 71:131–152.

Brzinsky-Fay, C. and Solga, H. (2016). Compressed, postponed, or disadvantaged? School-to-work-transition patterns and early occupational attainment in West Germany. *Research in Social Stratification and Mobility*, 46:21–36.

Bühlmann, P. and Wyner, A. J. (1999). Variable length Markov chains. *The Annals of Statistics*, 27(2):480–513.

Burgette, L. F. and Reiter, J. P. (2010). Multiple imputation for missing data via sequential regression trees. *American journal of epidemiology*, 172(9):1070–1076.

Buskirk, T. D. and Kolenikov, S. (2015). Finding respondents in the forest: A comparison of logistic regression and random forest models for response propensity weighting and stratification. *Survey Methods: Insights from the Field*, pages 1–17.

Chassan, M. and Concordet, D. (2023). How to test the missing data mechanism in a hidden markov model. *Computational Statistics & Data Analysis*, 182:107723.

Coulter, R. and Van Ham, M. (2013). Following people through time: An analysis of individual residential mobility biographies. *Housing Studies*, 28(7):1037–1055.

Croissant, Y. (2020). Estimation of random utility models in R: the mlogit package. *Journal of Statistical Software*, 95:1–41.

Cutler, A., Cutler, D. R., and Stevens, J. R. (2012). Random forests. In Zhang, C. and Ma, Y., editors, *Ensemble machine learning*, pages 157–175. Springer.

Daikeler, J., Bošnjak, M., and Lozar Manfreda, K. (2020). Web versus other survey modes: an updated and extended meta-analysis comparing response rates. *Journal of Survey Statistics and Methodology*, 8(3):513–539.

Dangubic, M. and Voorpostel, M. (2017). Refusal conversion in the Swiss household panel 1999-2015: An overview. *FORS Working Papers Series*, 17-2. Swiss Centre of Expertise in the Social Sciences, Lausanne.

Daniels, M. J. and Hogan, J. W. (2008). *Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis*. CRC press.

Daniels, M. J. and Hogan, J. W. (2014). Bayesian methods for incomplete data. In Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A., and Verbeke, G., editors, *Handbook of missing data methodology*, pages 91–116. CRC Press.

De Jong, V. M., Eijkemans, M. J., Van Calster, B., Timmerman, D., Moons, K. G., Steyerberg, E. W., and Van Smeden, M. (2019). Sample size considerations and predictive performance of multinomial logistic prediction models. *Statistics in medicine*, 38(9):1601–1619.

De Silva, A. P., Moreno-Betancur, M., De Livera, A. M., Lee, K. J., and Simpson, J. A. (2017). A comparison of multiple imputation methods for handling missing values in longitudinal data in the presence of a time-varying covariate with a non-linear association with time: a simulation study. *BMC medical research methodology*, 17(1):1–11.

Décieux, P. J., Mergener, A., Neufang, M. K., and Sischka, P. (2015). Implementation of the forced answering option within online surveys: do higher item response rates come at the expense of participation and answer quality? *Psihologija*, 48(4):311–326.

Demirtas, H., Freels, S. A., and Yucel, R. M. (2008). Plausibility of multivariate normality assumption when multiply imputing non-Gaussian continuous outcomes: a simulation assessment. *Journal of Statistical Computation and Simulation*, 78(1):69–84.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Devillanova, C., Raitano, M., and Struffolino, E. (2019). Longitudinal employment trajectories and health in middle life. *Demographic Research*, 40:1375–1412.

Deville, J.-C. and Saporta, G. (1983). Correspondence analysis, with an extension towards nominal time series. *Journal of econometrics*, 22(1-2):169–189.

Di Giulio, P., Impicciatore, R., and Sironi, M. (2019). The changing pattern of cohabitation: A sequence analysis approach. *Demographic Research*, 40(42):1211–1248.

Dillman, D. A. (2009). Some consequences of survey mode changes in longitudinal surveys. In Lynn, P., editor, *Methodology of longitudinal surveys*, pages 127–140. John Wiley & Sons.

Dixon, W. J. and Brown, M. B. (1983). *BMDP statistical software*, volume 1. University of California Press.

Donders, A. R. T., Van Der Heijden, G. J., Stijnen, T., and Moons, K. G. (2006). A gentle introduction to imputation of missing values. *Journal of clinical epidemiology*, 59(10):1087–1091.

Doove, L. L., Van Buuren, S., and Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational statistics & data analysis*, 72:92–104.

Eekhout, I., de Boer, R. M., Twisk, J. W., de Vet, H. C., and Heymans, M. W. (2012). Missing data: a systematic review of how they are reported and handled. *Epidemiology*, 23(5):729–732.

Eisenberg-Guyot, J., Peckham, T., Andrea, S. B., Oddo, V., Seixas, N., and Hajat, A. (2020). Life-course trajectories of employment quality and health in the US: a multichannel sequence analysis. *Social Science & Medicine*, 264:113327.

Elder, G. H., Kirkpatrick Johnson, M., and Crosnoe, R. (2003). The emergence and development of life course theory. In Mortimer, J. and Shanahan, M., editors, *Handbook of the Life Course*, Handbooks of Sociology and Social Research, pages 3–19. Springer.

Elzinga, C. H. (2005). Combinatorial representations of token sequences. *Journal of Classification*, 22(1):87–118.

Elzinga, C. H. (2015). Comment: on the association between sequences in GIMSA. *Sociological Methodology*, 45(1):45–51.

Elzinga, C. H. and Liefbroer, A. C. (2007). De-standardization of family-life trajectories of young adults: A cross-national comparison using sequence analysis. *European Journal of Population/Revue européenne de Démographie*, 23(3-4):225–250.

Elzinga, C. H. and Studer, M. (2015). Spell sequences, state proximities, and distance metrics. *Sociological Methods & Research*, 44(1):3–47.

Fasang, A. E. (2015). Comment: What's the added value? *Sociological Methodology*, 45(1):56–70.

Fasang, A. E. and Raab, M. (2014). Beyond transmission: Intergenerational patterns of family formation among middle-class american families. *Demography*, 51(5):1703–1728.

Freedman, D., Thornton, A., Camburn, D., Alwin, D., and Young-DeMarco, L. (1988). The life history calendar: A technique for collecting retrospective data. *Sociological methodology*, pages 37–68.

Friedman, J., Hastie, T., Tibshirani, R., et al. (2001). *The elements of statistical learning*. Springer.

Gabadinho, A. and Ritschard, G. (2016). Analyzing state sequences with probabilistic suffix trees: the PST R package. *Journal of statistical software*, 72(1):1–39.

Gabadinho, A., Ritschard, G., Müller, N. S., and Studer, M. (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software*, 40(4):1–37.

Gabadinho, A., Ritschard, G., Studer, M., and Müller, N. S. (2010). Indice de complexité pour le tri et la comparaison de séquences catégorielles. *Revue des nouvelles technologies de l'information RNTI*, E-19:61–66.

Gauthier, J.-A. (2015). Comment: How to make a long story short. *Sociological Methodology*, 45(1):70–73.

Gauthier, J.-A., Widmer, E. D., Bucher, P., and Notredame, C. (2010). 1. Multichannel sequence analysis applied to social science data. *Sociological methodology*, 40(1):1–38.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC.

Gemmill, A. (2019). From some to none? Fertility expectation dynamics of permanently childless women. *Demography*, 56(1):129–149.

Gitinabard, N., Heckman, S., Barnes, T., and Lynch, C. F. (2019). What will you do next? A sequence analysis on the student transitions between online platforms in blended courses. *arXiv preprint arXiv:1905.00928*.

Glynn, R. J. and Laird, N. M. (1986). Regression estimates and missing data: complete case analysis. *Cambridge MA: Harvard School of Public Health, Department of Biostatistics*.

Glynn, R. J., Laird, N. M., and Rubin, D. B. (1986). Selection modeling versus mixture modeling with nonignorable nonresponse. In Wainer, H., editor, *Drawing inferences from self-selected samples*, pages 115–142. Routledge.

Graham, J. W. (2012). *Missing data: Analysis and design*. Springer Science & Business Media.

Graham, J. W., Olchowski, A. E., and Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention science*, 8(3):206–213.

Groves, R. M. (2005). *Survey errors and survey costs*. John Wiley & Sons.

Guo, Y., Kopec, J. A., Cibere, J., Li, L. C., and Goldsmith, C. H. (2016). Population survey features and response rates: a randomized experiment. *American journal of public health*, 106(8):1422–1426.

Halpin, B. (2010). Optimal matching analysis and life-course data: The importance of duration. *Sociological Methods & Research*, 38(3):365–388.

Halpin, B. (2012). Multiple imputation for life-course sequence data. *Department of Sociology Working Paper Series*, University of Limerick.

Halpin, B. (2013). Imputing sequence data: Extensions to initial and terminal gaps, Stata's *mi*. *Department of Sociology Working Paper Series*, University of Limerick.

Halpin, B. (2015). Mict: Stata module to provide multiple imputation for categorical time-series.

Halpin, B. (2016a). Missingness and truncation in sequence data: a non-self-identical missing state. In Gilbert, R. and Matthias, S., editors, *Proceedings of the International Conference on Sequence Analysis and Related Methods*, pages 443–444, Lausanne.

Halpin, B. (2016b). Multiple imputation for categorical time series. *The Stata Journal*, 16(3):590–612.

Halpin, B. (2017). SADI: Sequence analysis tools for Stata. *The Stata Journal*, 17(3):546–572.

Han, S.-K. and Moen, P. (1999). Work and family over time: a life course approach. *The ANNALS of the American Academy of Political and Social Science*, 562(1):98–110.

Han, Y., Liefbroer, A. C., and Elzinga, C. H. (2017). Comparing methods of classifying life courses: sequence analysis and latent class analysis. *Longitudinal and Life Course Studies*, 8(4):319–341.

Hartley, H. O. (1958). Maximum likelihood estimation from incomplete data. *Biometrics*, 14(2):174–194.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of economic and social measurement, volume 5, number 4*, pages 475–492. National Bureau of Economic Research.

Helske, S. and Helske, J. (2017). Mixture hidden Markov models for sequence data: the seqHMM package in R. *arXiv preprint arXiv:1704.00543*.

Hennig, C. and Liao, T. F. (2010). Comparing latent class and dissimilarity based clustering for mixed type variables with application to social stratification. Technical report, Department of Statistical Science, University College London.

Hennig, C. and Lin, C.-J. (2015). Flexible parametric bootstrap for testing homogeneity against clustering and assessing the number of clusters. *Statistics and Computing*, 25(4):821–833.

Herzing, J. M., Elcheroth, G., Lipps, O., and Kleiner, B. (2019). Surveying national minorities. Technical report, University of Lausanne; FORS.

Honaker, J. and King, G. (2010). What to do about missing values in time-series cross-section data. *American journal of political science*, 54(2):561–581.

Honaker, J., King, G., and Blackwell, M. (2011). Amelia II: a program for missing data. *Journal of statistical software*, 45(1):1–47.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2:193–218.

Hubert, L. J. and Levin, J. R. (1976). A general statistical framework for assessing categorical clustering in free recall. *Psychological bulletin*, 83(6):1072.

Hughes, R. A., White, I. R., Seaman, S. R., Carpenter, J. R., Tilling, K., and Sterne, J. A. (2014). Joint modelling rationale for chained equations. *BMC medical research methodology*, 14(1):1–10.

Huque, M. H., Carlin, J. B., Simpson, J. A., and Lee, K. J. (2018). A comparison of multiple imputation methods for missing data in longitudinal studies. *BMC medical research methodology*, 18(1):1–16.

Jackson, Y., Courvoisier, D. S., Duvoisin, A., Ferro-Luzzi, G., Bodenmann, P., Chauvin, P., Guessous, I., Wolff, H., Cullati, S., and Burton-Jeangros, C. (2019). Impact of legal status change on undocumented migrants' health and well-being (parchemins): protocol of a 4-year, prospective, mixed-methods study. *BMJ open*, 9(5):e028336.

Jalovaara, M. and Fasang, A. (2017). From never partnered to serial cohabitors: union trajectories to childlessness. *Demographic Research*, 36:1703–1720.

Jamshidian, M. and Jalal, S. (2010). Tests of homoscedasticity, normality, and missing completely at random for incomplete multivariate data. *Psychometrika*, 75(4):649–674.

Jeffreys, H. (1998). *The theory of probability*. OUP Oxford.

Jowell, R., Roberts, C., Fitzgerald, R., and Eva, G. (2006). *Measuring attitudes cross-nationally: Lessons from the European Social Survey*. SAGE.

Kalaycioglu, O., Copas, A., King, M., and Omar, R. Z. (2016). A comparison of multiple-imputation methods for handling missing data in repeated measurements observational studies. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pages 683–706.

Kalton, G. and Flores-Cervantes, I. (2003). Weighting methods. *Journal of official statistics*, 19(2):81.

Kaufman, L. and Rousseeuw, P. J. (1990). Partitioning around medoids (program pam). *Finding groups in data: an introduction to cluster analysis*, 344:68–125.

Kiecolt-Glaser, J. K. and Wilson, S. J. (2017). Lovesick: How couples' relationships influence health. *Annual review of clinical psychology*, 13:421–443.

King, G., Honaker, J., Joseph, A., and Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American political science review*, 95(1):49–69.

Lacey, R., Stafford, M., Sacker, A., and McMunn, A. (2016). Work-family life courses and subjective wellbeing in the mrc national survey of health and development (the 1946 British birth cohort study). *Journal of population ageing*, 9(1):69–89.

Lall, R. (2016). How multiple imputation makes a difference. *Political Analysis*, 24(4):414–433.

Lay-Yee, R., Matthews, T., Moffitt, T., Poulton, R., Caspi, A., and Milne, B. (2021). Do socially isolated children become socially isolated adults? *Advances in Life Course Research*, 50:100419.

Lesnard, L. (2008). Off-scheduling within dual-earner couples: an unequal and negative externality for family time. *American Journal of Sociology*, 114(2):447–490.

Lesnard, L. (2010). Setting cost in optimal matching to uncover contemporaneous socio-temporal patterns. *Sociological Methods & Research*, 38(3):389–419.

Levy, R., Gauthier, J.-A., and Widmer, E. (2006). Entre contraintes institutionnelle et domestique: les parcours de vie masculins et féminins en suisse. *Canadian Journal of Sociology/Cahiers canadiens de sociologie*, pages 461–489.

Liao, T. F., Bolano, D., Brzinsky-Fay, C., Cornwell, B., Fasang, A. E., Helske, S., Piccarreta, R., Raab, M., Ritschard, G., Struffolino, E., et al. (2022). Sequence analysis: its past, present, and future. *Social science research*, 107:102772.

Liao, T. F. and Fasang, A. E. (2021). Comparing groups of life-course sequences using the bayesian information criterion and the likelihood-ratio test. *Sociological Methodology*, 51(1):44–85.

Link, M. W., Murphy, J., Schober, M. F., Buskirk, T. D., Hunter Childs, J., and Langer Tesfaye, C. (2014). Mobile technologies for conducting, augmenting and potentially replacing surveys: Executive summary of the aapor task force on emerging technologies in public opinion research. *Public Opinion Quarterly*, 78(4):779–787.

Lipps, O., Herzing, J. M., Pekari, N., Ernst Stähli, M., Pollien, A., Riedo, G., and Reveilhac, M. (2019). Incentives in surveys. Technical report, FORS, University of Lausanne.

Little, R. J. (1992). Regression with missing x's: a review. *Journal of the American statistical association*, 87(420):1227–1237.

Little, R. J. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421):125–134.

Little, R. J., Carpenter, J. R., and Lee, K. J. (2022). A comparison of three popular methods for handling missing data: complete-case analysis, inverse probability weighting, and multiple imputation. *Sociological Methods & Research*, page 00491241221113873.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

Little, R. J. and Schenker, N. (1995). Missing data. In Arminger, G., Clogg, C. C., and Sobel, M. E., editors, *Handbook of statistical modeling for the social and behavioral sciences*, pages 39–75. Springer.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404):1198–1202.

Liu, B., Widener, M. J., Smith, L. G., Farber, S., Minaker, L. M., Patterson, Z., Larsen, K., and Gilliland, J. (2022). Disentangling time use, food environment, and food behaviors using multi-channel sequence analysis. *Geographical Analysis*, 54(4):881–917.

Liu, C. and Rubin, D. B. (1994). The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. *Biometrika*, 81(4):633–648.

Liu, C., Rubin, D. B., and Wu, Y. N. (1998). Parameter expansion to accelerate EM: the PX-EM algorithm. *Biometrika*, 85(4):755–770.

Loosveldt, G. and Billiet, J. (2002). Item nonresponse as a predictor of unit nonresponse in a panel survey. *Journal of Official Statistics*, 18(4):545.

Loosveldt, G. and Carton, A. (2001). An empirical test of a limited model for panel refusals. *International Journal of Public Opinion Research*, 13(2):173–185.

Lorentzen, T., Bäckman, O., Ilmakunnas, I., and Kauppinen, T. (2019). Pathways to adulthood: sequences in the school-to-work transition in Finland, Norway and Sweden. *Social Indicators Research*, 141(3):1285–1305.

Lowe, M. R., Holbrook, C. M., and Hondorp, D. W. (2020). Detecting commonality in multidimensional fish movement histories using sequence analysis. *Animal Biotelemetry*, 8(1):1–14.

Luan, J., Zhang, C., Xu, B., Xue, Y., and Ren, Y. (2020). The predictive performances of random forest models with limited sample size and different species traits. *Fisheries Research*, 227:105534.

Lyberg, L. and Weisberg, H. (2016). *Total survey error: a paradigm for survey methodology*. SAGE.

Lynn, P. (2009a). *Methodology of longitudinal surveys*. John Wiley & Sons.

Lynn, P. (2009b). Methods for longitudinal surveys. In Lynn, P., editor, *Methodology of longitudinal surveys*, pages 1–18. John Wiley & Sons.

Malin, L. and Wise, R. (2018). Glass ceilings, glass escalators and revolving doors. In Ritschard, G. and Studer, M., editors, *Sequence Analysis and Related Approaches*, volume 10 of *Life Course Research and Social Policies*, pages 49–68. Springer.

Manzoni, A. and Mooi-Reci, I. (2018). Measuring sequence quality. In Ritschard, G. and Studer, M., editors, *Sequence Analysis and Related Approaches*, volume 10 of *Life Course Research and Social Policies*, pages 261–278. Springer.

Mattijssen, L. and Pavlopoulos, D. (2019). A multichannel typology of temporary employment careers in the Netherlands: identifying traps and stepping stones in terms of employment and income security. *Social science research*, 77:101–114.

McCullagh, P. and Nelder, J. A. (2019). *Generalized linear models*. Routledge.

McKnight, P. E., McKnight, K. M., Sidani, S., and Figueredo, A. J. (2007). *Missing data: A gentle introduction*. Guilford Press.

McMunn, A., Lacey, R., Worts, D., McDonough, P., Stafford, M., Booker, C., Kumari, M., and Sacker, A. (2015). De-standardization and gender convergence in work–family life courses in great britain: A multi-channel sequence analysis. *Advances in Life Course Research*, 26:60–75.

McVicar, D. and 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.

Melnykov, V. et al. (2016). ClickClust: An R package for model-based clustering of categorical sequences. *Journal of Statistical Software*, 74(i09).

Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ecm algorithm: A general framework. *Biometrika*, 80(2):267–278.

Meng, X.-L. and Van Dyk, D. (1997). The em algorithm—an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(3):511–567.

M'Kendrick, A. (1925). Applications of mathematics to medical problems. *Proceedings of the Edinburgh Mathematical Society*, 44:98–130.

Möhring, K. and Weiland, A. P. (2022). Couples' life courses and women's income in later life: a multichannel sequence analysis of linked lives in germany. *European Sociological Review*, 38(3):371–388.

Molenberghs, G., Beunckens, C., Sotto, C., and Kenward, M. G. (2008). Every missingness not at random model has a missingness at random counterpart with equal fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(2):371–388.

Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A., and Verbeke, G. (2014). *Handbook of missing data methodology*. Chapman & Hall/CRC handbooks of modern statistical methods. CRC Press.

Molenberghs, G. and Kenward, M. (2007). *Missing data in clinical studies*. John Wiley & Sons.

Morselli, D., Spini, D., Le Goff, J.-M., Gauthier, J.-A., Brändle, K., Mugnari, E., Dasoki, N., Roberts, C., Bernardi, L., Bühlmann, F., et al. Assessing the performance of the Swiss panel LIVES calendar: evidence from a pilot study. *LIVES working papers*, 28. Swiss National Centre of Competence in Research LIVES, Geneva.

Müller, N. S., Gabadinho, A., Ritschard, G., and Studer, M. (2008). Extracting knowledge from life courses: clustering and visualization. In *International Conference on Data Warehousing and Knowledge Discovery*, pages 176–185. Springer.

Müller, N. S., Sapin, M., Jacques-Antoine, G., Orita, A., and Widmer, E. D. (2012). Pluralized life courses? An exploration of the life trajectories of individuals with psychiatric disorders. *International Journal of Social Psychiatry*, 58(3):266–277.

Murtagh, F. and Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *Journal of classification*, 31(3):274–295.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453.

Nevalainen, J., Kenward, M. G., and Virtanen, S. M. (2009). Missing values in longitudinal dietary data: a multiple imputation approach based on a fully conditional specification. *Statistics in medicine*, 28(29):3657–3669.

Noghrehchi, F., Stoklosa, J., Penev, S., and Warton, D. I. (2021). Selecting the model for multiple imputation of missing data: just use an IC! *Statistics in Medicine*, 40(10):2467–2497.

Oris, M., Guichard, E., Nicolet, M., Gabriel, R., Tholomier, A., Monnot, C., Fagot, D., and Joye, D. (2016). Representation of vulnerability and the elderly. a total survey error perspective on the vlv survey. In *Surveying human vulnerabilities across the life course*, pages 27–64. Springer.

Oris, M. and Ritschard, G. (2014). Sequence analysis and transition to adulthood: an exploration of the access to reproduction in nineteenth-century East Belgium. In Blanchard, P., Bühlmann, F., and Gauthier, J.-A., editors, *Advances in sequence analysis: Theory, method, applications*, volume 2 of *Life Course Research and Social Policies*, pages 151–167. Springer.

Park, T. and Davis, C. S. (1993). A test of the missing data mechanism for repeated categorical data. *Biometrics*, 49(2):631–638.

Pasteels, I. and Mortelmans, D. (2015). Dyadic analysis of repartnering after divorce: Do children matter? *Journal of Family Research*, 10:143–164.

Pelletier, D., Assche, B.-V., Simard-Gendron, A., et al. (2020). Measuring life course complexity with dynamic sequence analysis. *Social Indicators Research*, 152(3):1127–1151.

Perkins, N. J., Cole, S. R., Harel, O., Tchetgen Tchetgen, E. J., Sun, B., Mitchell, E. M., and Schisterman, E. F. (2018). Principled approaches to missing data in epidemiologic studies. *American journal of epidemiology*, 187(3):568–575.

Piccarreta, R. (2017). Joint sequence analysis: association and clustering. *Sociological Methods & Research*, 46(2):252–287.

Piccarreta, R. and 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.

Piccarreta, R. and Studer, M. (2019). Holistic analysis of the life course: Methodological challenges and new perspectives. *Advances in Life Course Research*, 41:100251.

Pigott, T. D. (2001). A review of methods for missing data. *Educational research and evaluation*, 7(4):353–383.

Pollock, G. (2007). Holistic trajectories: a study of combined employment, housing and family careers by using multiple-sequence analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(1):167–183.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raab, M., Fasang, A. E., and Hess, M. (2018). Pathways to death: the co-occurrence of physical and mental health in the last years of life. *Demographic Research*, 38:1619–1634.

Raab, M., Fasang, A. E., Karhula, A., and Erola, J. (2014). Sibling similarity in family formation. *Demography*, 51(6):2127–2154.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850.

Rissanen, J. (1983). A universal data compression system. *IEEE Transactions on information theory*, 29(5):656–664.

Ritschard, G. (2021). Measuring the nature of individual sequences. *Sociological Methods & Research*, page 00491241211036156.

Ritschard, G., Bussi, M., and O'Reilly, J. (2018). An index of precarity for measuring early employment insecurity. In Ritschard, G. and Studer, M., editors, *Sequence Analysis and Related Approaches*, volume 10 of *Life Course Research and Social Policies*, pages 279–295. Springer.

Ritschard, G. and Studer, M., editors (2018a). *Sequence analysis and related approaches : Innovative methods and applications*, volume 10 of *Life Course Research and social Policies*. Springer, Cham, Switzerland.

Ritschard, G. and Studer, M. (2018b). Sequence analysis: Where are we, where are we going? In Ritschard, G. and Studer, M., editors, *Sequence Analysis and Related Approaches*, volume 10 of *Life Course Research and Social Policies*, pages 1–11. Springer.

Roberts, C., Herzing, J. M., Sobrino Piazza, J., Abbet, P., and Gatica-Perez, D. (2022). Data privacy concerns as a source of resistance to complete mobile data collection tasks via a smartphone app. *Journal of Survey Statistics and Methodology*, 10(3):518–548.

Robette, N., Bry, X., and Lelièvre, E. (2015). A "global interdependence" approach to multidimensional sequence analysis. *Sociological Methodology*, 45(1):1–44.

Ron, D., Singer, Y., and Tishby, N. (1996). The power of amnesia: learning probabilistic automata with variable memory length. *Machine learning*, 25(2):117–149.

Rossignon, F., Studer, M., Gauthier, J.-A., and Le Goff, J.-M. (2018). Sequence history analysis (sha): Estimating the effect of past trajectories on an upcoming event. In Ritschard, G. and Studer, M., editors, *Sequence Analysis and Related Approaches*, volume 10 of *Life Course Research and Social Policies*, pages 83–100. Springer.

Rothenbühler, M. and Voorpostel, M. (2016). Attrition in the swiss household panel: Are vulnerable groups more affected than others? In Oris, M., Roberts, C., Joye, D., and Ernst Stähli, M., editors, *Surveying Human Vulnerabilities across the Life Course*, pages 223–244. Springer.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Roux, J., Grimaud, O., and Leray, E. (2018). Multichannel sequence analysis: an innovative method to study patterns of care pathways. Application to multiple sclerosis based on French health insurance data. *Revue d'Épidémiologie et de Santé Publique*, 66:S430–S431.

Rouzinov, S. and Berchtold, A. (2022). Regression-based approach to test missing data mechanisms. *Data*, 7(2):16.

Royston, P. (2004). Multiple imputation of missing values. *The Stata Journal*, 4(3):227–241.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rubin, D. B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association*, 72(359):538–543.

Rubin, D. B. (1978). Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse. In *Proceedings of the survey research methods section of the American Statistical Association*, volume 1, pages 20–34. American Statistical Association.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley Series in Probability and Statistics. John Wiley & Sons.

Santos, J. M. and Embrechts, M. (2009). On the use of the adjusted rand index as a metric for evaluating supervised classification. In *Artificial Neural Networks–ICANN 2009: 19th International Conference, Limassol, Cyprus, September 14-17, 2009, Proceedings, Part II 19*, pages 175–184. Springer.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. CRC press.

Schafer, J. L. and Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research*, 33(4):545–571. PMID: 26753828.

Scherer, S. (2001). Early career patterns: a comparison of Great Britain and West Germany. *European sociological review*, 17(2):119–144.

Schoon, I. and Lyons-Amos, M. (2016). Diverse pathways in becoming an adult: the role of structure, agency and context. *Research in Social Stratification and Mobility*, 46:11–20.

Schwanitz, K. (2017). The transition to adulthood and pathways out of the parental home: a cross-national analysis. *Advances in Life Course Research*, 32:21–34.

Seaman, S. R., White, I. R., Copas, A. J., and Li, L. (2012). Combining multiple imputation and inverse-probability weighting. *Biometrics*, 68(1):129–137.

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., and Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using mice: a caliber study. *American journal of epidemiology*, 179(6):764–774.

Sirniö, O., Kauppinen, T. M., and Martikainen, P. (2017). Intergenerational determinants of joint labor market and family formation pathways in early adulthood. *Advances in Life Course Research*, 34:10–21.

Spallek, M., Haynes, M., and Jones, A. (2014). Holistic housing pathways for Australian families through the childbearing years. *Longitudinal and Life Course Studies*, 5(2):205–226.

Spiess, M., Kleinke, K., and Reinecke, J. (2021). Proper multiple imputation of clustered or panel data. *Advances in Longitudinal Survey Methodology*, pages 424–446.

Spini, D., Morselli, D., Elcheroth, G., Gauthier, J.-A., Le Goff, J.-M., Dasoki, N., Tillmann, R., and Rossignon, F. (2019). The LIVES-FORS cohort survey: a longitudinal diversified sample of young adults who have grown up in Switzerland. *Longitudinal and Life Course Studies*, 10(3):399–410.

SPSS, I. (2017). Ibm spss 25.0 for windows [computer software]. *Chicago, IL: SPSS*.

Stoltzfus, J. C. (2011). Logistic regression: a brief primer. *Academic emergency medicine*, 18(10):1099–1104.

Struffolino, E., Bernardi, L., and Larenza, O. (2020). Lone mothers' employment trajectories: a longitudinal mixed-method study. *Comparative Population Studies*, 45.

Studer, M. (2013). WeightedCluster Library Manual: A practical guide to creating typologies of trajectories in the social sciences with R. *LIVES Working papers*, 24. Swiss National Centre of Competence in Research LIVES, Geneva.

Studer, M. (2015). Comment: on the use of globally interdependent multiple sequence analysis. *Sociological Methodology*, 45(1):81–88.

Studer, M. (2018). Divisive property-based and fuzzy clustering for sequence analysis. In Ritschard, G. and Studer, M., editors, *Sequence Analysis and Related Approaches*, volume 10 of *Life Course Research and Social Policies*, pages 223–239. Springer.

Studer, M. (2019). Validating sequence analysis typologies using bootstrapping. *LIVES Working papers*, 80. Swiss National Centre of Competence in Research LIVES, Geneva.

Studer, M., Liefbroer, A. C., and Mooyaart, J. E. (2018a). Understanding trends in family formation trajectories: an application of Competing Trajectories Analysis (CTA). *Advances in Life Course Research*, 36:1–12.

Studer, M. and Ritschard, G. (2016). What matters in differences between life trajectories: A comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 179(2):481–511.

Studer, M., Ritschard, G., Gabadinho, A., and Müller, N. S. (2011). Discrepancy analysis of state sequences. *Sociological Methods & Research*, 40(3):471–510.

Studer, M., Struffolino, E., and Fasang, A. E. (2018b). Estimating the relationship between time-varying covariates and trajectories: the sequence analysis multistate model procedure. *Sociological Methodology*, 48(1):103–135.

Sturgis, P., Allum, N., and Brunton-Smith, I. (2009). Attitudes over time: The psychology of panel conditioning. In Lynn, P., editor, *Methodology of longitudinal surveys*, pages 113–126. John Wiley & Sons.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, 82(398):528–540.

Thurkow, N. M., Bailey, J. S., and Stamper, M. R. (2000). The effects of group and individual monetary incentives on productivity of telephone interviewers. *Journal of Organizational Behavior Management*, 20(2):3–25.

Tillmann, R., Voorpostel, M., Kuhn, U., Lebert, F., Ryser, V.-A., Lipps, O., Wernli, B., and Antal, E. (2016). The Swiss household panel study: observing social change since 1999. *Longitudinal and Life Course Studies*, 7(1):64–78.

Tillé, Y. and Wilhelm, M. (2017). Probability sampling designs: principles for choice of design and balancing. *Statistical Science*, pages 176–189.

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16(3):219–242. PMID: 17621469.

Van Buuren, S. (2015). Fully conditional specification. In Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A., and Verbeke, G., editors, *Handbook of missing data methodology*, pages 267–294. CRC Press.

Van Buuren, S. (2018). *Flexible Imputation of Missing Data*. Chapman & Hall/CRC Interdisciplinary Statistics. CRC Press.

Van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C. G., and Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of statistical computation and simulation*, 76(12):1049–1064.

Van Buuren, S. and Groothuis-Oudshoorn, K. (1999). *Flexible multivariate imputation by MICE*. TNO.

Van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, 45:1–67.

Vehovar, V., Toepoel, V., and Steinmetz, S. (2016). Non-probability sampling. In Wolf, C., Fu, Y.-c., Smith, T., and Joye, D., editors, *The Sage handbook of survey methodology*, volume 1, pages 327–343. SAGE.

Von Hippel, P. T. (2013). Should a normal imputation model be modified to impute skewed variables? *Sociological Methods & Research*, 42(1):105–138.

Von Hippel, P. T. (2020). How many imputations do you need? A two-stage calculation using a quadratic rule. *Sociological Methods & Research*, 49(3):699–718.

Voorpostel, M. (2010). Attrition patterns in the Swiss household panel by demographic characteristics and social involvement. *Swiss Journal of Sociology*, 36(2):359–377.

Voorpostel, M., Lipps, O., and Roberts, C. (2021). Mixing modes in household panel surveys: recent developments and new findings. *Advances in longitudinal survey methodology*, pages 204–226.

Voorpostel, M., Tillmann, R., Lebert, F., Kuhn, U., Lipps, O., Ryser, V.-A., and Wernli, B. (2016). Swiss household panel user guide (1999–2015). FORS, Lausanne.

Wahrendorf, M., Marr, A., Antoni, M., Pesch, B., Jöckel, K.-H., Lunau, T., Moebus, S., Arendt, M., Brüning, T., Behrens, T., et al. (2018). Agreement of self-reported and administrative data on employment histories in a German cohort study: A sequence analysis. *European Journal of Population*, 35(2):329–346.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

Warrens, M. J. and van der Hoef, H. (2022). Understanding the adjusted rand index and other partition comparison indices based on counting object pairs. *Journal of Classification*, 39(3):487–509.

White, I. R., Royston, P., and Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine*, 30(4):377–399.

Widmer, E. D. and Ritschard, G. (2009). The de-standardization of the life course: are men and women equal? *Advances in Life Course Research*, 14(1-2):28–39.

Wu, H. and Leung, S.-O. (2017). Can likert scales be treated as interval scales?—a simulation study. *Journal of Social Service Research*, 43(4):527–532.

Wu, W., Jia, F., and Enders, C. (2015). A comparison of imputation strategies for ordinal missing data on likert scale variables. *Multivariate behavioral research*, 50(5):484–503.

Yeh, H.-W., Chan, W., and Symanski, E. (2012). Intermittent missing observations in discrete-time hidden Markov models. *Communications in Statistics-Simulation and Computation*, 41(2):167–181.

Yeh, H.-W., Chan, W., Symanski, E., and Davis, B. R. (2010). Estimating transition probabilities for ignorable intermittent missing data in a discrete-time Markov chain. *Communications in Statistics—Simulation and Computation®*, 39(2):433–448.

Young, R. and Johnson, D. R. (2015). Handling missing values in longitudinal panel data with multiple imputation. *Journal of Marriage and Family*, 77(1):277–294.

Yuan, Y. C. (2010). Multiple imputation for missing data: Concepts and new development (version 9.0). *SAS Institute Inc, Rockville, MD*, 49(1-11):12.

# Appendix A

# Technical elements

This appendix contains statistical details related to Chapter 2 that focuses on the state of the art. In particular, we develop the concepts of bias, mechanisms of missing data, likelihood and Bayesian methods, and multiple imputation.

To formally define the concepts, we consider longitudinal data, which is the main focus of this thesis. We denote by $Y_{itj}$, the $j^{th}$ measurement of the $i^{th}$ individual at time $t$, with $i = 1, \ldots, n$, $j = 1, \ldots, m$ and $t = 1, \ldots, T$. Due to missingness, some values are potentially not observed. Therefore, we define the missing matrix $M$ as $M_{itj} = 1$ if $Y_{itj}$ is missing and $M_{itj} = 0$ if $Y_{itj}$ is observed. One can split the complete data $Y$ into observed $Y^{obs}$ and missing $Y^{mis}$ parts. For each individual $i$, $(Y_i, M_i)$ is called the "full data", with its density denoted as $f(Y_i, M_i|\theta, \gamma)$, where $\theta$ is the vector of parameters of the model for $Y_i$ and $\gamma$ is the vector of parameters modelling the missing matrix $M_i$. To simplify things, we suppose that the interest of the statistical analysis is the parameter $\theta$.

### Bias

If we denote by $\hat{\theta}$ the estimator of the parameter $\theta$, computed on the sample, the bias is defined as

$$\text{Bias}[\hat{\theta}] = \mathbb{E}[\hat{\theta}] - \theta.$$

Concretely, the bias is the difference between the parameter of interest and the average value of the estimator computed on all possible samples drawn from the population (Van Buuren, 2018).

### Mechanisms

The missing data are

- MCAR if $f(M_i|Y_i^{obs}, Y_i^{mis}) = f(M_i)$,

- MAR if $f(M_i|Y_i^{obs}, Y_i^{mis}) = f(M_i|Y_i^{obs})$,

- MNAR if $f(M_i|Y_i^{obs}, Y_i^{mis})$ depends at least on $Y_i^{mis}$.

When the data are MCAR, or MAR with disjoint parameter spaces for $\theta$ and $\gamma$, the mechanism is referred to as "ignorable". With Bayesian methods, the additional property that the prior distribution $f(\theta, \gamma)$ is factorisable as $f(\theta)f(\gamma)$ is needed (Little and Rubin, 2019).

### Likelihood and Bayesian methods

Likelihood methods are based on the observed likelihood

$$L(\theta, \gamma|Y^{obs}, M) = c\prod_{i=1}^{n}\int f(Y_i, M_i|\theta, \gamma)dY_i^{mis}. \tag{A.1}$$

In the case of ignorable data, the observed likelihood is equal to

$$L(\theta, \gamma|Y^{obs}, M) = \prod_{i=1}^{n} f(M_i|Y_i^{obs}, \gamma)f(Y_i^{obs}|\theta). \tag{A.2}$$

Therefore, the estimation of $\theta$ is independent of the missing distribution, and only the distribution of the complete data needs to be specified.

With Bayesian methods, distributions are assumed for the parameters. These methods are based on Bayes's theorem, which states that

$$f(\theta, \gamma|Y^{obs}, M) = \frac{f(Y^{obs}, M|\theta, \gamma)f(\theta, \gamma)}{f(Y^{obs}, M)}. \tag{A.3}$$

When the hypothesis of ignorability for the Bayesian inference is satisfied, we have that

$$f(\theta, \gamma) \propto \prod_{i=1}^{n} f(Y_i^{obs}|\theta)f(\theta)f(M_i|Y_i^{obs}, \gamma)f(\gamma). \tag{A.4}$$

Therefore, as for likelihood methods, the estimation of $\theta$ is independent of the missing distribution.

Selection and pattern-mixture models are the main strategies for non-ignorable missing data. With selection models, the joint distribution of $Y$ and $M$ is modeled as

$$f(M, Y|\theta, \gamma) = \prod_{i=1}^{n} f(Y_i|\theta)f(M_i|Y_i, \theta), \tag{A.5}$$

while with pattern-mixture models, it is set as

$$f(M, Y|\eta, \psi) = \prod_{i=1}^{n} f(Y_i|M_i, \eta) f(M_i|\psi). \tag{A.6}$$

**Multiple imputation**

There is a close connection between the Bayesian framework and multiple imputation. The whole idea of multiple imputations (Rubin, 1978) derives from the equation

$$f(\theta|Y_{obs}) = \int f(\theta|Y_{mis}, Y_{obs}) f(Y_{mis}|Y_{obs}) dY_{mis}. \tag{A.7}$$

By drawing several values for $Y_{mis}$ from the conditional distribution $f(Y_{mis}|Y_{obs})$, we can approximate this equation by

$$f(\theta|Y_{obs}) \approx \frac{1}{M} \sum_{i=1}^{M} f(\theta|Y_{mis}^{(m)}, Y_{obs}), \tag{A.8}$$

where $Y_{mis}^{(m)}$ is a draw for the missing values. From this approximation, Rubin (1987) derived rules to combine the results from several completed datasets when a normally distributed parameter $\theta$ is the main focus. If we denote by $(\hat{\theta}_i, V_i)$, respectively, the parameter and its variance estimates computed on the $i^{th}$ completed dataset, the pooled estimate $\hat{\theta}$ is set as the mean value

$$\hat{\theta} = \frac{1}{M} \sum_{i=1}^{M} \hat{\theta}_i.$$

The pooled variance $T$ is defined as

$$T = \hat{V} + (1 + 1/M)B,$$

where

$$\hat{V} = \frac{1}{M} \sum_{i=1}^{M} \hat{V}_i$$

is the estimated within imputation variance and

$$B = \frac{1}{M-1} \sum_{i=1}^{M} (\hat{\theta}_i - \hat{\theta})^2$$

is the estimated between imputation variance. In the pooled variance $T$, $B$ is inflated by a factor $1/M$ to consider that the number of imputations is finite.

# Appendix B

# Statistical models

## B.1 Multinomial model

### Multinomial model

The multinomial model, which is a type of generalised linear model (McCullagh and Nelder, 2019), is a standard way to model a categorical output with more than two categories. For a categorical random variable $Y$ with $K$ categories and a predictors $X$, the multinomial model is expressed as:

$$P(Y = k|X = x) = \frac{e^{\beta_{0k}+x^t\beta_k}}{1+\sum_{l=1}^{K-1} e^{\beta_{0l}+x^t\beta_l}}, \text{ for } k = 1, \ldots, K-1$$
$$P(Y = K|X = x) = \frac{1}{1+\sum_{l=1}^{K-1} e^{\beta_{0l}+z^t\beta_l}}.$$

For a sample $\{x_i, y_i\}_{i=1}^n$, the parameters $\beta_k, k = 1, \ldots, K-1$ are estimated by maximising the conditional log-likelihood

$$l(\beta) = \sum_{i=1}^{n} \log P(Y = y_i|X = x_i),$$

where $\beta = (\beta_1, \ldots, \beta_{k-1})^t$. In the *mlogit* package (Croissant, 2020), the Newton-Raphson algorithm is used to estimate the parameters. Starting from an initial guess $\beta^{(0)}$ for the parameter, it estimates at each step

$$\beta^{(t+1)} = \beta^{(t)} - \left\{\frac{\partial^2 l}{\partial\beta\partial\beta}(\beta^{(t)})\right\}^{-1} \frac{\partial l}{\partial\beta}(\beta^{(t)})$$

until a convergence is reached. However, the inversion of the Hessian matrix, namely the second derivatives, can be quite time-consuming and is even not possible in some cases.

**Neural networks**

It is composed of hidden units $Z_1, \cdots, Z_m$, which are defined as (Friedman et al., 2001)

$$Z_i := \sigma(X) = \frac{1}{1 + e^{-(\alpha_{0i} + \alpha_i X)}}. \tag{B.1}$$

Then, in a similar way to the multinomial model, the output is modelled as

$$P(Y = k|Z = z) := g_k(z) = \frac{e^{\beta_{0k} + z^t \beta_k}}{\sum_{l=1}^{K} e^{\beta_{0l} + z^t \beta_l}}, \text{ for } k = 1, \ldots, K. \tag{B.2}$$

To estimate the parameters $\alpha$'s and $\beta$'s, called weights, the deviance

$$-\sum_{i=1}^{n} \sum_{k=1}^{K} y_{ik} \log f_k(x_i),$$

where $y_{ik} = 1$ if $y_i = k$ and otherwise and $f_k(x) = g_k(\sigma(x))$, is minimised. An iterative algorithm is used to solve this minimization problem.

This model is applied to fit multinomial models in the *nnet* package. Its application instead of the *mlogit* function, allowed us to fasten the computation in the *seqimpute* package that implements the MICT algorithm.

## B.2 Random forest

The random forest model builds a predetermined number of trees on bootstrap samples (Breiman, 2001). A bootstrap sample is generated by randomly drawing with replacement $n$ observations, where $n$ is the size of the sample. A tree is a partition of the values of the predictor space into non-overlapping parts $P_1, \ldots, P_l$, obtained with a recursive binary splitting. The predictor space is first split into two parts, also called regions, $P_1(j, t) = \{x|x_j \leq t\}$ and $P_2(j, t) = \{x|x_j > t\}$, which are split again into two subparts, and so on. The process usually ends when every observation that belongs to a given region has the same predicted class. Every time we must determine a split, a random sample of predictors is drawn among all of them. This random draw reduces the correlation between the trees (Friedman et al., 2001). The Gini index

$$G = \sum_{i=1}^{l} \sum_{k=0}^{1} p_{ik}(1 - p_{ik}),$$

where $p_{ik}$ is the proportion of the observations which have predicted class $k$ in the region $i$, is used to determine the feature to split and the division point by minimizing For a new observation, each tree is applied and estimated probabilities are obtained by the proportion of

trees among the forest that predict each class.

## B.3 Variable-length Markov chains

Variable-length Markov chains are a type of Markov chains. Markov chains are used to model longitudinal categorical data (Gabadinho and Ritschard, 2016). We denote by $x := x_1 x_2 \cdots x_l$, a sequence of length $l$, which is the realisation of $l$ random variables $X_1, \ldots, X_l$. Markov chains are

Concretely, for each context $c_1, \ldots, c_k$, there exist a $L$, $0 \leq L \leq k$, with $P(\sigma|c_1, \ldots, c_k) \approx P(\sigma|c_{k-L}, \ldots, c_k)$. Therefore, going from contexts of maximum length, the algorithm determine, for each context $c = c_1, \ldots, c_k$, if $\widehat{P}(\sigma|c)$ can be approximated by $\widehat{P}(\sigma|suf(c))$, where $suf(c) = c_2, \ldots, c_k$. The estimated probabilities $\widehat{P}(\sigma|c)$ are the share of time the state $\sigma$ follows the subsequence $c$ any time that $c$ is observed in the dataset. Two gain functions are mainly used to determine if $\widehat{P}(\sigma|c)$ can be approximated by $\widehat{P}(\sigma|suf(c))$. The first one is based on the ratio $\frac{\widehat{P}(\sigma|c)}{\widehat{P}(\sigma|suf(c))}$ and a cutoff value $C$ provided by the user

$$G_1(c) = \sum_{\sigma \in A} \mathbb{1}\left[\frac{\hat{P}(\sigma|c)}{\hat{P}(\sigma|suf(c))} \geq C \text{ or } \frac{\hat{P}(\sigma|c)}{\hat{P}(\sigma|suf(c))} \leq \frac{1}{C}\right] \geq 1 \tag{B.3}$$

Therefore, if for every element $\sigma$ of the alphabet, the ratio $\frac{\widehat{P}(\sigma|c)}{\widehat{P}(\sigma|suf(c))}$ is close to 1, $\widehat{P}(\sigma|c)$ will be approximated by $\widehat{P}(\sigma|suf(c))$. The second gain function is defined as

$$G_2(c) = N(c) \sum_{\sigma \in A} \widehat{P}(\sigma|c) \log\left(\frac{\widehat{P}(\sigma|c)}{\widehat{P}(\sigma|suf(c))}\right) > C, \tag{B.4}$$

where $N(c)$ is the number of times $c$ appears in the dataset. The threshold $C$ is defined as

$$C_\alpha = \frac{1}{2} qschiq(1 - \alpha, df), \tag{B.5}$$

where *qschiq* is the quantile function of a $\chi^2$ variable and *df* is set to the number of states in the alphabet minus one.

# Appendix C

# Men results

Concerning men, as hypothesised, there were no clear (linear) links between professional and family channels. The cohabitational status and child channels were the most interrelated ones, according to Cronbach's (0.5). Unlike with women, the pair composed of the health issues and cohabitational status channels as well as the pair of the professional status and cohabitational status channels produced Cronbach's values larger than 0.1 (0.17 and 0.11, respectively). Although these values were still small, we chose to investigate these combinations of channels, as interpreting raw Cronbach's values to evaluate joint channels can be unclear. Moreover, this provided information on how the two approaches (MSA and EA) behave when channels are weakly linked.

For the pair of health status and cohabitational channels, no clustering under 10 groups takes the association between channels into account since both ASWw and HC are in the 95% intervals of the values generated by the null model. Therefore, it is not sensible to cluster together the health status and cohabitational channels. In a similar way, for the pair of professional and cohabitational status, no clustering built with MSA, regardless of the dissimilarity measure, is a joint typology. On the other hand, for EA, the five groups clustering built with optimal matching where substitution costs were based on transition costs, is significant for both cluster quality indices according to the permutation test. However, this cluster is not satisfactory. It consists of a large group that contains more than 70% of the sequences and four small clusters (Figure C.1). Moreover, two clusters are heterogeneous (ASWw by groups close to 0). This highlights two points. First, a clustering may be satisfactory for some criteria but not others. Then, EA may be more prone to create artificial joint clusterings than MSA.

The derived typologies for the cohabitational and child channels were slightly different from those extracted from the women sequences. The two-cluster solutions, which separated the dataset according to whether an individual had a child or not, still gave the best ASWw and HC values in most cases. The classification in seven groups built with standard optimal matching was the most sensible clustering when MSA is applied (Figure C.2). One has to be

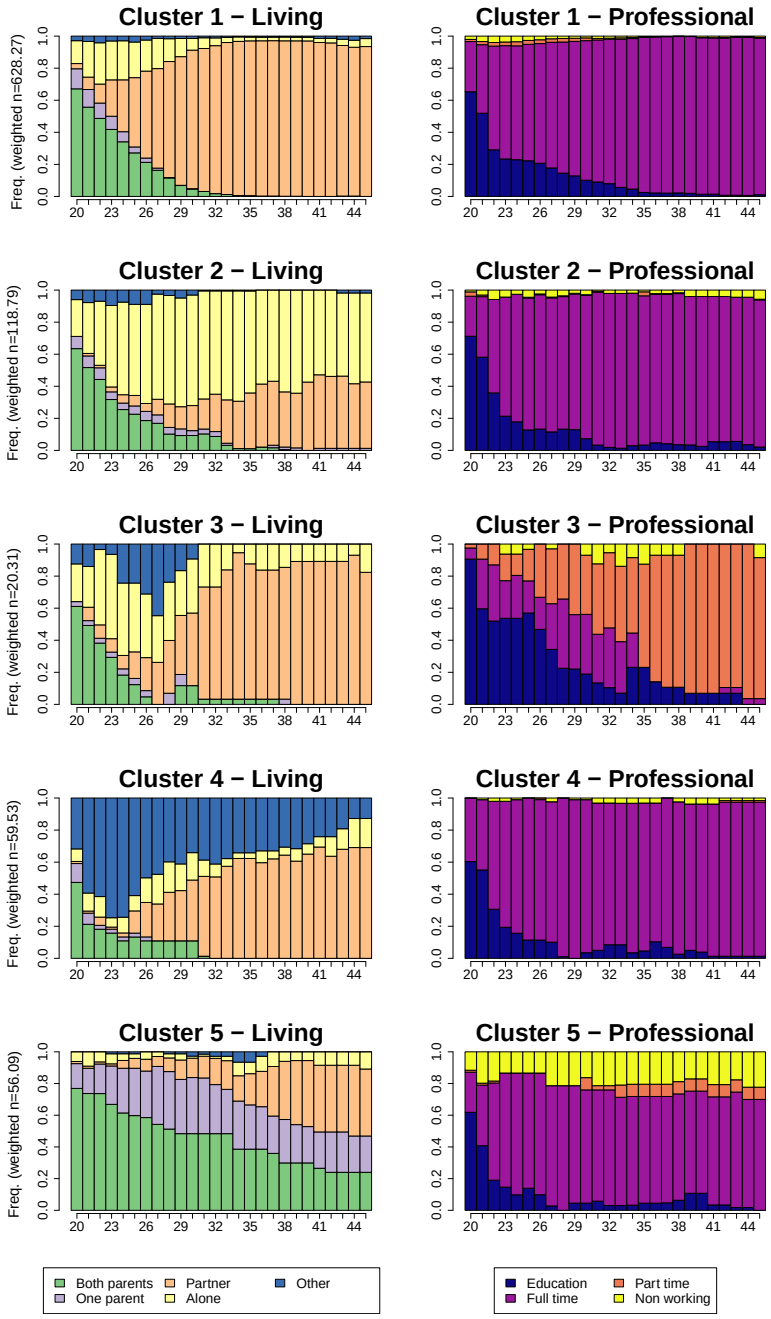



Figure C.1: Men subset: Chronograms of the five-group typology of professional and cohabitational status channels obtained with MSA with substitution costs based on transition rates.

careful because some clusters are not well separated. Concerning EA, two clusterings stand out: the eight groups clustering built with standard OM and the seven groups clustering built with Hamming distance. The first one is defined mainly by differences in duration (Figure C.3), and the second one by differences in timing (Figure C.4). Both clusterings have clusters that are not well separated and/or small. Therefore, one should not over-interpret these groups.

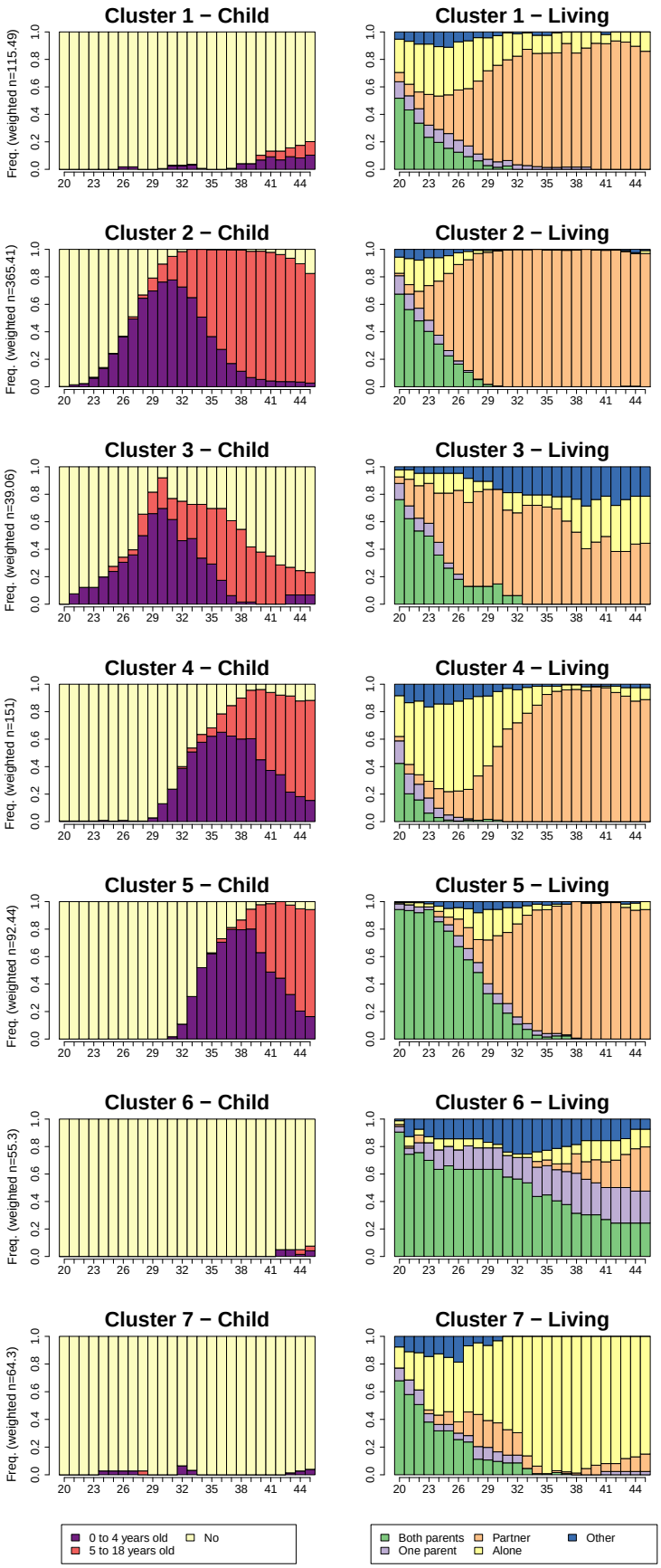



Figure C.2: Men subset: Chronograms of the seven-group typology of child and cohabitational status channels obtained with MSA with standard optimal matching.

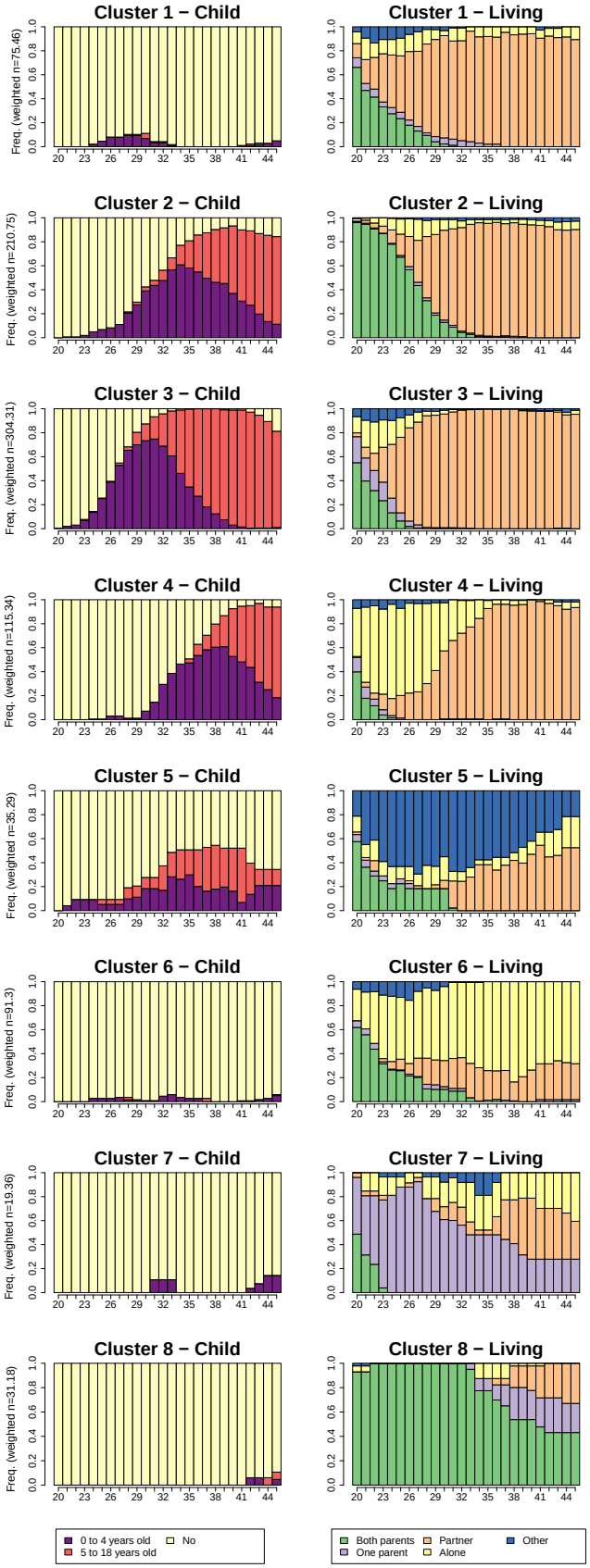



Figure C.3: Men subset: Chronograms of the eight-group typology of child and cohabitational status channels obtained with MSA with standard optimal matching.

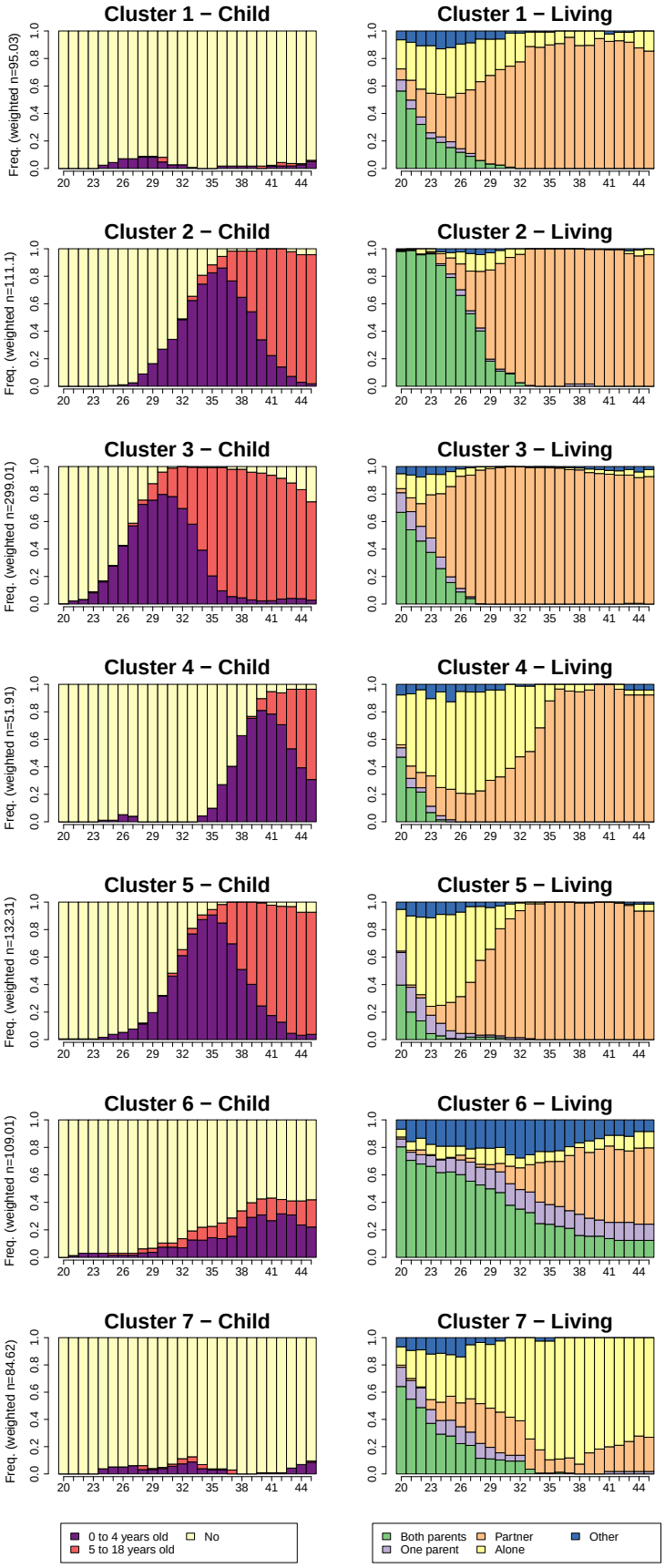



Figure C.4: Men subset: Chronograms of the seven-group typology of child and cohabitational status channels obtained with EA with Hamming distance.

# Appendix D

# Detail of the variables

| Variable | Description of the variable | Variable | Description of the variable |
|---|---|---|---|
| SEX14 | Sex | P14W31 | CMJ: Number of employees |
| AGE14 | Age in year of interview | P14W32 | CMJ: Private or Public employer |
| CIVSTA14 | Civil status in year of interview | P14W602 | CMJ: restructuring |
| OWNKID14 | Number of children born | P14W34A | CMJ: Position: Management, supervision |
| EDUCAT14 | Highest level of education achieved, grid + individual 11 codes | P14W36 | CMJ: Job limitation in time |
| EDCAT14 | Highest level of education achieved, grid + individual 17 codes | P14W37 | CMJ: Job limitation: Type of contract |
| EDU_1_14 | Highest level of education achieved, individual questionnaire 11 codes | P14W38 | CMJ: Job limitation: Duration of contract |
| EDUGR14 | Highest level of education achieved, grid 11 codes | P14W39 | CMJ: Part-time or Full-time |
| EDGR14 | Highest level of education achieved, grid 19 codes | P14W42 | CMJ: Percentage of part-time work |
| ISCED14 | International Standard Classification of Education ISCED 1997 | P14W43 | CMJ: Part-time work: Reason |
| EDYEAR14 | Years of Education based on ISCED Classification | P14W46 | CMJ: Work reference: Number of hours |
| OCCUPA14 | Actual occupation, from grid | P14W85 | CMJ: Company: Number of employees |
| NAT_1_14 | First nationality | P14W84 | CMJ: Commuting: Number of minutes daily |
| NAT_2_14 | Second nationality | P14W71A | CMJ: Type of working hours |
| P14E01 | First education completed after compulsory schooling | P14W216 | CMJ: Night work |
| P14E32 | First education completed after compulsory schooling: start: year | P14W218 | CMJ: Saturdays or Sundays work |
| P14E33 | First education completed after compulsory schooling: end: year | P14W603 | CMJ: Rhythm of work: intensity |
| P14E02 | Second education completed after compulsory schooling | P14W604 | CMJ: Work conditions: stress |
| P14E03 | Education completed since last wave | P14W605 | CMJ: Work conditions: noise, dirtiness |
| P14E04 | Education completed since last wave: type | P14W606 | CMJ: Work conditions: tiring posture |
| P14E05 | Second education completed: Type | P14W607 | CMJ: using a computer |
| P14E34 | Second education completed: start: year | P14W74 | CMJ: N contractual hours per week |
| P14E35 | Second education completed: end: year | P14W77 | CMJ: Number of hours worked per week |
| P14E06 | Third education completed after compulsory schooling | P14W80 | CMJ: Work at home |
| P14E07 | Third education completed: Type | P14W86A | CMJ: Job security: Estimation |
| P14E36 | Third education completed: start: year | P14W87 | CMJ: Job with supervisory tasks |
| P14E37 | Third education completed: end: year | P14W90 | CMJ: Number of persons under supervision |
| P14E31 | Tertiary-level education: field | P14W91 | CMJ: Job with participation in decision making |
| P14E08 | Education: Doctorate obtained | P14W92 | CMJ: Satisfaction: Income |
| P14E14 | Education: Current training at a school | P14W93 | CMJ: Satisfaction: Work conditions |
| P14E15 | Education: Current training: Type | P14W94 | CMJ: Satisfaction: Work atmosphere |
| P14E16 | First language: Personal use | P14W229 | CMJ: Satisfaction: interest in tasks |
| P14E17 | Second language: Personal use | P14W230 | CMJ: Satisfaction: amount of work |
| P14E18 | Professional training courses: Last 12 months | P14W615 | CMJ: Satisfaction: hierarchical superiors |
| WSTAT14 | Working status | P14W616 | CMJ: Satisfaction: promotion |
| P14W01 | Employment: work with payment last week | P14W228 | CMJ: Satisfaction: Job in general |
| P14W02 | Employment: work without payment last week | P14W95 | CMJ: Comparison: work conditions |
| P14W03 | Employment: job although not work last week | P14W96 | CMJ: Comparison: work atmosphere |
| P14W04 | Seeking job: Last four weeks | P14W100 | CMJ: Qualifications for job |
| P14W05 | Taken steps to find a job: Last 4 weeks | P14W101 | CMJ: Risk of unemployment: Next 12 months |
| P14W613 | Unemployed: regional job centre | PAUG14R4 | PROFESSIONAL INTEGRATION |
| P14W06 | Job offered: Earliest starting date | P14F50 | Interference work <-> private activities/family obligations |
| P14W500 | Currently not working: ready to accept an interesting job | P14F51 | Exhausted after work to do what you would like |
| P14W11 | Chance of finding the right job: Next 12 months | P14F52 | How difficult to disconnect from work |
| P14W12 | Currently not working: First reason | IS1MAJ14 | ISCO classification: Main current job: 1-digit-position |
| P14W13 | Currently not working: Second reason | IS2MAJ14 | ISCO classification: Main current job: 2-digit-position |
| P14W14 | Currently not working: Third reason | IS3MAJ14 | ISCO classification: Main current job: 3-digit-position |
| P14W17 | Number of jobs or employers: Last week | IS4MAJ14 | ISCO classification: Main current job: 4-digit-position |
| P14W610 | Total number of hours worked in all jobs | CSPMAJ14 | Swiss socio-professional category: Main job |
| P14W18 | Change of job or employer: Last 12 months | GLDMAJ14 | Goldthorpes class schema: Main job |
| P14W21 | Change of employer: frequency | ESECMJ14 | European Socio-Economic Classification (ESeC) main current job |
| P14W23 | Change of job: month | TR1MAJ14 | Treiman prestige scale 1 for main job |
| P14W25 | Change of job: year | CAIMAJ14 | CAMSIS scale with ISCO for main job |
| P14W66 | Change of employer: month | WR3MAJ14 | Wrights class structure: Main job |
| P14W69 | Change of employer: year | NOGA2M14 | Current main job: Nomenclature of economic activities |
| P14W600 | Change of job or employer: first reason | P14W608 | First regular job: at what age |
| P14W601 | Change of job or employer: second reason | P14W609 | Number of years spent in paid work |
| P14W612 | Current main job: access | P14W154 | Non-active persons: Paid Job Last year: yes,no |
| P14W28 | CMJ: Current main job: Occupation | P14W177 | Active persons: Change in job: yes, no |
| P14W29 | CMJ: Type of employment | | |

Table D.1: Description of the variables of the LCS subsample.

# Appendix E

## E.1 States with an higher probability

| Trajectory | States |
|---|---|
| Professional | *education* |
| | *non-working* |
| Cohabitational (4 states) | with *child* |
| | *Other* |
| Cohabitational (8 states) | living *alone* |
| | with *child* |
| | with *one parent* |
| | with *partner* |
| Civil status | *separated* |
| | *divorced* |
| | *single*, never married |
| Health Satisfaction | *low* |
| | *average* |
| School-to-work transition | further education (*FE*) |
| | higher education (*HE*) |
| | *training* |

Table E.1: For each datasets, the states with an higher probability to trigger a missing value for the second and third process of missing data are detailed.

## E.2 Details on the data

The objective was to create datasets for life course research that did not have any missing data and possessed typical characteristics. We used six datasets, with five of them derived from the Swiss household panel (SHP) (Voorpostel et al., 2016) and one from a study by McVicar and Anyadike-Danes (2002).

Three datasets were based on the retrospective data collected through life-history calendars in the SHP. To compensate for attrition, a second refreshment sample was added to the SHP in 2013. During the first wave of data collection for this new sample, retrospective data were collected, and from these data, we created three datasets, which include two datasets capturing cohabitational status trajectories and one dataset capturing professional status trajectories.

#### Cohabitational status

The goal was to build two datasets of cohabitational status. The two datasets consist of the same trajectories, but coded differently. This was done to assess the coding detail on the quality of the imputations.

We built trajectories of cohabitational status between the ages of 15 and 40. Among the 6088 individuals in the sample, 3854 were older than 40 at the time of the data collection, and 3710 had no missing data for their cohabitational status.

Concretely, individuals were asked to mark potential cohabitational periods for a list of people, including father, mother, sister(s) or brother(s), half-brother(s) or half-sister(s), alone, partner or spouse, child or children, other people from kinship, friend(s) or housemate(s) and other. We regrouped all potential combinations into eight and four groups to create the two datasets. For the coding in eight states, individuals were put in the *Both parent* state if they were living with both their parents and potentially any other siblings, *One parent* state if they were living with one parent and potentially other kinship, *Alone* if they were living alone, *Partner* if they were living with their partner, but without children, *Child* if they were living with their child (but without a partner and without parent(s)), *Partner and Child* if they were living with a partner at a least one child, *Relatives* if they were living with any relative (other than parent(s), partner or child) and *Other* if they were living with other individuals other than relatives.

To obtain the coding in 4 states of the cohabitational status from the coding in 8 states, the states *Both parents* and *One parent* were fused into the state *Parent(s)*, the states *Child* and *Partner and Child* into *Child* and finally the states *Alone*, *Relatives* and *Other* into *Other*.

### Professional status

We built trajectories of retrospective professional status between the ages of 15 and 40. The life-history calendar assessed the professional activities but not the educational one, which was measured during the second wave of collection for the second refreshment sample.

Out of the 6088 individuals in the sample, 4932 completed the life-history calendar and responded to the second wave. Among them, 3473 were older than 40 at the time of the retrospective collection and 3382 did not have any missing data.

Information about the percentage worked was gathered with the life-history calendar. To simplify, we considered only two categories for the professional activities: full-time or part-time.

The civil status and satisfaction with health status datasets were built using prospective data collected through the Swiss household panel, which began in 1999 and has annual data collection. For both civil status and satisfaction with health status, only individuals with no missing data during the first 21 waves were selected.

### Satisfaction with health status

Out of the 7,799 individuals who participated in the first wave of the Swiss household panel, 1,264 answered all 21 waves of data collection. Of these, only five individuals had missing values for at least one wave of the satisfaction with health status variable, resulting in a final sample of 1,259 individuals.

The satisfaction with health status variable was measured on a scale ranging from 0 (not at all satisfied) to 10 (completely satisfied) in each wave. To create more balanced categories, we recoded this variable into four categories: low (0 to 4), average (5 and 6), high (7 and 8), and very high (9 and 10). This categorisation was chosen to ensure that there were enough observations in the low category, given that most people tend to report being quite satisfied with their health status.

### Civil status

The civil status dataset includes individuals who have information about their civil status in each of the 21 waves. This information is obtained not only from the questionnaire provided to each individual but also from the questionnaire filled by the household reference person. Therefore, the civil status dataset contains more observations (2324) than the satisfaction with the health status dataset (1259).

Concerning the coding, we fused the *registered partnership* with the *married* state and the *dissolved partnership* with the *separated* one.

### Mvad

The mvad datasets comes from a study by McVicar and Anyadike-Danes (2002) on transition from school to work. In addition to some covariates, it contains the monthly labour market activities of young individuals in a cohort survey, followed from July 1993 to June 1999. The data were collected through two interviews that collected retrospective data. 712 responded to both interviews and, hence, does not have any missing data.

The dataset is available in the *TramineR* (Gabadinho et al., 2011) package in *R*. We kept all 712 trajectories and kept the original coding.

## E.3 Formulae for the criteria

We detail the formulae applied for the computation of the three criteria. For an imputed dataset and the corresponding complete dataset on which missing data were simulated, the three criteria are computed the following way.

Let $A$ be the set of states that appear in the dataset, $T$ the total number of time points and $N$ the total number of sequences in the dataset.

The *timing* criteria is defined as

$$\sum_{t=1}^{T}\sum_{s\in A}\frac{|\widehat{n}_s^{(t)} - n_s^{(t)}|}{NT},$$

where $\widehat{n}_i^{(t)}$ is the number of sequences of the imputed dataset that are in state $s$ in time $t$ and $n_i^{(t)}$ is the number of sequences of the original dataset that are in state $s$ in time $t$.

The *duration* criteria is defined as

$$\sum_{s\in A}\frac{|\widehat{mls}_s - mls_s|}{NT},$$

where $\widehat{mls}_s$ is the mean length of the spells in state $s$ in the imputed dataset and $mls_s$ is the mean length of the spells in state $s$ in the original dataset.

The *sequencing* criteria is defined as

$$\sum_{s\in A}\frac{n_s^{DSS}}{n^{DSS}}\sum_{q\in A}|\widehat{p}_{sq}^{DSS} - p_{sq}^{DSS}|,$$

where the $\widehat{p}_{sq}^{DSS}$ is the probability to switch from state $s$ to state $q$ in the dataset of sequences of distinct successive states built from the imputed dataset and, $p_{sq}^{DSS}$, $n_s^{DSS}$, $n^{DSS}$ are, respectively, the probability to switch from state $s$ to state $q$, the total number of state $s$ and the total number of states in the dataset of sequences of distinct successive states built from the original dataset.

## E.4 Normalisation of the criteria

For a dataset, missing data generation model and criteria, let $C$ be the collection of all the values obtained across all imputation methods and parametrisations applied. Each value $c \in C$ is modified as

$$\frac{c - min(C)}{max(C) - min(C)}.$$

This new set of values is standardised (mean 0 and unit variance).

## E.5 Total score by algorithm

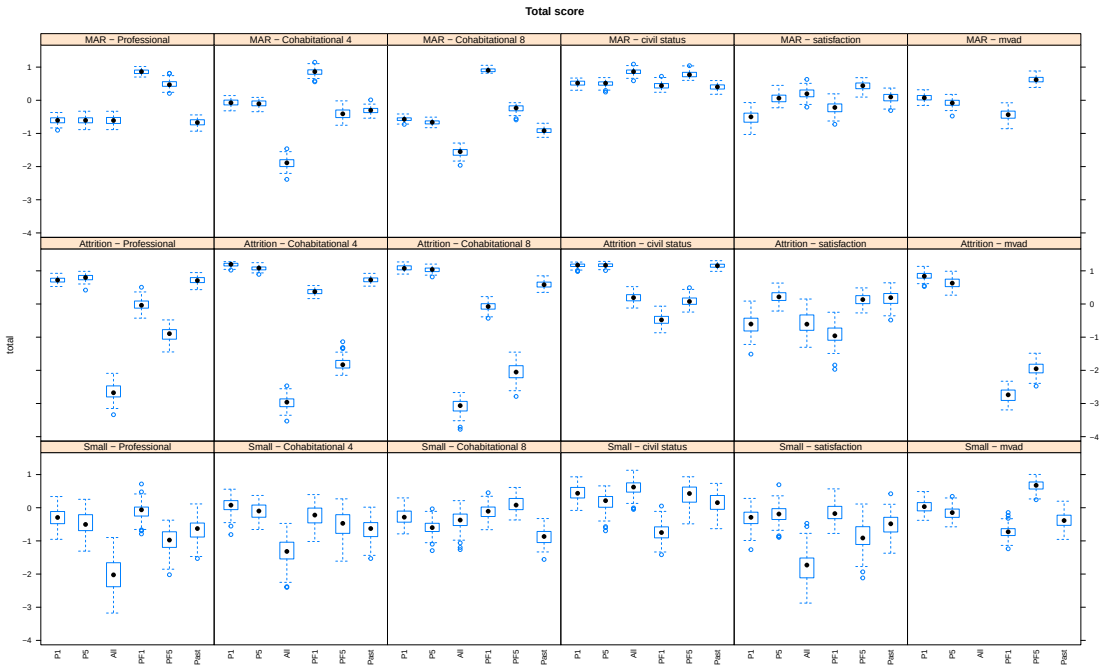



Figure E.1: FCS multinomial - Total score for each parametrisation on each of the 18 scenarios. The different parametrisations are: using one observation in the past (P1), five observations in the past (P5), all past observations (Past), one observation both in past and future (PF1), five observations both in past and future (PF5) and all observations (All)

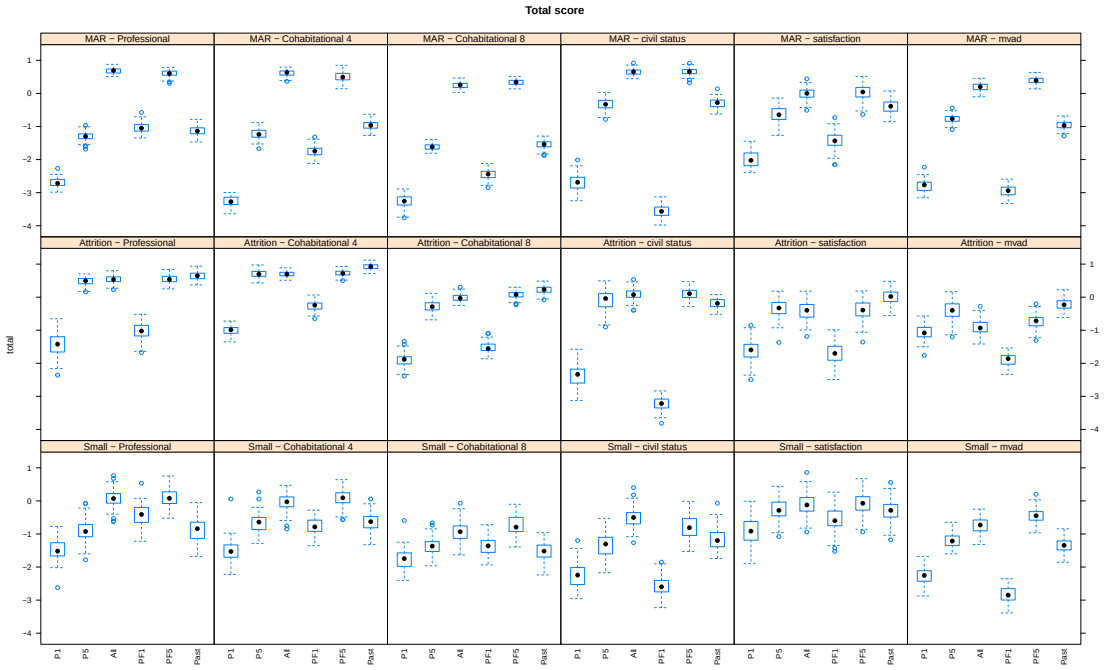



Figure E.2: FCS random forest - Total score for each parametrisation on each of the 18 scenarios. The different parametrisations are: using one observation in the past (P1), five observations in the past (P5), all past observations (Past), one observation both in past and future (PF1), five observations both in past and future (PF5) and all observations (All).

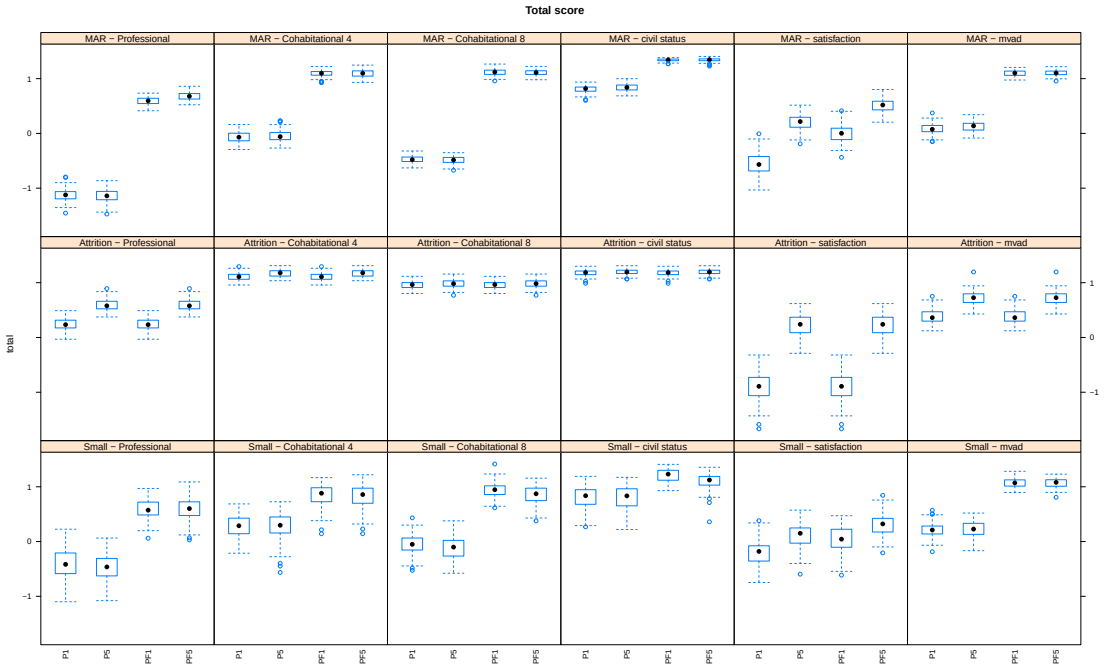



Figure E.3: MICT multinomial - Total score for each parametrisation on each of the 18 scenarios. The different parametrisations are: using one observation in the past (P1), five observations in the past (P5), one observation both in past and futur (PF1) and five observations both in past and futur (PF5).

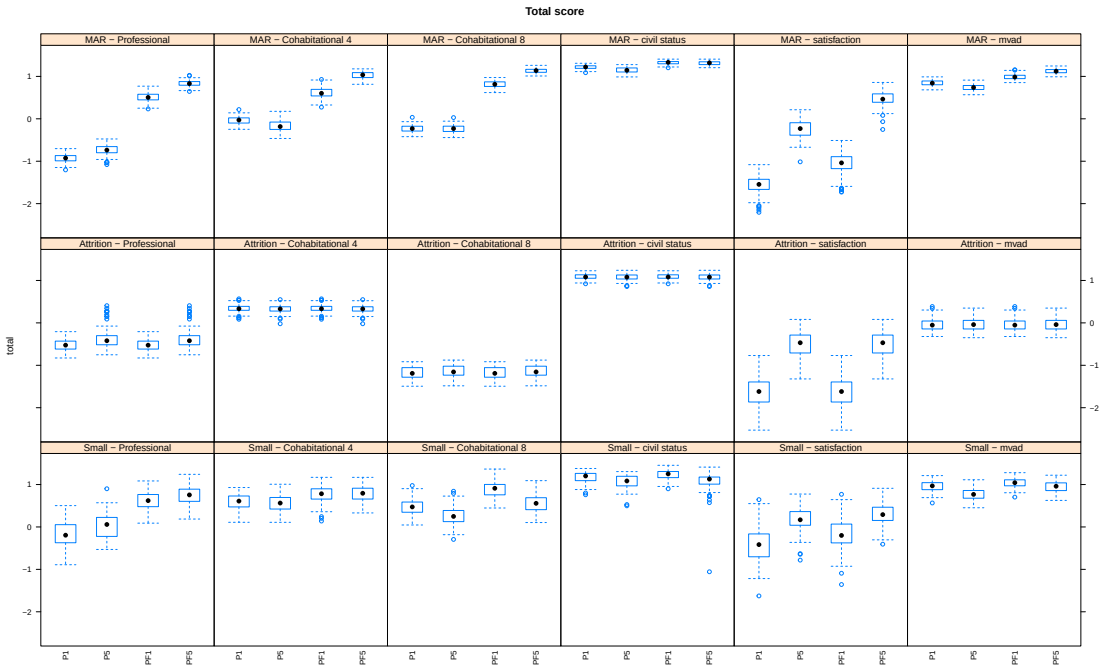



Figure E.4: MICT random forest - Total score for each parametrisation on each of the 18 scenarios. The different parametrisations are: using one observation in the past (P1), five observations in the past (P5), one observation both in the past and future (PF1), and five observations both in the past and future (PF5).

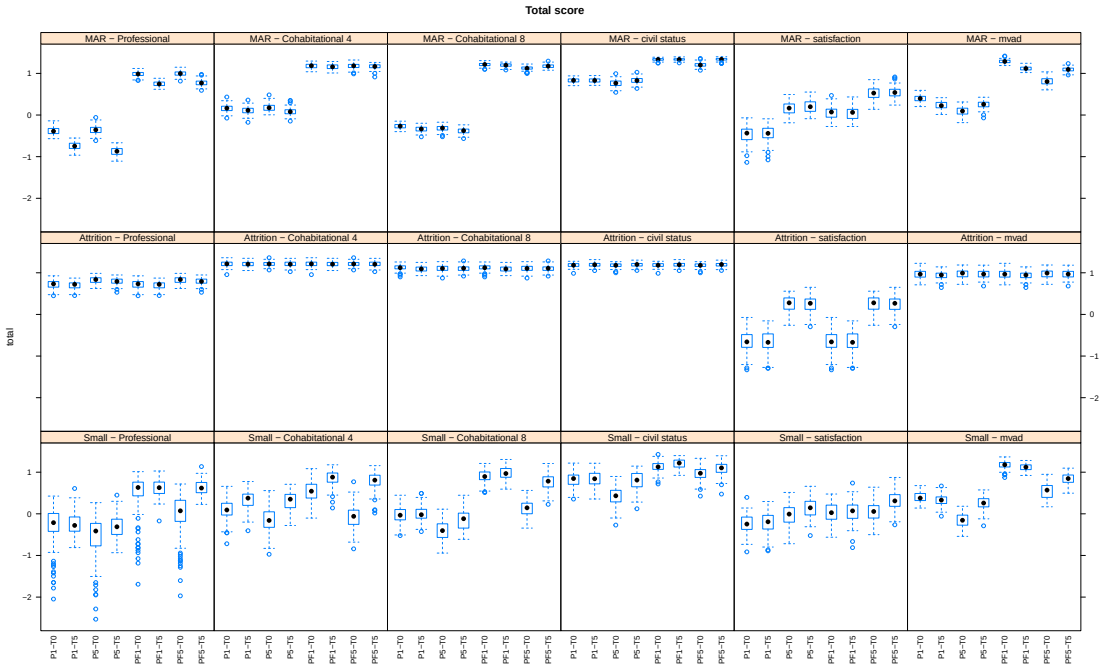



Figure E.5: MICT-timing multinomial - Total score for each parametrisation on each of the 18 scenarios. Concerning the parametrisations, time frames of width 0 and 5, denoted respectively by T0 and T5, are combined with four different choices of predictors: one observation in the past (P1), five observations in the past (P5), one observation in the past and future (PF1) and five observations in the past and future (PF5).

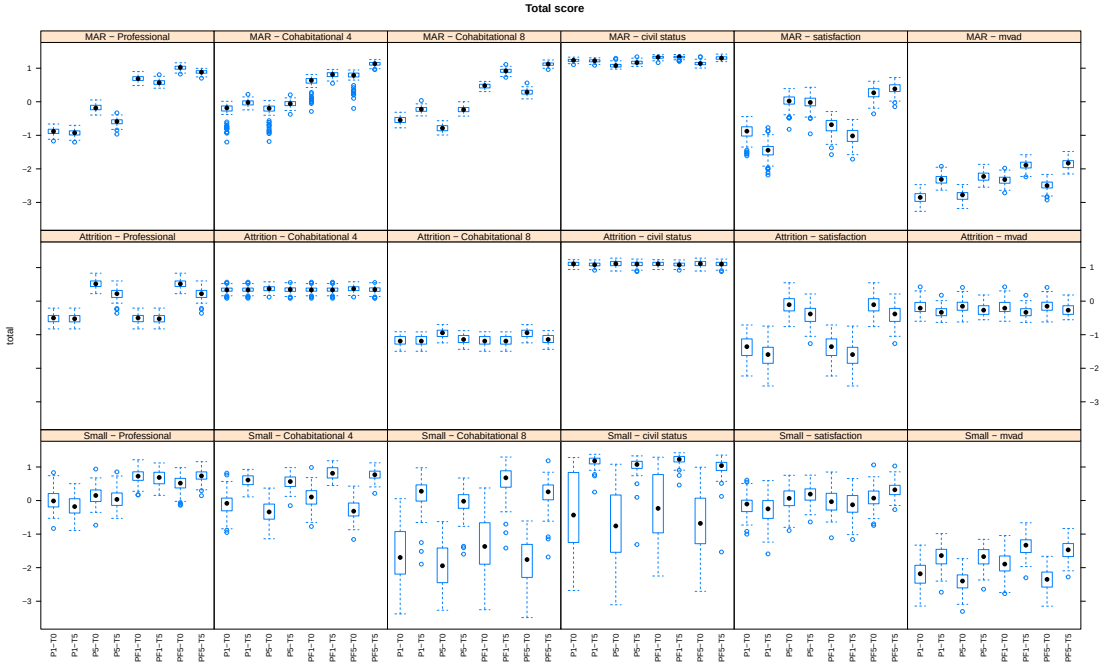



Figure E.6: MICT-timing random forest - Total score for each parametrisation on each of the 18 scenarios. Concerning the parametrisations, time frames of width 0 and 5, denoted respectively by T0 and T5, are combined with four different choices of predictors: one observation in the past (P1), five observations in the past (P5), one observation in the past and future (PF1) and five observations in the past and future (PF5).

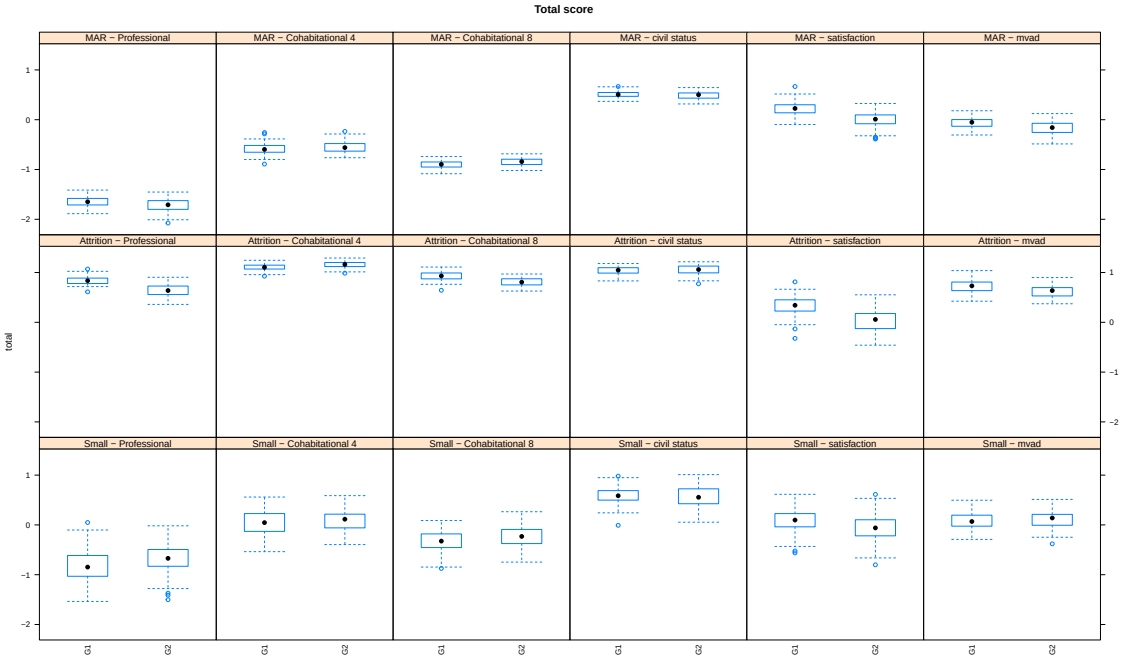



Figure E.7: VLMC - Total score for the results obtained with both the gain functions (G1 and G2) on each of the 18 scenarios.

# Appendix F

# Additional results for the multichannel comparison

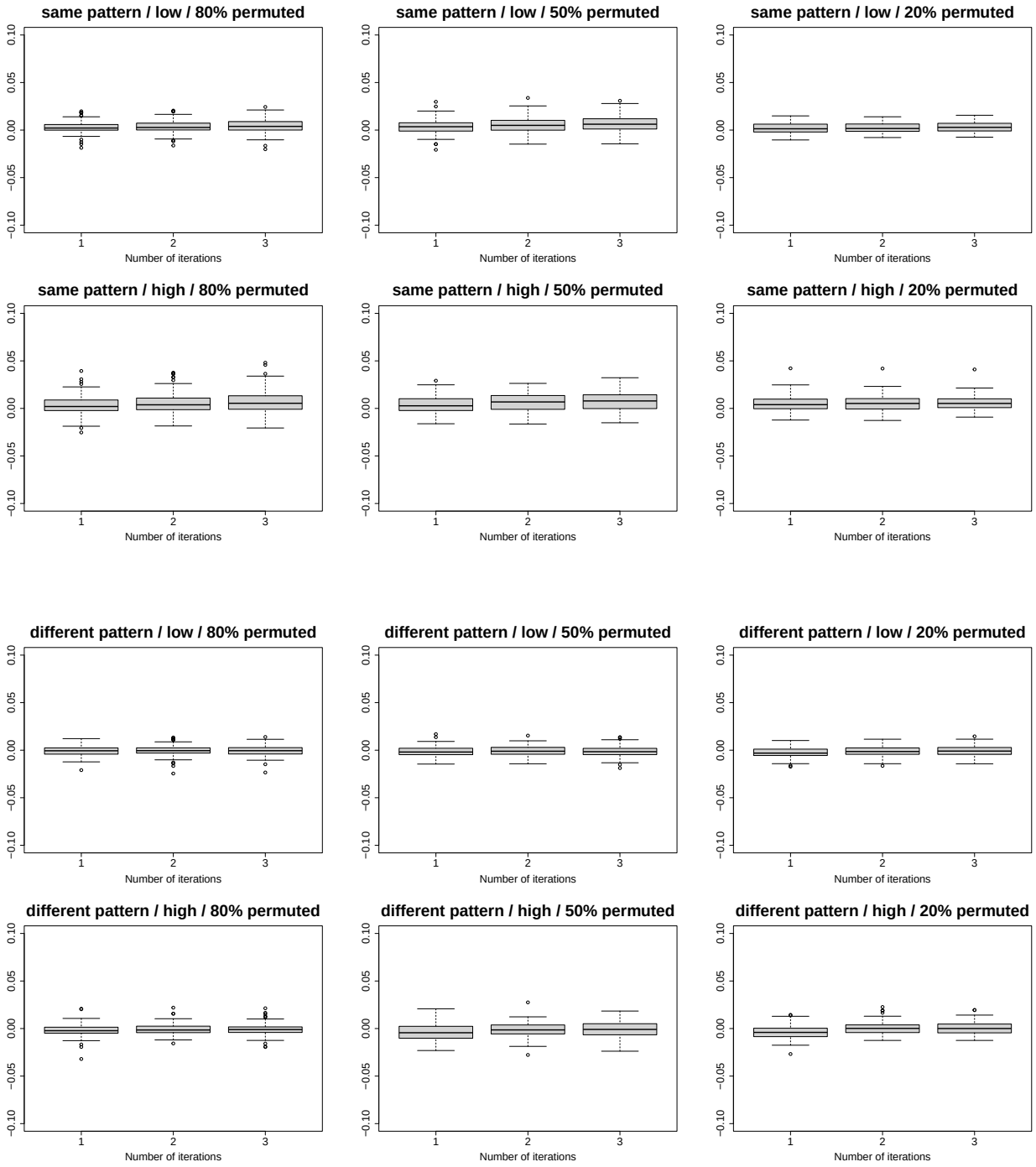



Figure F.1: MAR mechanism - Boxplots of the criteria relative to the local association, , obtained from handling missing data on the civil status dataset with 1 to 3 iterations of the *MICT-multichannel* algorithm. Each subplot corresponds to a scenario of missing data generation and is labelled as "type of pattern / rate of missing data / % of sequences from the duplicated dataset permuted".

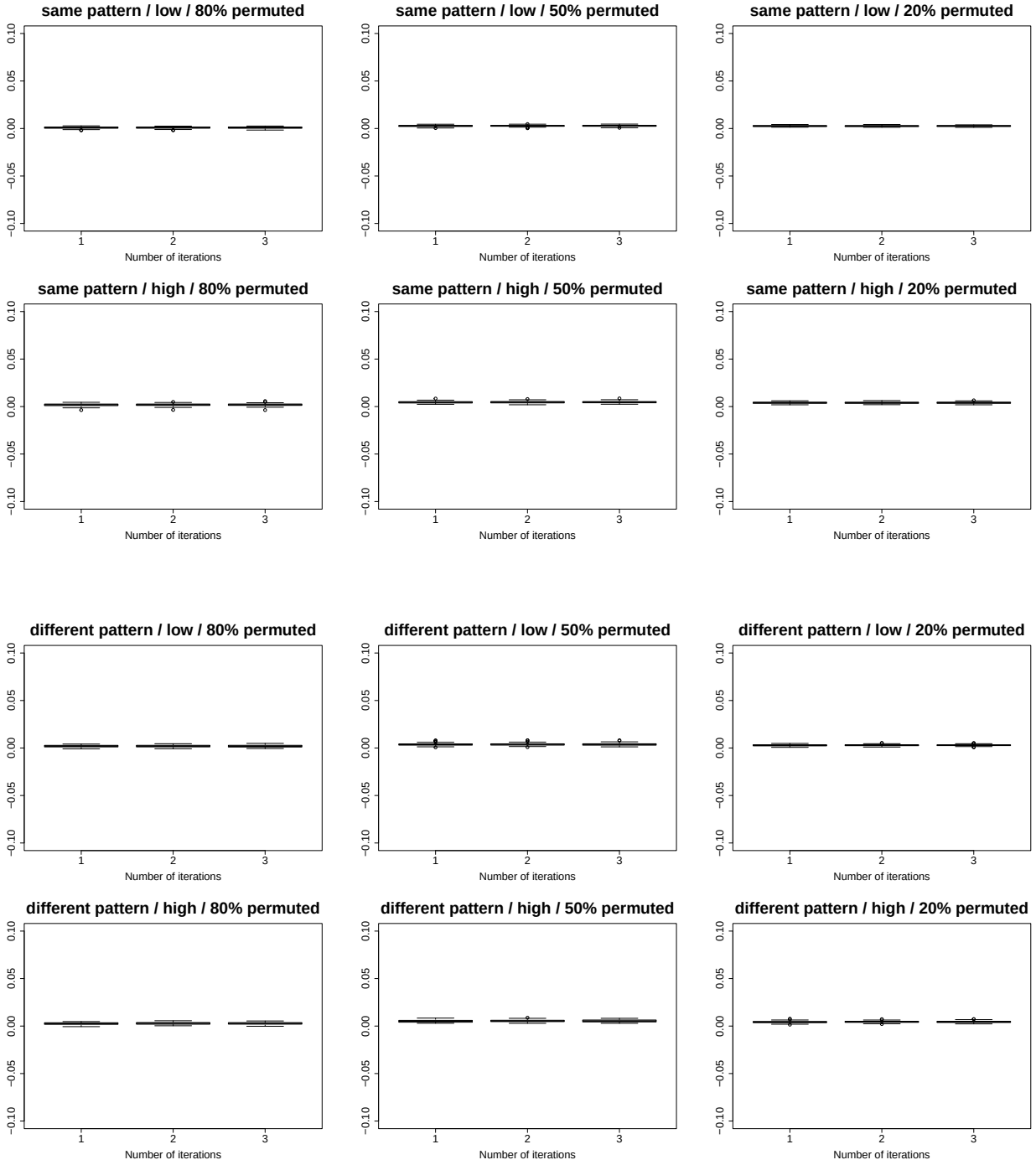



Figure F.2: MAR mechanism - Boxplots of the Cronbach's $\alpha$ bias, obtained from handling missing data on the civil status dataset with 1 to 3 iterations of the *MICT-multichannel* algorithm. Each subplot corresponds to a scenario of missing data generation and is labelled as "type of pattern / rate of missing data / % of sequences from the duplicated dataset permuted".

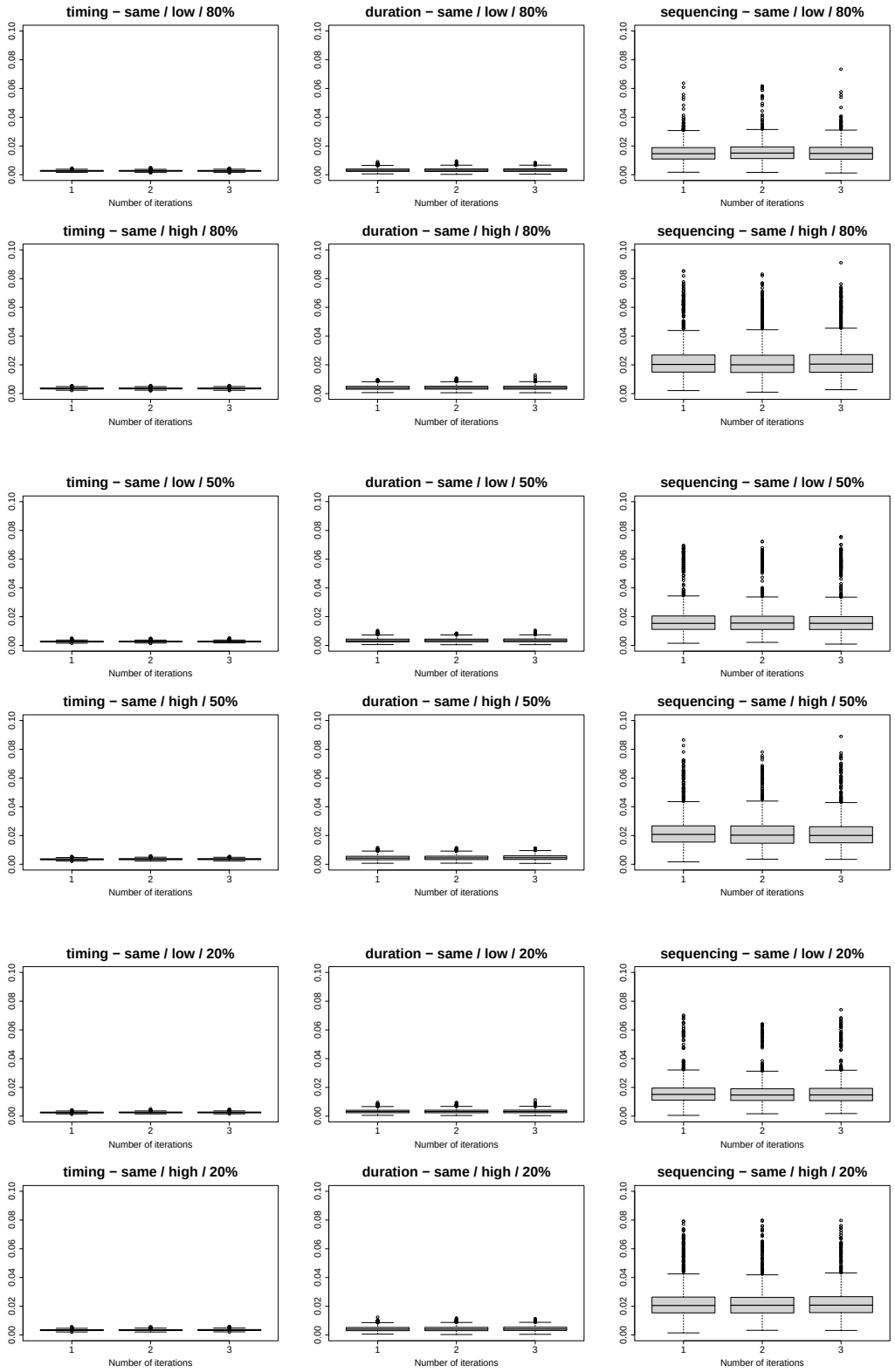



Figure F.3: MAR mechanism - Boxplots of the longitudinal characteristics bias, obtained from handling missing data on the civil status dataset with 1 to 3 iterations of the *MICT-multichannel* algorithm. Each row of subplots corresponds to a scenario of missing data generation with a same pattern of missing values and is labelled as "longitudinal characteristic - type of pattern / rate of missing data / % of sequences from the duplicated dataset permuted".

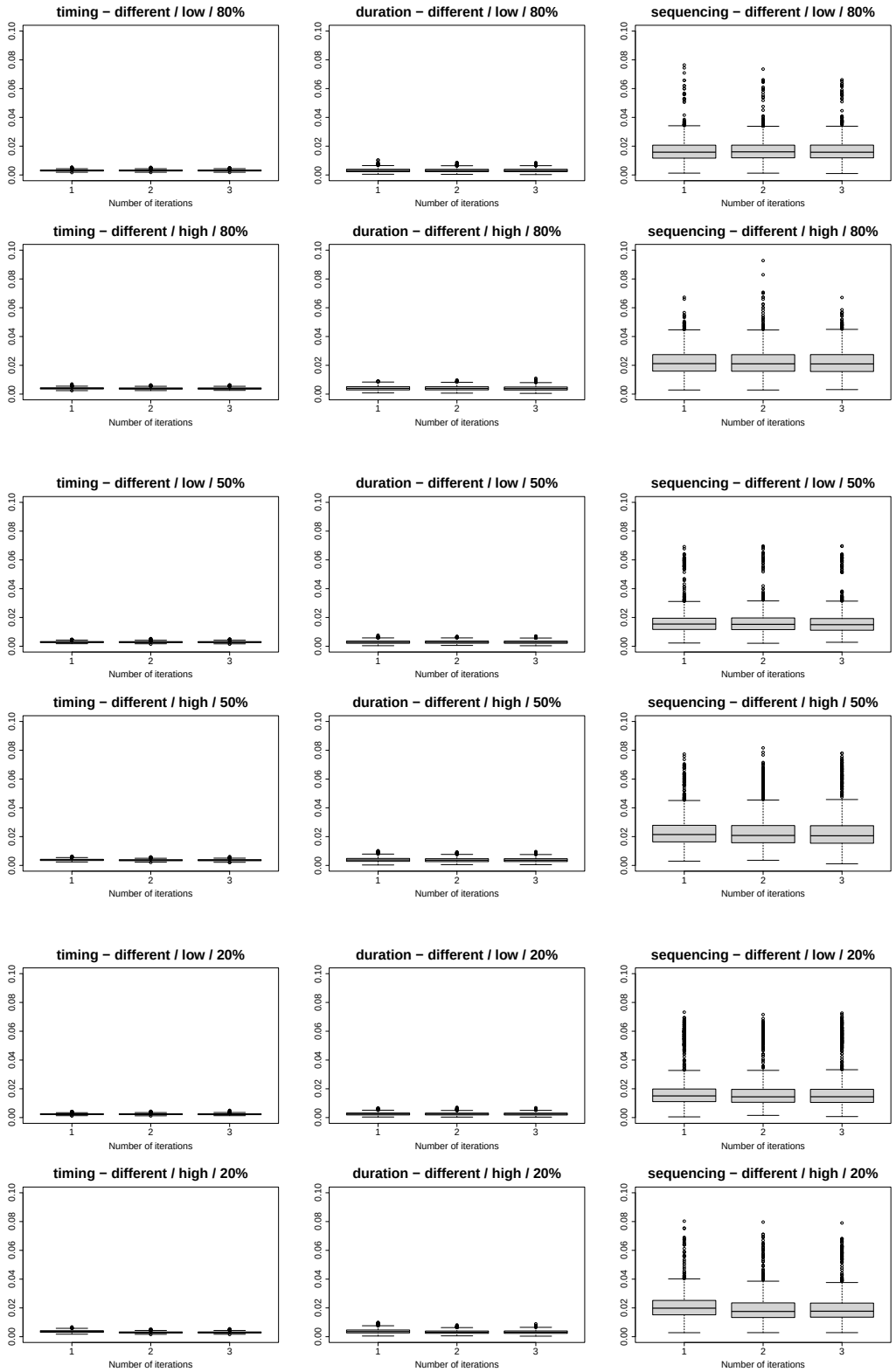



Figure F.4: MAR mechanism - Boxplots of the longitudinal characteristics bias, obtained from handling missing data on the civil status dataset with 1 to 3 iterations of the *MICT-multichannel* algorithm. Each row of subplots corresponds to a scenario of missing data generation with a different pattern of missing values and is labelled as "longitudinal characteristic - type of pattern / rate of missing data / % of sequences from the duplicated dataset permuted".

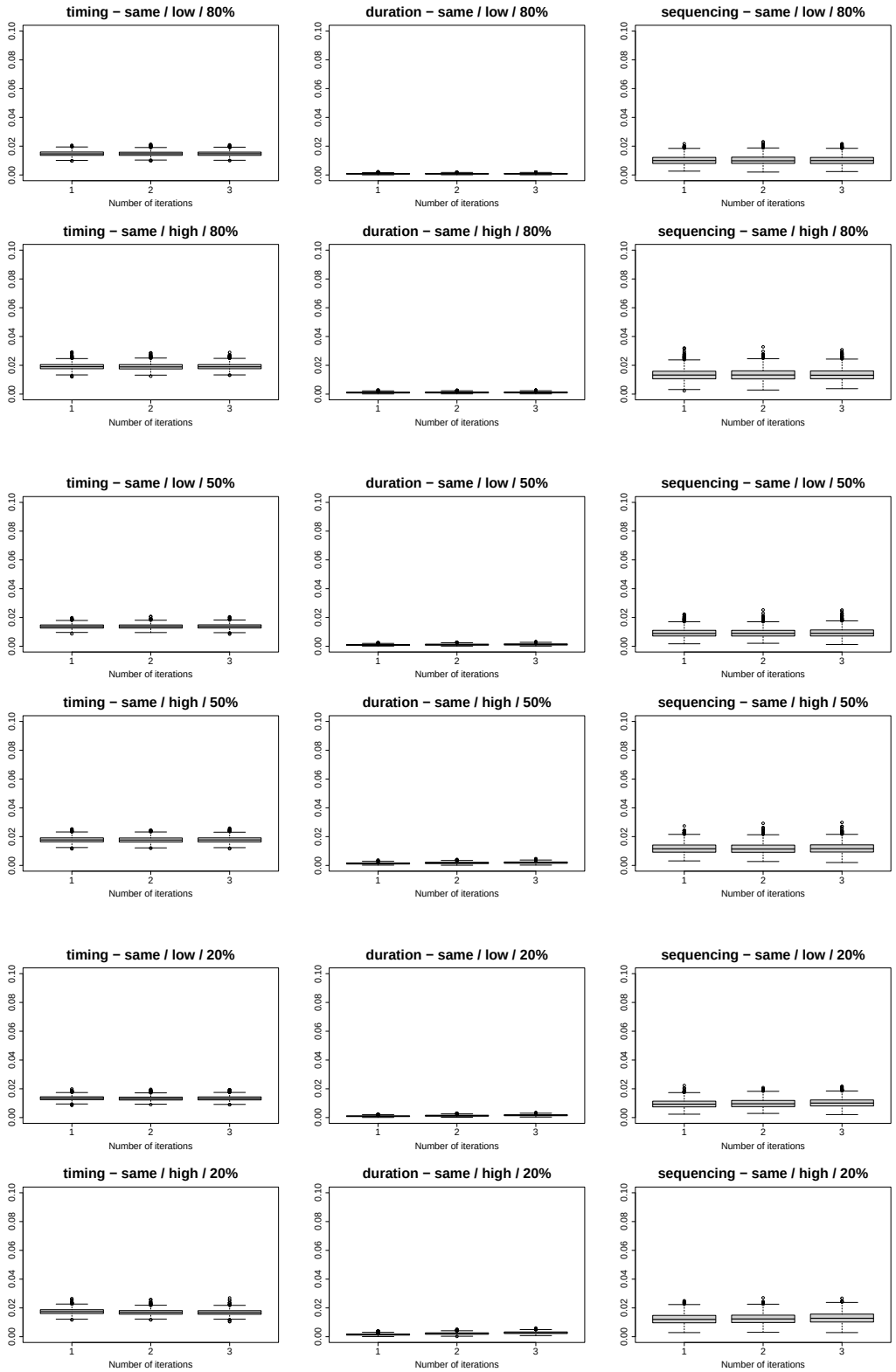



Figure F.5: MAR mechanism - Boxplots of the longitudinal characteristics bias, obtained from handling missing data on the satisfaction with health status dataset with 1 to 3 iterations of the *MICT-multichannel* algorithm. Each row of subplots corresponds to a scenario of missing data generation with a same pattern of missing values and is labelled as "longitudinal characteristic - type of pattern / rate of missing data / % of sequences from the duplicated dataset permuted".

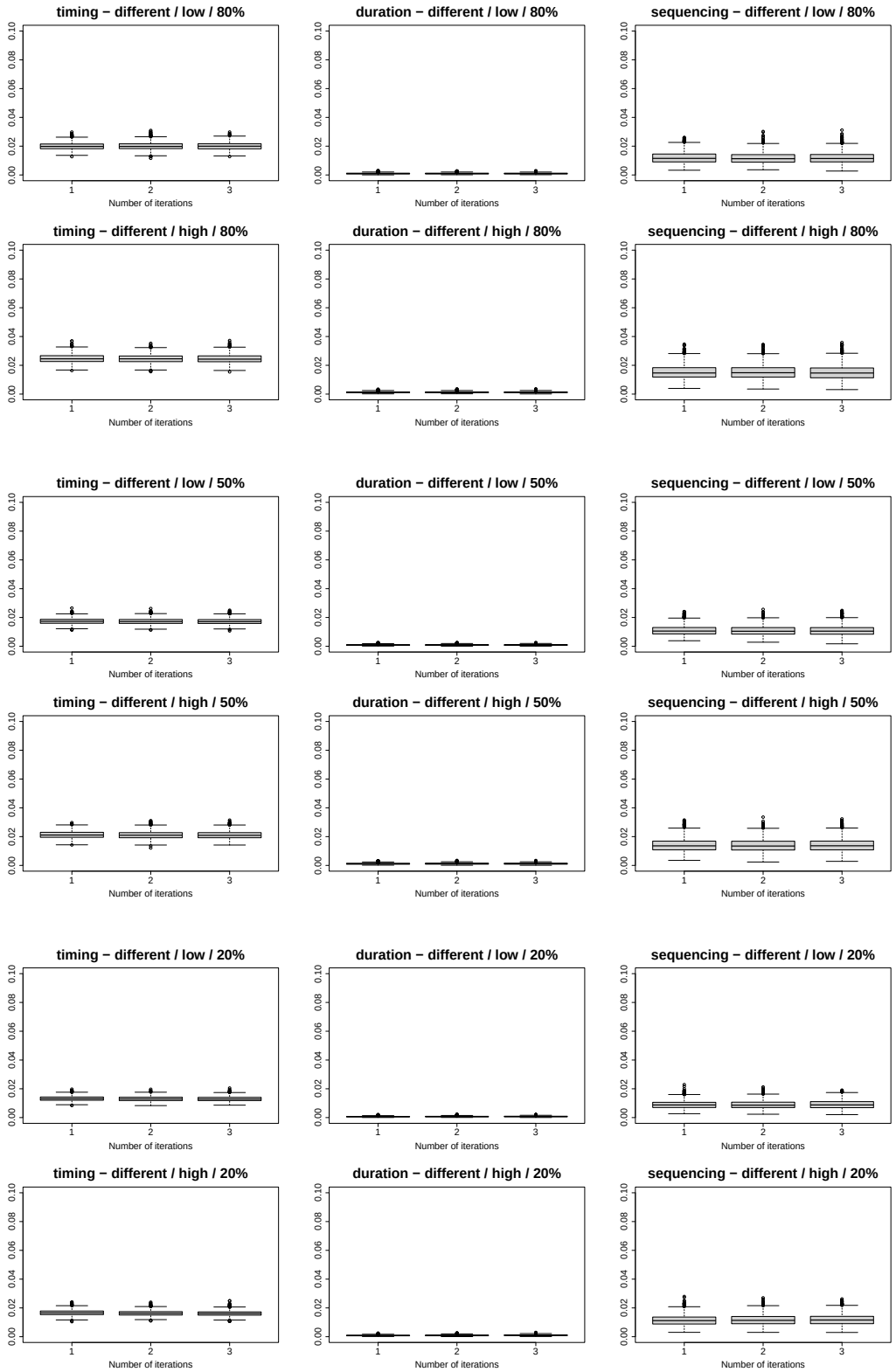



Figure F.6: MAR mechanism - Boxplots of the longitudinal characteristics bias, obtained from handling missing data on the satisfaction with health status dataset with 1 to 3 iterations of the *MICT-multichannel* algorithm. Each row of subplots corresponds to a scenario of missing data generation with a different pattern of missing values and is labelled as "longitudinal characteristic - type of pattern / rate of missing data / % of sequences from the duplicated dataset permuted".

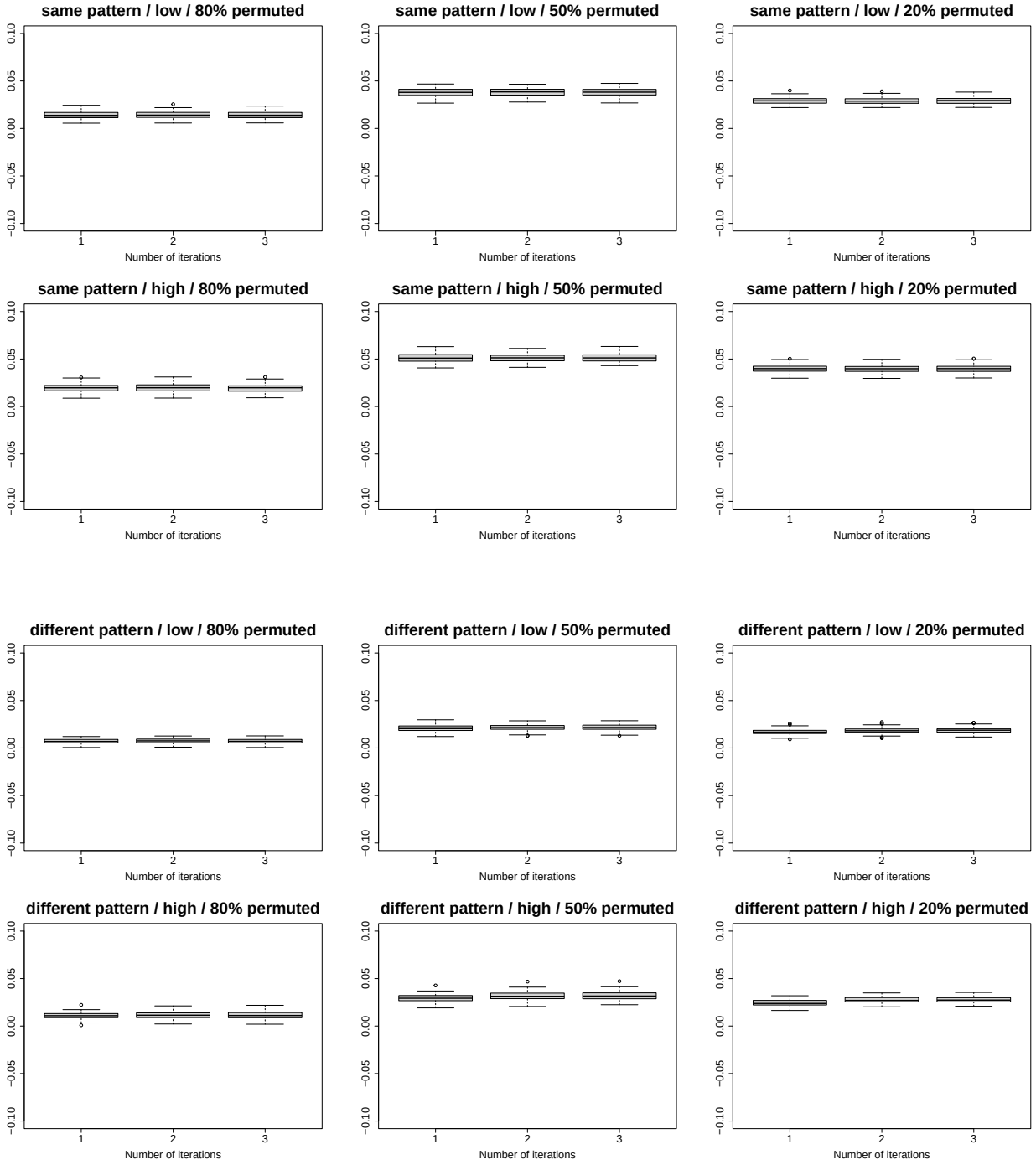



Figure F.7: MAR mechanism - Boxplots of the Cronbach's $\alpha$ bias, obtained from handling missing data on the professional status dataset with 1 to 3 iterations of the *MICT-multichannel* algorithm. Each subplot corresponds to a scenario of missing data generation and is labelled as "type of pattern / rate of missing data / % of sequences from the duplicated dataset permuted".

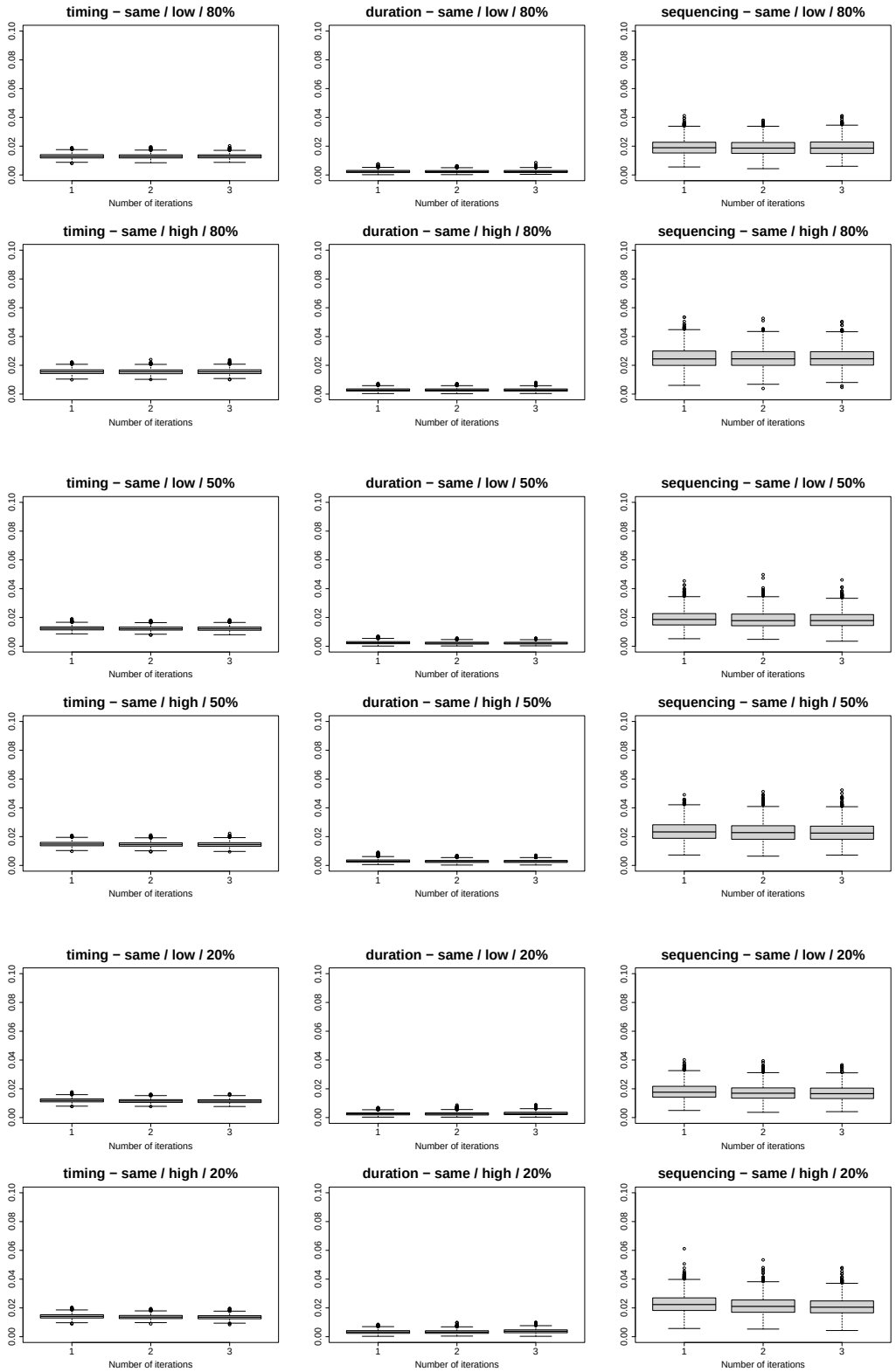



Figure F.8: MAR mechanism - Boxplots of the longitudinal characteristics bias, obtained from handling missing data on the professional status dataset with 1 to 3 iterations of the *MICT-multichannel* algorithm. Each row of subplots corresponds to a scenario of missing data generation with a same pattern of missing values and is labelled as "longitudinal characteristic - type of pattern / rate of missing data / % of sequences from the duplicated dataset permuted".

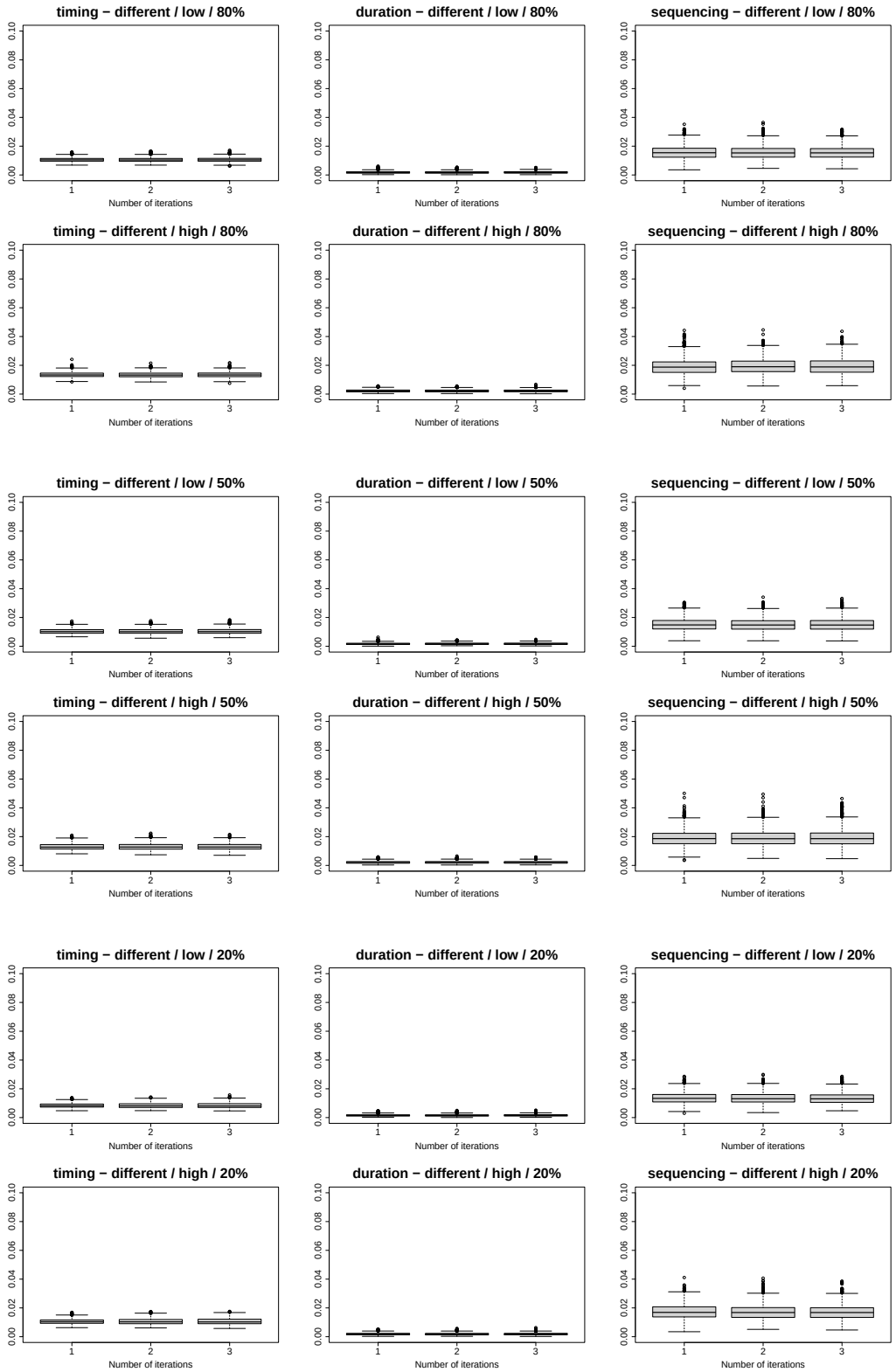



Figure F.9: MAR mechanism - Boxplots of the longitudinal characteristics bias, obtained from handling missing data on the professional status dataset with 1 to 3 iterations of the *MICT-multichannel* algorithm. Each row of subplots corresponds to a scenario of missing data generation with a different pattern of missing values and is labelled as "longitudinal characteristic - type of pattern / rate of missing data / % of sequences from the duplicated dataset permuted".

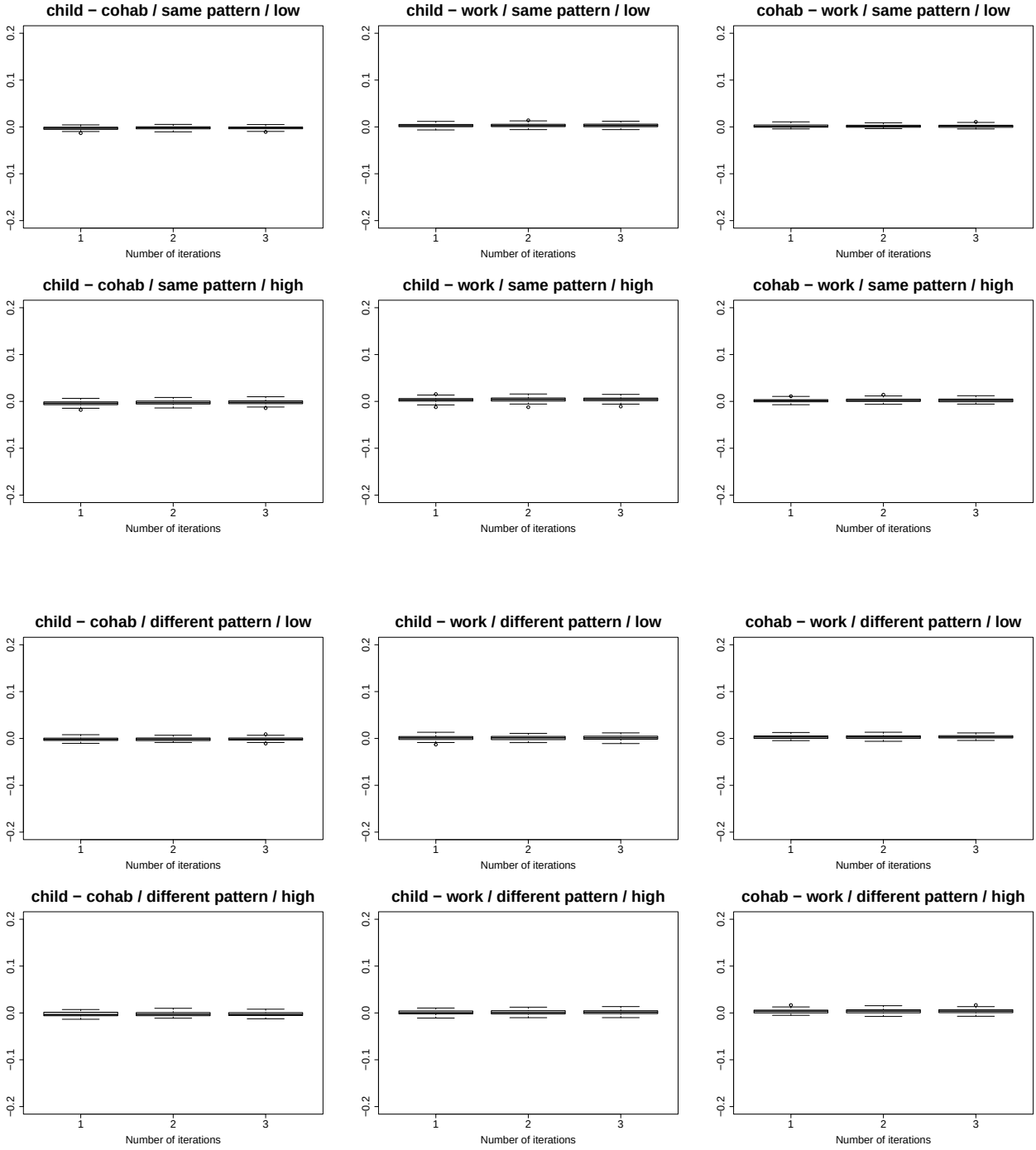



Figure F.10: MAR mechanism - Boxplots of the bias of the Cramer's V , for each pair of channels, obtained from handling missing data with 1 to 3 iterations of the *MICT-multichannel* algorithm. Each subplot corresponds to a scenario of missing data generation. It is labelled as "channels considered / type of pattern / rate of missing data".

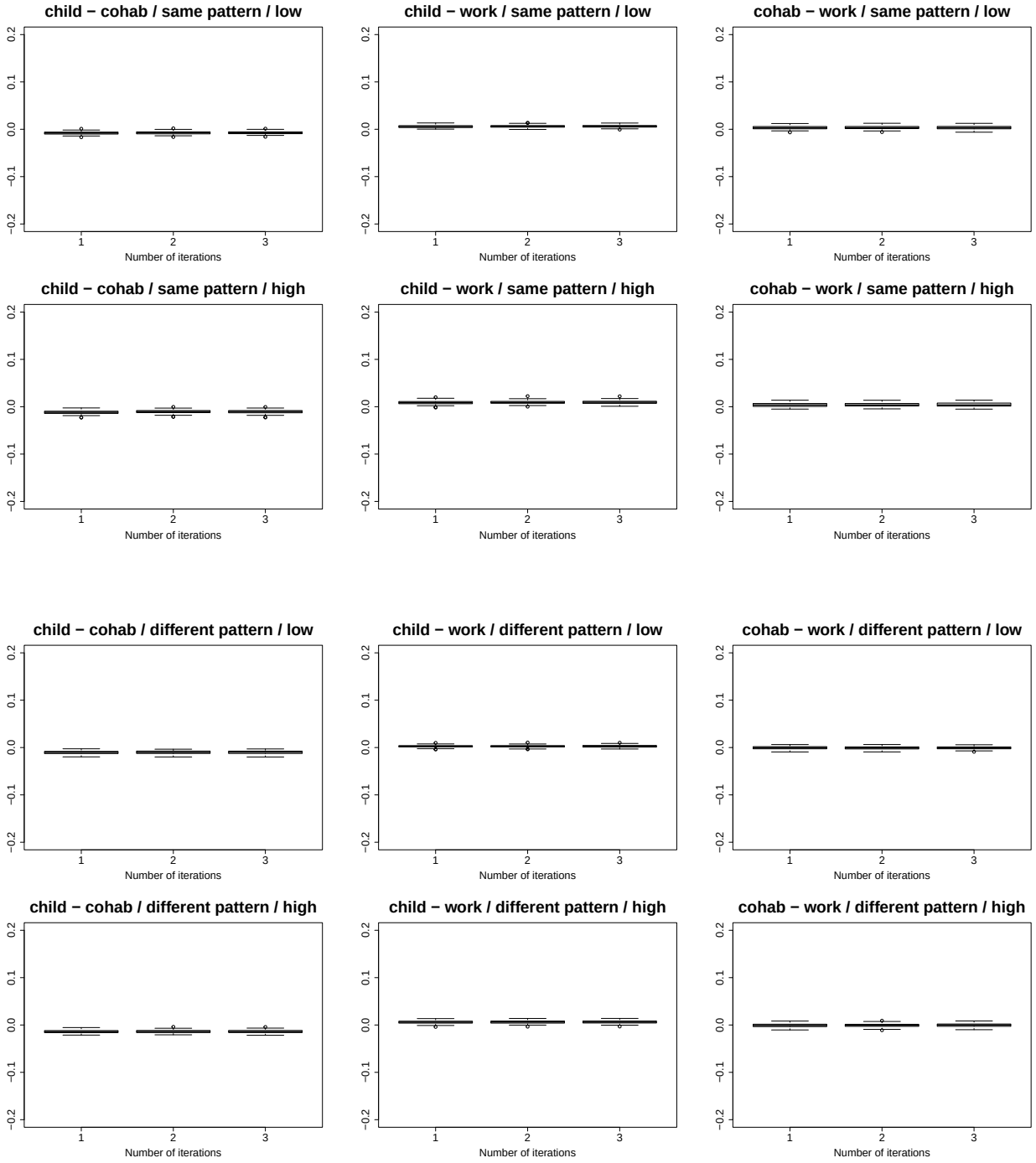



Figure F.11: MAR mechanism - Boxplots of the bias of the Cronbach's $\alpha$ computed, for each pair of channels, obtained from handling missing data with 1 to 3 iterations of the *MICT-multichannel* algorithm. Each subplot corresponds to a scenario of missing data generation. It is labelled as "channels considered / type of pattern / rate of missing data".

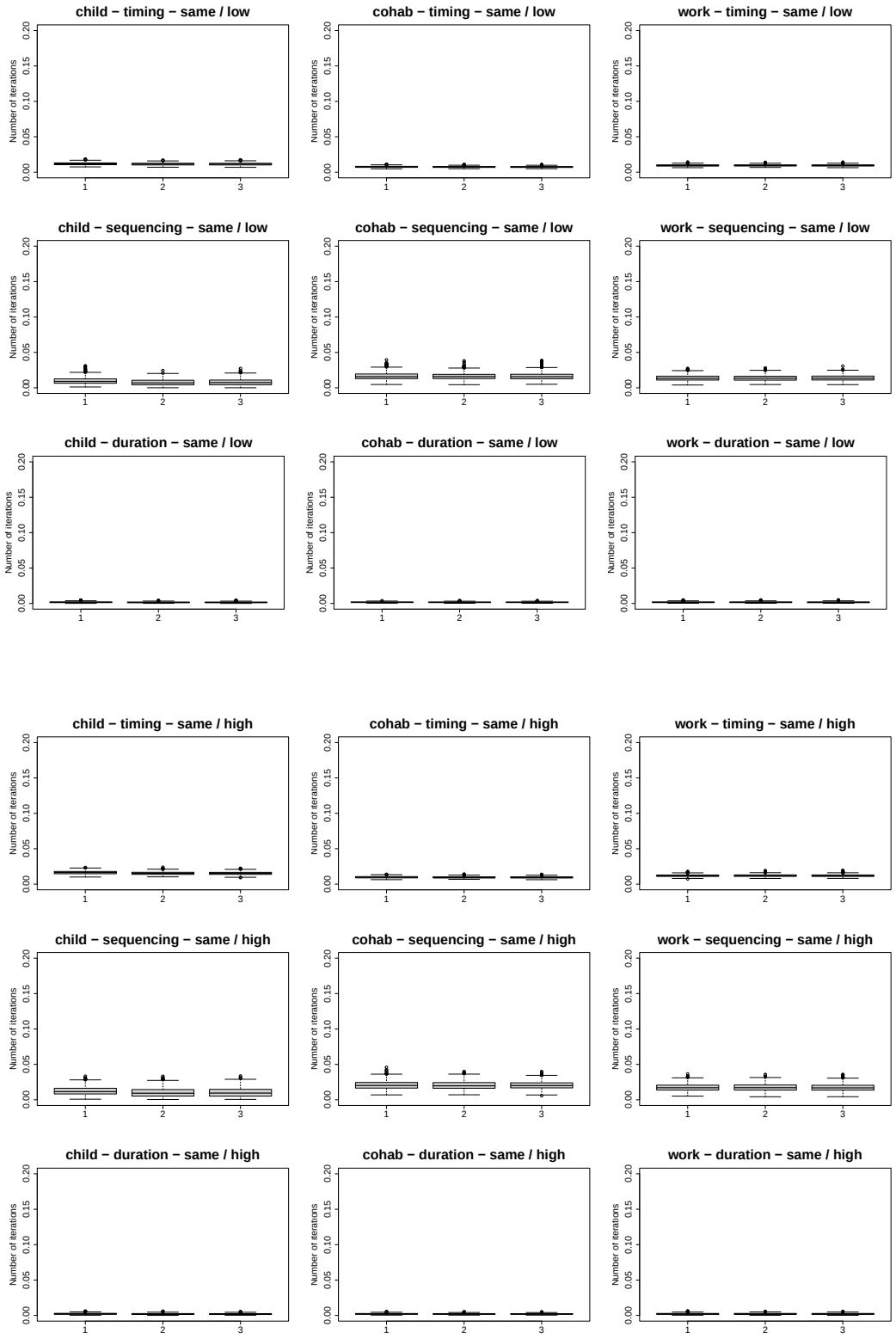



Figure F.12: Case of a MAR mechanism and same patterns of missing data. Boxplots of the bias on the criteria, obtained with the five considered methods, namely *CCA*, *MICT* applied to each channel separately (labelled as "MICT-s"), *MICT-multichannel* (labelled as "MICT-m"), *FCS* and *two-fold FCS* (labelled as "2folds FCS"). Each subplot corresponds to a scenario of missing data generation. It is labelled as "channel considered - criterion - type of pattern / rate of missing data".

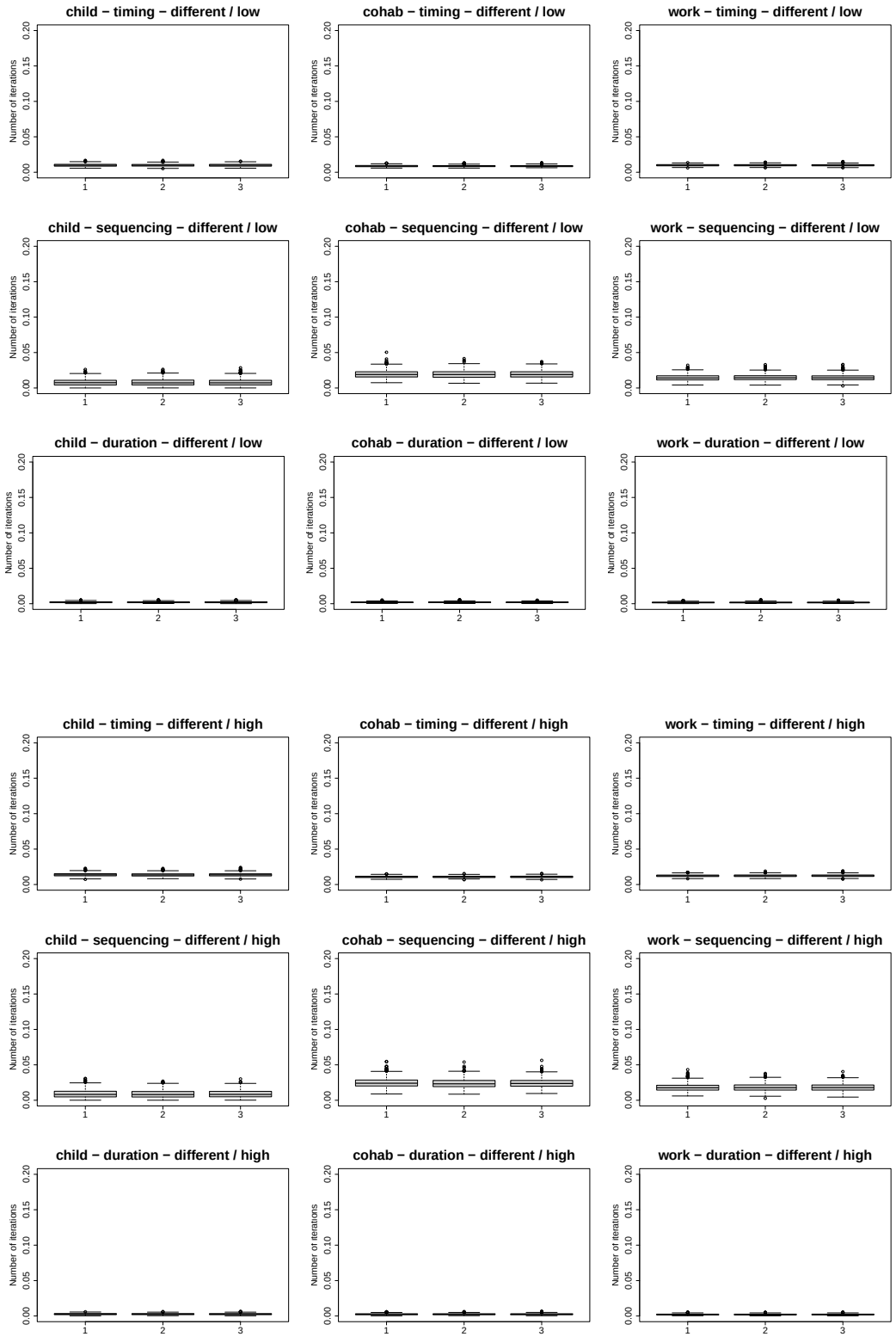



Figure F.13: Case of a MAR mechanism and different patterns of missing data. Boxplots of the bias on the longitudinal criteria, obtained with the five considered methods, namely *CCA*, *MICT* applied to each channel separately (labelled as "MICT-s"), *MICT-multichannel* (labelled as "MICT-m"), *FCS* and *two-fold FCS* (labelled as "2folds FCS"). Each subplot corresponds to a scenario of missing data generation. It is labelled as "channel considered - criterion - type of pattern / rate of missing data".


Figure F.14: MAR mechanism - Boxplots of the bias of the Cramer's V , for each pair of channels, obtained from handling missing data with *MICT-multichannel* algorithm and different orders for the channels. Each subplot corresponds to a scenario of missing data generation. It is labelled as "channels considered / type of pattern / rate of missing data" Each boxplot is labelled based on the order, where "ch" correspond to child, "co" to cohabitational status and "wo" for work status. Therefore, as an example, "ch-co-wo" means that the child channel was imputed first, then the cohabitational one and finally the professional one.

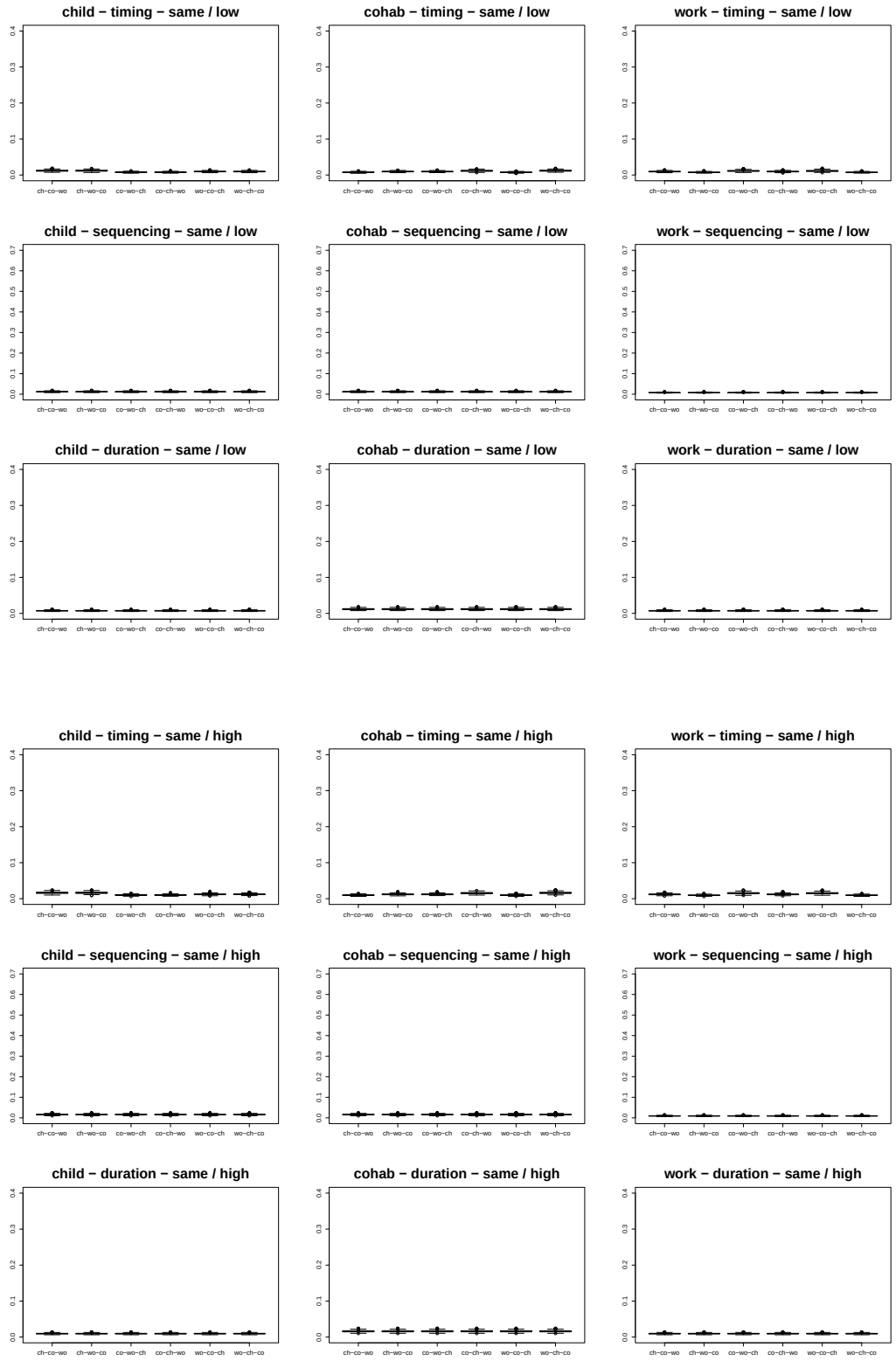



Figure F.15: MAR mechanism - Boxplots of the bias on the longitudinal criteria, from handling missing data with *MICT-multichannel* algorithm and different order for the channels. Each subplot corresponds to a scenario of missing data generation with same patterns of missing data. It is labelled as "channel considered / type of pattern / rate of missing data" Each boxplot is labelled based on the order, where "ch" correspond to child, "co" to cohabitational status and "wo" for work status. Therefore, as an example, "ch-co-wo" means that the child channel was imputed first, then the cohabitational one and finally the professional one.

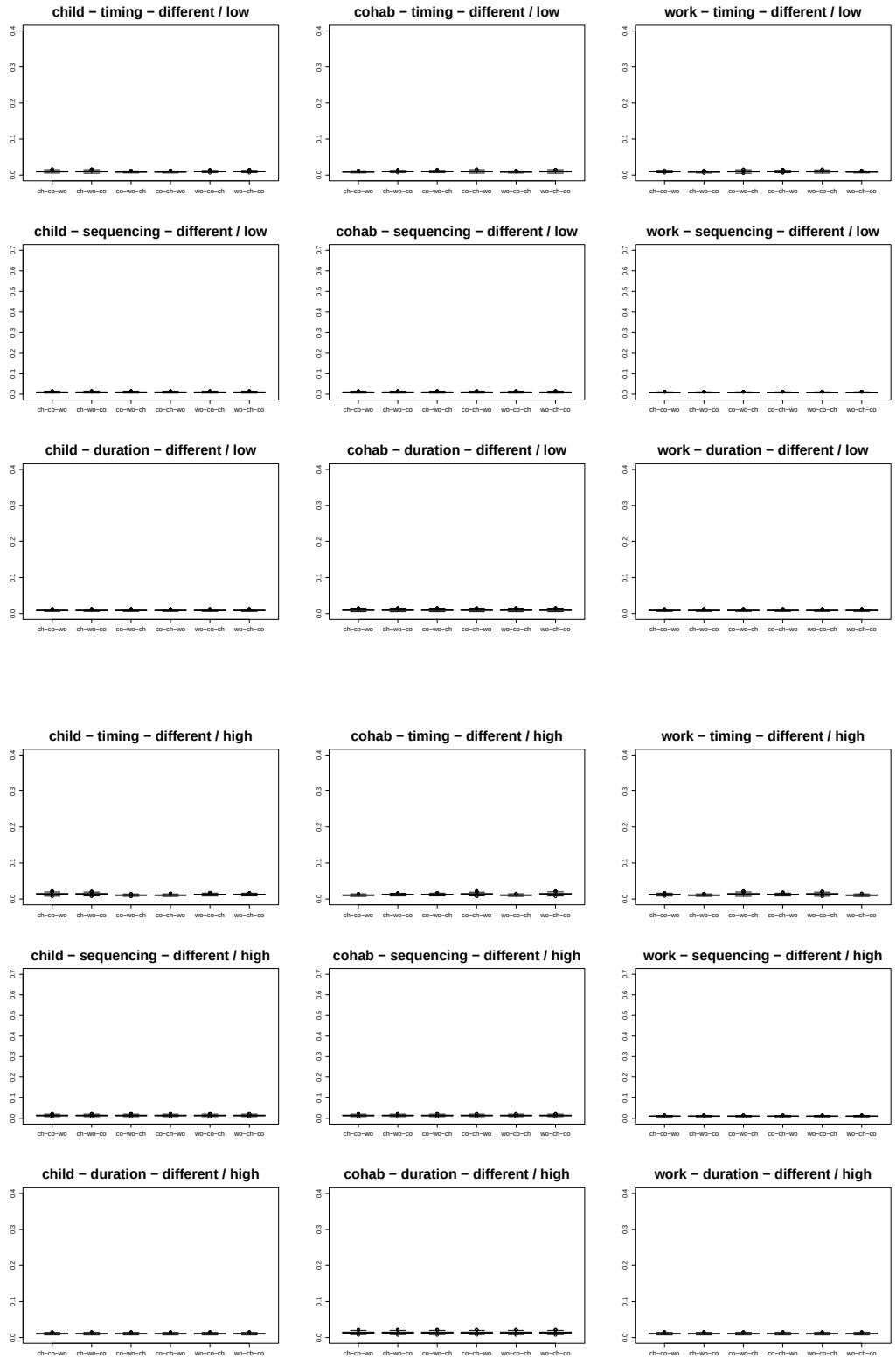



Figure F.16: MAR mechanism - Boxplots of the bias on the longitudinal criteria, from handling missing data with *MICT-multichannel* algorithm and different order for the channels. Each subplot corresponds to a scenario of missing data generation with different patterns of missing data. It is labelled as "channel considered / type of pattern / rate of missing data" Each boxplot is labelled based on the order, where "ch" correspond to child, "co" to cohabitational status and "wo" for work status. Therefore, as an example, "ch-co-wo" means that the child channel was imputed first, then the cohabitational one and finally the professional one.

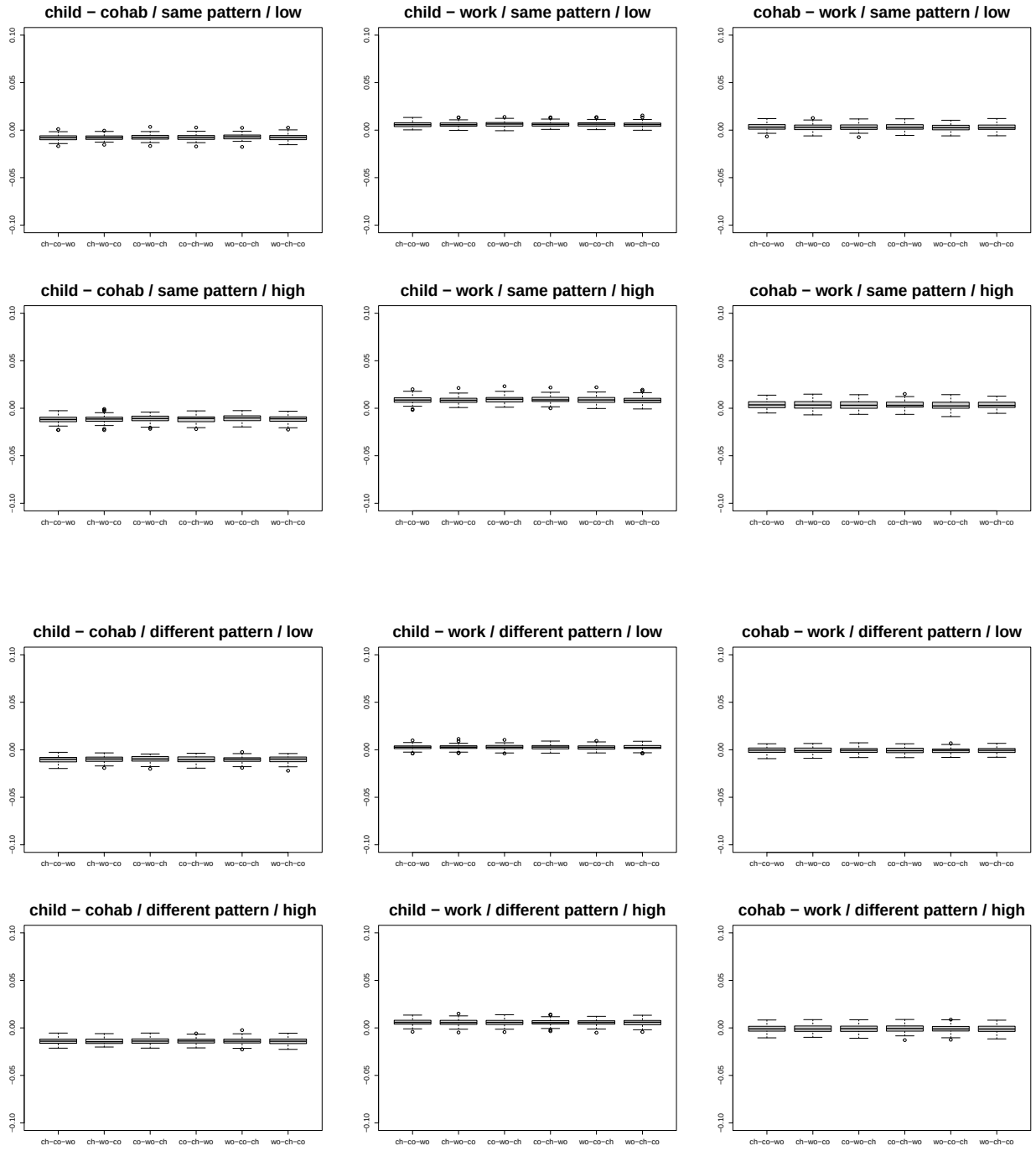



Figure F.17: MAR mechanism - Boxplots of the bias of the Cronbach's $\alpha$ computed at time 18, for each pair of channels, obtained from handling missing data with *MICT-multichannel* algorithm and different orders for the channels. Each subplot corresponds to a scenario of missing data generation. It is labelled as "channels considered / type of pattern / rate of missing data" Each boxplot is labelled based on the order, where "ch" correspond to child, "co" to cohabitational status and "wo" for work status. Therefore, as an example, "ch-co-wo" means that the child channel was imputed first, then the cohabitational one and finally the professional one.