UNIVERSITÄT TRIER

Data Fusion in Official Statistics: An Evaluation of Classical versus Statistical Learning Approaches

Dissertation

approved by Faculty IV of Trier University for the award of the academic degree

Doctor rerum politicarum (Dr. rer. pol.)

by

Jannik Schaller

Supervisor:	Prof. Dr. Ralf Münnich
Reviewer:	Prof. Dr. Ralf Münnich
	Prof. Dr. Hans Kiesl
Date of defence:	18 October 2023

Trier, 2024

This thesis was supported by the German Research Foundation (DFG) within the framework of the research project 'Multisectoral Regional Microsimulation Model (MikroSim)' (FOR 2559).

Acknowledgements

My first sincere gratitude goes to my first supervisor Ralf Münnich for his scientific and professional support, for the trust placed in me and for the great opportunity to carry out this research as part of an ongoing DFG research project. I also want to thank Hans Kiesl for agreeing to be my second supervisor and for his valuable comments. Furthermore, I am grateful to Florian Meinfelder for numerous scientific discussions. His expertise in data fusion has repeatedly provided valuable input that have often led to new ideas or to concrete suggestions on specific issues. I would also like to thank Susanne Rässler, whose enthusiasm and scientific expertise sparked my interest in statistics and missing data techniques early on.

This thesis was completed at the Federal Statistical Office of Germany (Destatis) as part of the DFG research group FOR 2559 'Multisectoral Regional Microsimulation Model (MikroSim)'. I am grateful for the funding provided by the German Research Foundation (DFG).

A special thank you goes to the entire team at the Federal Statistical Office of Germany, especially at the Institute for Research and Development in Federal Statistics and the Research Data Centre, as well as to the entire MikroSim research group. In particular, I would like to thank Hanna Brenzel, Jana Emmenegger, Jannek Mühlhan, Hariolf Merkle, Simon Schmaus, Sarah Bohnensteffen and Martin Palm. I am especially grateful for the excellent, constructive and joyful collaboration, the mutual support, for example in joint research work or proofreading, as well as for the regular academic exchange, which always brought valuable input with new ideas and new motivation.

Finally, I would like to express my particular gratitude to my family and friends, who are the greatest and most reliable support in all situations in life. Their always open ear and their ability to cheer me up with encouraging words during academic lows or calm me down with grounding words during academic highs were essential for completing this thesis.

Contents

German Summary

Datenfusionen sind in der amtlichen Statistik von stetig zunehmender Relevanz. Das Ziel einer Datenfusion besteht darin, zwei oder mehr Datenquellen über statistische Verfahren miteinander zu verbinden, um verschiedene Merkmale, die nicht zusammen in einer Datenquelle beobachtet wurden, gemeinsam auswerten zu können. Ein direktes Verknüpfen amtlicher Datenquellen anhand eindeutiger Identifikatoren ist aufgrund methodischer und rechtlicher Restriktionen häufig nicht möglich. Zielführende Datenfusionsmethoden sind daher von zentraler Bedeutung, um die vielfältigen Datenquellen der amtlichen Statistik effektiver nutzen und verschiedene Merkmale gemeinsam analysieren zu können. Allerdings fehlt es der Literatur an umfassenden Evaluationen dahingehend, welche Fusionsansätze unter welchen Datenkonstellationen vielversprechende Ergebnisse liefern. Das zentrale Ziel der vorliegenden Arbeit besteht deshalb darin, eine konkrete Bandbreite möglicher Fusionsalgorithmen, die neben klassischen Imputationsansätzen auch Verfahren des Statistical und Machine Learning umfasst, in ausgewählten Datenkonstellationen zu bewerten.

Zur Spezifikation und Identifikation dieser Datenkontexte werden daten- und imputationsbezogene Szenarientypen einer Datenfusion eingeführt: Explizite Szenarien, implizite Szenarien und Imputationsszenarien. Aus diesen drei Szenarientypen werden für die amtliche Statistik besonders relevante Fusionsszenarien als Grundlage für die Simulationen und Evaluationen ausgewählt. Als explizite Szenarien dienen die Erfüllung oder Verletzung der zentralen Annahme bedingter Unabhängigkeit (CIA) sowie variierende Größenverhältnisse der zu fusionierenden Stichproben. Beide Aspekte dürften sich direkt, also explizit, auf die Performance verschiedener Fusionsmethoden auswirken. Als implizite Szenarien werden die addierte Stichprobengröße der zu fusionierenden Datenquellen sowie das Skalenniveau der zu imputierenden Variable betrachtet. Beide Aspekte legen aufgrund der Datenbeschaffenheit die Anwendbarkeit bestimmter Fusionsmethoden nahe oder schließen diese aus. Als Imputationsszenarien dienen die univariate oder simultane, multivariate Imputationslösung sowie die Imputation künstlich generierter oder bereits zuvor beobachteter Werte im Falle von metrischen Merkmalen.

Bezüglich der konkreten Bandbreite möglicher Fusionsalgorithmen werden mit Distance Hot Deck (DHD), dem Regressionsmodell (RM) und Predictive Mean Matching (PMM) drei klassische Imputationsansätze betrachtet. Mit Decision Trees (DT) und Random Forest (RF) werden wiederum zwei prominente, baumbasierte Verfahren aus dem Statistical Learning-Bereich im Kontext der Datenfusion diskutiert. Derartige Prädiktionsverfahren zielen jedoch darauf ab, individuelle Werte möglichst präzise vorherzusagen, was mit dem vordergründigen Anspruch einer Datenfusion, der Reproduktion gemeinsamer Verteilungen, kollidieren kann. Zudem umfassen DT und RF lediglich univariate Imputationslösungen und es werden, im Falle metrischer Variablen, künstlich generierte statt real beobachtete Werte imputiert. Daher wird mit Predictive Value Matching (PVM) ein neues, Statistical Learning-basiertes Nächste-Nachbar-Verfahren vorgestellt, welches die verteilungstechnischen Nachteile von DT und RF überwinden könnte, eine uni- und multivariate Imputationslösung bietet und darüber hinaus, bezüglich metrischer Merkmale, reale und zuvor beobachtete Werte imputiert. Sämtliche Prädiktionsverfahren können dem neuen PVM-Ansatz zugrundeliegen. Im Rahmen dieser Arbeit wird PVM auf Basis von Decision Trees (PVM-DT) und Random Forest (PVM-RF) betrachtet.

Die zugrundeliegenden Fusionsmethoden werden in umfassenden Simulationen und Evaluationen untersucht. Dabei fokussiert sich die Evaluation der verschiedenen Datenfusionsverfahren auf die ausgewählten Fusionsszenarien. Die Grundlage hierfür bilden zwei konkrete und aktuelle Anwendungsfälle der Datenfusion in der amtlichen Statistik, die Fusion von EU-SILC und Household Budget Survey einerseits sowie von Einkommensteuerstatistik und Mikrozensus andererseits. Beide Anwendungsfälle weisen wesentliche Unterschiede hinsichtlich verschiedener Fusionsszenarien auf und dienen somit dem Zweck, eine Vielzahl von Datenkonstellationen abzudecken. Aus beiden Anwendungsfällen werden Simulationsdesigns entwickelt, wobei insbesondere die expliziten Szenarien in die Simulationen eingearbeitet werden.

Entlang der Ergebnisse erweist sich unter Erfüllung der CIA insbesondere PVM-RF als vielversprechender und universeller Fusionsansatz. Denn PVM-RF liefert sowohl für kategoriale, als auch für metrische zu imputierende Variablen zufriedenstellende Ergebnisse und bietet zudem, unabhängig vom Skalenniveau, eine uni- und multivariate Imputationslösung. Auch PMM stellt eine adäquate Fusionsmethode dar, jedoch nur in Bezug auf metrische Merkmale. Ebenfalls implizieren die Ergebnisse, dass die Anwendung der Statistical Learning-Methoden Chance und Risiko zugleich ist. Bei CIA-Verletzung können potentielle, auf Korrelationen bezogene Übertreibungseffekte von DT und RF, teilweise auch von RM, nützlich sein. Die übrigen Verfahren induzieren hingegen bei Verletzung der CIA schlechte Ergebnisse. Unter Erfüllung der CIA besteht jedoch das Risiko, dass die Prädiktionsmethoden RM, DT und RF Zusammen-

hänge überschätzen. Die Größenverhältnisse der zu fusionierenden Studien weisen wiederum einen eher untergeordneten Einfluss auf die Performance von Fusionsmethoden aus. Dies ist eine wichtige Implikation dahingehend, dass nicht zwangsläufig, wie bisher üblich, der größere Datensatz als Spenderstudie dienen muss.

Die Ergebnisse der Simulationen und Evaluationen münden in konkrete Implikationen dahingehend, welche Datenfusionsmethoden unter den ausgewählten Daten- und Imputationskonstellationen verwendet und betrachtet werden sollten. Von diesen Implikationen profitiert die Wissenschaft im Allgemeinen sowie die amtliche Statistik im Besonderen. Denn sie bieten für künftige Datenfusionsvorhaben wichtige Anhaltspunkte, um zu beurteilen, welche konkrete Datenfusionsmethode adäquate Ergebnisse entlang der in dieser Arbeit untersuchten Datenkonstellationen liefern könnte. Ebenfalls bietet die Arbeit mit PVM eine vielversprechende, methodische Innovation für künftige Datenfusionen sowie für Imputationsprobleme im Allgemeinen.

List of Figures

List of Tables

List of Symbols

Unless otherwise defined, the following notations apply:

b	Index for bootstrap samples
B	Number of bootstrap samples
β	Regression coefficient
$D_{i,j}$	Distance between recipient unit i and donor unit j
$\mathrm{E}(\cdot)$	Expectation
ε	Vector of residuals
$f(\cdot)$	Density function/ arbitrary function
i	Index for entities from the recipient data
j	Index for entities from the donor data
k	Index of classes for categorical variables
K	Number of classes for categorical variables
m	Index of regions for Decision Trees
M	Number of regions for Decision Trees
N_{R_m}	Number of observations in region R_m
n_{don}	Sample size of the donor data
n_{rec}	Sample size of the recipient data
p	Number of common $\mathbf{X}$ variables
p_{don}	Number of specific $\mathbf{Z}$ variables
p_{rec}	Number of specific $\mathbf{Y}$ variables
r	Index for the specific $\mathbf{Z}$ variables
R	Regions for Decision Trees
s	Split point for Decision Trees

t	Index for the common **X** variables
X	Common variables
Y	Specific variables from the recipient data
Z	Specific variables from the donor data

Chapter 1

Introduction

Data fusion, also known as statistical matching, aims to provide an integrated microdata set that allows for a joint analysis of variables from (at least) two different data sources, while each of the data sources originally served a different purpose.[1] Merging data via fusion procedures is a central and increasingly relevant topic in official statistics. This is because often the characteristics relevant for a particular scientific analysis or evaluation are not found in one but in different data sources. A direct linkage of the datasets is often not possible due to legal restrictions and in favour of relieving the respondents. The data must therefore be fused by means of statistical procedures. In the scientific literature, however, a comprehensive evaluation of a multitude of possible fusion methods in different data contexts of official statistics is missing so far. Hence, one important innovation of this thesis is that concrete scenarios are defined and the performance of a variety of possible data fusion methods is evaluated based on these scenarios. In addition, a new and promising data fusion algorithm is introduced. To further motivate the investigations on data fusions in this work, we start with a general overview on official statistics data sources in Germany and then turn to identifying specific research gaps in the data fusion literature.

1.1 Principles of Official Statistics

Official statistics in Germany, like other National Statistical Institutes (NSIs), collect and provide a variety of different data on diverse topics. These range from socio-demographic and socio-economic surveys such as the Census or the Microcensus to health or environmental

[1] It should be noted that throughout this work the terms *data fusion* and *statistical matching* are used synonymously, as are the terms *fuse* and *match*.

statistics and statistics from the construction industry or the transport sector, to name just a few.[2] Accordingly, official statistics in Germany and other countries provide data on almost all socially relevant topics. However, the single official statistics data sources of Germany and other NSIs are often committed to a particular objective. The European Union Statistics on Income and Living Conditions (EU-SILC) (see e.g. Eurostat 2020), for example, is dedicated to the detailed measurement of the income and various income components of private households, while the Household Budget Survey (HBS)[3] (see e.g. Eurostat 2015) provides extensive information on the consumption expenditures of private households. While both studies allow a detailed analysis of the respective components, EU-SILC regarding the income and HBS concerning the consumption expenditures, it is apparent that a joint analysis of the income variables from EU-SILC and the consumption information from HBS is not viable with the single datasets.

From the perspective of official statistics, the concentration of various official statistics on a specific topic has its reasonable and justified background. On the one hand, a challenge with surveys in general is to reduce the response burden on participants. The associated intention is to motivate as many respondents as possible to provide complete and accurate information on the given topics, which will benefit the data quality, and to participate again in future surveys (see e.g. Giesen et al. 2018). The consequence of relieving the burden on respondents, however, is that surveys in official statistics often refer to a specific set of topics, with the exception of the extensive surveys for the German Microcensus, which are, though, subject to the obligation to provide information. On the other hand, particularly in Germany, every official statistic requires a legal basis that specifies the information to be collected (§ 5 para. 1 cl. 1 and § 9 para. 1 Bundesstatistikgesetz). The more subject areas that are to be collected within a statistic, the less likely political majorities are due to the legal data protection regulations in Germany, which in turn are necessary for the corresponding legal basis of an official statistic. In the context of data protection principles, of central relevance for official statistics in Germany is the so called *Volkszählungsurteil* (population census judgement) of the Bundesverfassungsgericht from 15 December 1983 (Bundesverfassungsgericht, decisions volume 65: 1-75). In light of the census scheduled for spring 1983 in Germany, the Bundesverfassungsgericht emphasised the right to informational self-determination in its ruling, thus making a landmark decision for official statistics in Germany with regard to the necessary data protection (Hornung and Schnabel 2009).

[2] An overview of available official microdata in Germany can be found, for example, on the homepage of the research data centres of the Statistical Offices of the Federal Republic and the Federal States at `https://www.forschungsdatenzentrum.de/en`.

[3] In Germany, the HBS equals the *Einkommens- und Verbrauchsstichprobe* (EVS).

The principles of data protection also affect the possibilities of combining information from the diverse official data sources. To jointly analyse characteristics stemming from different official statistics data sources, record linkage (see e.g. Herzog et al. 2010), that is, merging data files via (unique) identifiers available in the respective datasets, seems to be the most feasible approach. However, record linkage is generally not allowed with official statistics data sources due to legal restrictions and confidentiality reasons (§ 21 Bundesstatistikgesetz). Few exceptions exist, for example with the official economic statistics (§ 13a Bundesstatistikgesetz). As a consequence, direct and unique identifiers that are standardised across different official data sources are usually not available. Besides, from a methodological point of view, it is obvious that record linkage is hardly possible for samples that typically consist of different survey units. Hence, this would at best be conceivable for merging large samples or complete surveys, where the studies to be merged contain (at least mostly) the same survey units.

Compliance with these data protection aspects stipulated by politics and society and the resulting legal restrictions are elementary for official statistics, as this is the basic prerequisite for the trust of political decision-makers and citizens in official statistics. At the same time, however, these limitations are encountering a steadily growing demand for data in science, especially in the economic and social sciences as well as in medicine. Similarly, many socially relevant policy decisions are increasingly derived from data-driven, scientific studies. Not least the Covid-19 pandemic has shown the high relevance of reliable, high-quality data, the analysis potential of which can be further increased by linking different data sources. However, the required compliance with the necessary data protection regulation implies losses in the analytical potential of official statistics data sources and is thus contrary to the high and constantly growing data needs of science, and is also partly contrary to the efforts of official statistics to provide high-quality data on the population that are valuable for scientific analyses. This results in a fundamental trade-off between the data protection and the analysis potential of official statistics data sources.

This trade-off is also inherent if the scientific need arises to jointly analyse characteristics and variables that were collected in different data sources and have not (sufficiently) been jointly observed in a single official statistics dataset. Hence, an adequate linkage of the corresponding data sources is of central importance in order to achieve the required analysis, since record linkage is not possible due to legal and, in the case of samples, additionally due to methodological limitations. We will see in the following that such efforts to combine information from different data sources reflect current debates in official statistics.

In this instance, data fusion is a significant tool to provide an integrated microdata source where the different types of information relevant for the respective analyses are artificially joined on an individual or household level (Meinfelder and Schaller 2022). An appropriate fusion of different data sources could overcome the described trade-off between confidentiality principles and analysis potential to a considerable extent and thus be an opportunity for official statistics to unite both components more effectively. This emphasises the importance to investigate possible methods for different data fusion purposes for the current challenges in official statistics. For this could contribute to address the increasing data demands of science as well as the interest of official statistics in providing diverse data while taking legal and methodological restrictions into account. Thus, the main objective in this work is to analyse and evaluate potential data fusion methods to adequately fuse official statistics data sources in order to exploit the manifold potentials of the different official statistics data sources more effectively.

1.2 Research Gaps on Data Fusion

The literature, however, on concrete and practical data fusion methods to obtain an integrated microdata source appears diffuse and tends to leave practitioners from official statistics and other application areas with ambiguities rather than a concrete plethora of fusion options tailored to their respective data fusion objective and the underlying data situation. This could be due to the fact that many contributions on the data fusion literature are often devoted to a specific use case of data fusion (see e.g. Dalla Chiara et al. 2019; Lamarche et al. 2020), the central assumption and the natural uncertainty arising from it (see e.g. Kamakura and Wedel 1997; Moriarity and Scheuren 2001; D'Orazio et al. 2006a; Kiesl and Rässler 2006; Conti et al. 2012; Endres et al. 2019), as well as specific types of data fusion procedures (see e.g. Rodgers 1984; van der Putten et al. 2002; Gilula et al. 2006). In addition, since several years and with growing computing capacities, approaches from the field of statistical learning[4] have also been gaining popularity and are increasingly being considered for various statistical problems in general, but also sporadically discussed in the context of data fusion (see D'Ambrosio et al. 2012; D'Orazio 2019; Spaziani et al. 2019). Despite or perhaps because of this comprehensive amount of data fusion literature, it remains unclear which concrete data fusion implementation is suited under specific data circumstances and contexts, and to what extent statistical learning approaches can

[4]While the term *machine learning* seems to be more widespread in the literature, in this work we only rely on the term *statistical learning* since we apply and evaluate the original machine learning algorithms in a statistical context.

achieve a meaningful data fusion result. In this respect, the literature on data fusion methods lacks several aspects.

First, many comparative studies of different data fusion approaches are limited to specific archetypes of data fusion. These comprise in particular comparisons between nearest neighbour methods and general (multiple) imputation algorithms (see e.g. Rässler 2002; Lamarche et al. 2020; Meinfelder and Schaller 2022) and more recently between nearest neighbour methods and statistical learning approaches (D'Orazio 2019). Apart from the fact that such comparative studies typically refer to a concrete data situation or to simulated data, the evaluations of the respective fusion procedures investigated are often devoted to the marginal, univariate distribution of a variable to be fused, especially in the context of official statistics (see e.g. Webber and Tonkin 2013; Serafino and Tonkin 2017). However, the preservation of marginal distributions does not give any hint about whether the ultimate aim of a data fusion, the preservation of joint distributions of variables originally not jointly observed, is adequately fulfilled in the fused data file (Kiesl and Rässler 2006). Likewise, certain data fusion methods may perform desirably in specific fusion scenarios, such as when the central assumption is met or when the number of donors is high, but poorly in other scenarios, such as when the basic assumption is violated (Rodgers 1984) or when the number of donors is low (Andridge and Little 2010). Therefore, a comprehensive evaluation of a concrete plethora of possible fusion options in different data contexts is missing in the literature, to our knowledge, especially with regard to the central objective of data fusion, the reproduction of the joint distribution of characteristics originally not jointly observed. In this respect, one innovation of this thesis is to specify and classify selected data fusion scenarios and to investigate potential data fusion methods in terms of their performance in different scenarios.

Second, concerning the aim to establish and evaluate a concrete plethora of data fusion approaches, emerging algorithms from the field of statistical learning could serve as potential data fusion methods as well. However, such approaches are typically dedicated on minimising a certain error term, thus focusing on predicting individual values as accurately as possible rather than gaining to adequately represent distributional properties from a population. However, distribution-based aspects are essential for statistical applications. The vulnerability of statistical learning methods to establish unrealistic distributions (see e.g. Hastie et al. 2009: 312) could also affect a data fusion outcome. Furthermore, statistical learning approaches are only applicable to the univariate, variable-by-variable imputation of variables to be fused, and artificial values instead of previously observed values are imputed in the metric case. In this respect, it is essential to evaluate the performance of such learning-based methods in the data

fusion context and discuss its potentials for different data situations.

Third, the literature further lacks a general extension of statistical learning approaches that could alleviate the problems of adequately reproducing (joint) distributions, provide advantages in imputing real values and include a multivariate imputation solution. Therefore, a new nearest neighbour imputation algorithm based on statistical learning is introduced in this work. We coin this method *Predictive Value Matching* (PVM). An extension of the proposed PVM approach to the simultaneous and multivariate imputation of more than one variable is also developed, whereas the original statistical learning procedures do generally not involve a multivariate imputation solution. In addition, a corresponding and already parallelised `R` function to fuse different (large) datasets using PVM for both the univariate and multivariate cases is provided.

Bringing together the three claims derived from the described research gaps involves the evaluation of potential data fusion algorithms in different data situations including the implementation and evaluation of PVM. This already suggests the specification of a concrete plethora of data fusion approaches to be evaluated alongside PVM. Hence, on the one hand we consider three main archetypes of classical imputation techniques. In this regard, the Distance Hot Deck (DHD) method is a prominent representative of covariate-based nearest neighbour approaches that match observations using only the common variables observed in both datasets by applying a specified distance measure. This type of method represents a quite traditional data fusion approach that appears to be the default algorithm for data fusions in practice. Besides DHD, two other classical imputation methods are considered that reflect prominent imputation methods but are rarely discussed as dedicated data fusion techniques. One is the Regression Model (RM) and therefore a model-based approach, and the other is the semi-parametric Predictive Mean Matching (PMM) algorithm. On the other hand, besides these three classical approaches, two prominent tree-based prediction methods from the field of statistical learning are examined, namely Decision Trees (DT) and Random Forest (RF). In addition, various prediction methods can underlie the general learning-based PVM algorithm. In this thesis, we apply PVM using both tree-based methods (DT and RF) and thus consider the PVM method based on single Decision Trees (PVM-DT) and on Random Forest (PVM-RF).

In order to evaluate these approaches in various data fusion constellations, different data and imputation scenarios are identified and classified along possible consequences for the targeted data fusion. For this purpose, we introduce and distinguish three types of scenarios, namely explicit, implicit and imputation scenarios. In particular, the explicit scenarios are incorporated into the simulations, since these scenarios are expected to have a direct impact on the

performance of the data fusion procedures. It is far beyond the scope of this thesis to cover all potential data fusion constellations. We will therefore focus on selected scenarios as basis for the evaluations. As selected explicit scenarios, we consider the compliance or violation of the Conditional Independence Assumption (CIA) underlying common data fusion applications, as well as different sampling ratios of the datasets to be fused. The implicit scenarios involve the overall sample size of the datasets to be matched and the scale level of the variable that is to be imputed, both of which could limit or suggest the use of certain data fusion algorithms. The imputation scenarios considered refer to the concrete imputation step and comprises the univariate or multivariate imputation solution already mentioned as well as the imputation of previously observed or artificial values. Focusing on these selected scenarios, which cover a broad range of potential data fusion problems, we are able to evaluate the underlying data fusion procedures in specific data contexts.

As basis for the simulations and evaluations, we rely on two concrete but fundamentally different data fusion use cases. Both data fusion applications considered in this thesis represent current data fusion problems in official statistics. One is the data fusion of EU-SILC and HBS. Here, EU-SILC is to be extended with consumption characteristics from HBS in order to jointly analyse the various income information from EU-SILC with the comprehensive consumption details from HBS. This fusion constellation can be seen as a classical scenario, given that samples are to be fused and the larger sample, HBS in this case, serves as donor study that donates its information to the smaller sample, that is, EU-SILC as recipient study that is to be extended with the consumption information (see e.g. Donatiello et al. 2014; Uçar et al. 2016; Albayrak and Masterson 2017; Lamarche et al. 2020). The second data fusion application is the data fusion of the German Tax Statistics (TS), an income register covering all taxpayers in Germany, and the German Microcensus (MC), a 1 % sample of the population. The aim here is to jointly analyse the high-quality income information from the register-based Tax Statistics with socio-demographic variables obtained from the Microcensus in order to improve income modelling. The fundamental difference of this data fusion use case compared to the classical fusion constellation of EU-SILC and HBS is that a full register should be enhanced by variables from a large survey sample, which is still considerably smaller than the register data. Thus, the Tax Statistics serves as recipient study, while the smaller Microcensus represents the donor data file. The two data fusion examples are each used to create simulation designs that allow for an evaluation of the considered data fusion methods along both use cases, with the simulation designs each being further manipulated to incorporate different data scenarios. This is intended to address the identified research gaps and thus to stimulate the research on concrete data fusion

procedures tailored to current debates in official statistics.

The outline of this thesis is as follows: Chapter 2 first contains a general overview of data fusions. Thereby, the underlying definition of data fusion, the central assumption as well as different validation levels are discussed. Also presented in this chapter are different types of scenarios and specific scenarios selected as the basis for the evaluations. Chapters 3 and 4 are dedicated to the description and discussion of the data fusion methods underlying this thesis. Chapter 3 presents and discusses the classical imputation methods, while Chapter 4 is devoted to the alternative statistical learning approaches including PVM. In this respect, Chapters 2, 3, and 4 consist of an integrated literature review of previous methodological research on data fusion in general and on fusion algorithms in particular. Chapters 5 and 6 then evaluate the presented fusion procedures along the selected scenarios, each based on one of the two data fusion use cases. Chapter 5 is dedicated to the data fusion of EU-SILC and HBS, whereas Chapter 6 deals with the data fusion of TS and MC. In the Chapters 5 and 6, the motivation as well as the underlying fusion scenario are first described, before the research and simulation designs are specified and extended by additional scenarios. In addition, Chapter 6 also involves an empirical evaluation of the respective data fusion methods, which seems useful in the fusion context of TS and MC. Then, the results of the simulations and evaluations are presented and discussed. Chapter 7 summarises and classifies the corresponding results. Here, concrete implications for data fusions in official statistics will also be derived for different underlying fusion scenarios.

Chapter 2

Data Fusion: Overview and Scenarios

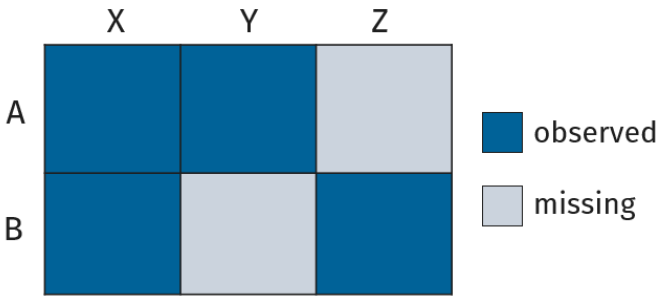
In order to address the aforementioned research objectives, first an overall glance on data fusion is provided in this chapter. This involves introducing data fusion as a specific missing data pattern and specifying the basic notation used throughout this work. It follows a discussion on the central assumption for data fusions, the Conditional Independence Assumption, an overview on validation levels of a data fusion and on different data fusion scenarios in official statistics.

2.1 Data Fusion as A Specific Missing Data Pattern

With regard to the concrete definition of data fusion, we follow the suggestion of Rubin (1986) to consider data fusion as file concatenation. In this respect, data fusion is defined as a specific missing-by-design pattern that occurs through 'stacking' of two or more originally independent data sources (see e.g. Rässler 2002: ch. 4). Figure 2.1 illustrates the respective missing data pattern in case of stacking two originally independent data sources A and B. Here, we obtain a group of variables **X** that were observed in both studies as well as a set of variables **Y** and **Z** that were initially not part of both studies. We thus denote variables that were observed in both studies as **X**, while we further refer to variables required for the analysis that were only observed in study A (but unobserved in study B) as **Y** and, analogously, variables relevant for the analyses that were only part of study B (but unobserved in study A) as **Z** (Meinfelder and Schaller 2022).

Accordingly, in Figure 2.1 the blank parts are missing, illustrating that study A lacks information on the **Z** variables (upper part of the stacked dataset) and B on the **Y** variables (lower part of the stacked dataset). In the context of data fusion, the **X** variables are also referred to as *common*

variables, while the variables **Y** and **Z** that were originally not jointly observed are also called *specific variables*. As the resulting missing data pattern in Figure 2.1 is actually artificially generated (through stacking of two or more datasets), it is often referred to as 'missing-by-design pattern' (see e.g. Koller-Meinfelder 2009: 9). The analytic aim of a data fusion, however, typically refers to the joint distribution of the variables **Y** and **Z** that were originally not jointly observed (see e.g. Kiesl and Rässler 2006).



Source: Meinfelder (2013: 85).

Figure 2.1: Data Fusion as Specific Missing Data Pattern

The missing data mechanism (see Rubin 1976; Little and Rubin 2019) is Missing Completely At Random (MCAR) as long as both studies are samples from the same population (D'Orazio et al. 2006b: 4-7; Meinfelder 2013). For details on different missing data mechanism, see for example Little and Rubin (2019: ch. 1). Since the missing data mechanism is considered ignorable as long as at least Missing At Random (MAR) and distinctness (see Rubin 1976) holds, likelihood-based inference for different quantities of interest would still be possible. For details on ignorability, see for example Rässler (2002: 75-78). However, MCAR can rapidly be challenged in practice. Koller-Meinfelder (2009: 9-10), for example, points out that MCAR no longer applies in the case of unit non-response underlying the data to be matched that is caused by various missing data mechanisms. Furthermore, D'Orazio et al. (2006b: 4) suggest that the data already no longer come from the same population if their drawing takes place at different times. Thus, while the MAR assumption should be treated with caution, it is nevertheless implicitly assumed by the Conditional Independence Assumption (CIA) presented in the following section, since the CIA encases the MAR assumption and therefore also ignorability (Koller-Meinfelder 2009: 10).

The schematic illustration in Figure 2.1 suggests that both missing parts of the stacked dataset could be imputed via data fusion methods. However, in practical implementations, typically only one fusion direction is targeted, that is, either the missing **Y** variables in B or the missing **Z** characteristics in A are imputed, but not both missing parts. We label the study that is to be

enhanced by the missing information as the *recipient study*, while the other study that 'donates' this information is referred to as the *donor study* (see e.g. Gabler 1997; van der Putten et al. 2002; Meinfelder and Schaller 2022).

Throughout this work, especially when presenting the relevant data fusion methods and also with regard to the concrete data fusion use cases of matching EU-SILC with HBS and TS with MC, we consider study A as the recipient study and B as the donor study. It follows that the respective aim is to adequately impute the missing $\mathbf{Z}$ variables of the donor data file B in the recipient study A in order to allow for a joint analysis of $\mathbf{Y}$ and $\mathbf{Z}$ in the resulting integrated microdata file A. Therefore, $\mathbf{Z}$ always represents the variable(s) to be fused (or matched) to the recipient data. Thus, after imputation we obtain an artificial distribution for $\widetilde{\mathbf{Z}}$ within the matched data file.

Note that in addition to this micro-based approach of providing an integrated data file via suited data fusion methods, macro-based approaches also exist. These seek to estimate merely the joint distribution $f(\mathbf{X},\mathbf{Y},\mathbf{Z})$ or at least some properties of it based on the datasets to be matched, rather than generating a complete microdata file (see D'Orazio et al. 2006b: ch. 2.1). However, the subsequent analysis of interest is more plausible and easier to realise in practice with a fused microdata file, which is why micro-based approaches predominate in data fusion research as well as in official statistics (D'Orazio et al. 2006b: 2-3). Consequently, this work focuses on data fusion methods to provide an integrated microdata file.

Concerning the notation, it should be noted that $\mathbf{X} = (X_1,\ldots,X_p)$, $\mathbf{Y} = (Y_1,\ldots,Y_{p_{rec}})$ and $\mathbf{Z} = (Z_1,\ldots,Z_{p_{don}})$ represent matrices, while p indicates the number of common $\mathbf{X}$ variables, p_{rec} is the number of specific $\mathbf{Y}$ variables from the recipient data file and p_{don} reflects the number of specific $\mathbf{Z}$ variables from the donor data file. In this respect, t is used as an index for vectors from the set of the common $\mathbf{X}$ variables and r in turn as an index for a specific variable from $\mathbf{Z}$. Hence, X_t with $t = 1,\ldots,p$ reflect one of the p common $\mathbf{X}$ variables, Z_r with $r = 1,\ldots,p_{don}$ one of the p_{don} specific $\mathbf{Z}$ variables. In addition, $\mathbf{x}_i = (x_{i1},\ldots,x_{ip})$, $\mathbf{y}_i = (y_{i1},\ldots,y_{ip_{rec}})$ and $\mathbf{z}_i = (z_{i1},\ldots,z_{ip_{don}})$ all represent row vectors of an observational unit i. Lower case letters not in bold are used for scalars. An exception is the notation $D_{i,j}$ for distance scalars, as this corresponds to the common notation in the data fusion literature. Furthermore, n_{rec} denotes the sample size of the recipient data file and n_{don} that of the donor data.

2.2 Conditional Independence Assumption (CIA)

With regard to the schematic illustration in Figure 2.1 it is already apparent that identifying assumptions are required for the (typically unobserved) joint distribution of $f(\mathbf{Y},\mathbf{Z})$. In this respect, of considerable importance for data fusions is the *Conditional Independence Assumption* (CIA), which implicitly underlies potential data fusion algorithms. This implicit assumption was first pointed out by Sims (1972) in a comment on Okner (1972) and discussed in more detail by Rodgers (1984).

The CIA states that the specific variables $\mathbf{Y}$ and $\mathbf{Z}$ are independent if conditioned on the common $\mathbf{X}$ variables. In other words, any association between $\mathbf{Y}$ and $\mathbf{Z}$ is therefore a function of $\mathbf{X}$. This implies $f(\mathbf{Y}\,|\,\mathbf{X},\mathbf{Z}) = f(\mathbf{Y}\,|\,\mathbf{X})$ and $f(\mathbf{Z}\,|\,\mathbf{X},\mathbf{Y}) = f(\mathbf{Z}\,|\,\mathbf{X})$, respectively (Meinfelder and Schaller 2022). With regard to the associations between the variables $\mathbf{Y}$ and $\mathbf{Z}$, the artificial variance covariance matrix of $\mathbf{Y}$ and $\mathbf{Z}$ from the fused microdata file is obtained by

$$\widetilde{\mathrm{cov}}(\mathbf{Y},\mathbf{Z}) = \mathrm{cov}(\mathbf{Y},\mathbf{Z}) - \mathrm{E}\left(\mathrm{cov}(\mathbf{Y},\mathbf{Z}\,|\,\mathbf{X})\right), \tag{2.1}$$

while the CIA yields $\mathrm{E}\left(\mathrm{cov}(\mathbf{Y},\mathbf{Z}\,|\,\mathbf{X})\right) = 0$, that is, a mean correlation of 0 if conditioned on $\mathbf{X}$ (Meinfelder 2013: 86). For a detailed derivation of Equation (2.1) see Rässler (2002: 23-24). Thus, from (2.1) it becomes clear that the true association between $\mathbf{Y}$ and $\mathbf{Z}$ is only reproduced adequately in the matched dataset if $\mathbf{Y}$ and $\mathbf{Z}$ are on average uncorrelated given $\mathbf{X}$. As conventional data fusion methods, including the algorithms discussed in this work, establish this conditional independence in the fused microdata file,[1] the CIA is implicitly presumed when conducting a data fusion. Thus, the artificial distribution of $\tilde{f}(\mathbf{Y},\mathbf{Z})$ in the fused dataset is subject to the CIA. In addition, as already indicated in the previous section, the CIA encases the MAR assumption and thus also ignorability (Koller-Meinfelder 2009: 10). However, a scenario where $\mathbf{Y}$ and $\mathbf{Z}$ are truly independent if conditioned on $\mathbf{X}$ neither seems realistic in several practical applications,[2] nor can this assumption be tested on the given data sources to be fused with conventional statistical approaches, as the joint distribution of $\mathbf{Y}$ and $\mathbf{Z}$ is typically unobserved in the respective studies (Kiesl and Rässler 2006).

Hence, various strategies have been introduced to overcome the problems related with the shortcomings of the CIA. Several publications proposed to incorporate auxiliary information about

[1] For an illustration, see Kiesl and Rässler (2005).

[2] One simple example: Suppose we aim to match income (Y) and educational level (Z) based on one common variable, age (X). Here, the CIA would mean that there is no correlation between income and educational level between observation units of the same age, but this is unrealistic.

the joint distribution of **Y** and **Z** (see e.g. Singh et al. 1993; D'Orazio et al. 2006b: ch. 3; Zhang 2015; Fosdick et al. 2016). This insinuates a horizontal overlap between the studies to be fused, since in this case a subset of the observation units yields information on both the **Y** and the **Z** variables. This may circumvent the CIA, but it does not necessarily improve the data fusion outcome. In this respect, findings from Kamgar et al. (2020) show that adding a horizontal overlap can be more prone to other sources of error, such as a misspecified imputation model, compared to the original data fusion pattern illustrated in Figure 2.1. Rather, adding further common **X** variables, which is equivalent to a 'wider' vertical overlap, led to improved fusion results relative to the horizontal overlap and the original data fusion pattern (Kamgar et al. 2020).

Another approach to address the shortcomings of the CIA is to account for the uncertainty of a data fusion result (see e.g. Conti et al. 2012; Endres et al. 2019). In the context of inferential analysis, a general approach to account for the uncertainty caused by the missing information is multiple imputation (MI). This approach is based on imputing missing values several times, which ultimately leads to an adjusted total variance for a given parameter estimator, taking into account the uncertainty caused by the missing information (Rubin 1978, 1987). Rässler (2002) extensively examined multiple imputation approaches in the data fusion context. Rao and Shao (1992), on the other hand, proposed a Jackknife variance estimator for single imputation in order to adjust the variance for missing values that were completed using hot deck imputation. With regard to the MAR assumption included within the CIA, Pfeffermann and Sikov (2011) and Little and Rubin (2019: ch. 15) discussed imputation methods under non-ignorable missing data.

However, such alternative approaches that account for the CIA problems have been less considered for practical implementations in official statistics to provide an integrated microdata file. The predominant analysis objective in official statistics is of a descriptive rather than an inferential nature, which is why we also focus in particular on single imputation throughout this work. Nevertheless, suitable strategies are relevant for official statistics that, despite the problems of the CIA, can induce an appropriate fusion result for single imputation and adequately reproduce the descriptive parameters of interest.

In this respect, an approach that is closely connected to include auxiliary information is to incorporate characteristics that are strongly related to the specific variables **Y** and **Z** (Donatiello et al. 2016). For example, in the context of the data fusion of EU-SILC and HBS, where information on income (**Y**) from EU-SILC and consumption expenditures (**Z**) from HBS are to be matched (see Ch. 5), Donatiello et al. (2016) has suggested including a variable very

similar to income or consumption and present in both datasets as a common variable in the fusion process. In this instance, the rudimentary observed income information from the HBS alongside one of the comprehensive income variables from EU-SILC have been incorporated as a common variable in the matching process. The rationale behind is that high correlations between the common **X** variables and the specific **Y** and **Z** variables weakens the impact of the CIA. In this respect, Kiesl and Rässler (2006) also show that for any pair of specific variables (Y_l, Z_r), the Fréchet-Hoeffding bounds represent a theoretical lower and upper limit for the marginal joint cumulative distribution function $F(Y_l, Z_r)$. Extremely high correlations between the common **X** variables and the specific **Y** and **Z** variables to be fused result in tight Fréchet-Hoeffding bounds for $F(Y_l, Z_r)$, thus indicating that the CIA (at least approximately) holds (Kiesl and Rässler 2006).

In conclusion, the CIA is a strong and often unrealistic assumption, the violation of which typically leads to biased data fusion results. We do not explicitly focus on alternative methods based on auxiliary information or uncertainty measurements that circumvent or adjust for the CIA problems. Instead, we focus on appropriate strategies according to Donatiello et al. (2016) by including suited variables in the fusion process, which supports the performance of some data fusion methods and allows to examine potential effects of the CIA.

2.3 Validation Levels of a Data Fusion

Now that data fusion has been introduced as a specific missing data pattern and the central assumption has been illuminated, this section is dedicated to different validity levels of a data fusion. This will be of importance for the evaluation of the proposed data fusion algorithms in this work. The validity levels of a data fusion has first been introduced by Rässler (2002: 29-32). Taking into account that, in line with Figure 2.1, dataset A represents the recipient file that is to be enhanced by the **Z** variables, while data source B reflects the donor study, the four levels of validity according to Rässler (2002: 29-32) and Kiesl and Rässler (2005, 2006) can be formulated as follows:

1. Preservation of individual values. This implies that the imputed **Z** values equal the true but unknown individual values of the specific **Z** variable(s) in the recipient data file. That is, for any Z_r variable, $\hat{z}_i, \ldots, \hat{z}_{n_{rec}} = z_i, \ldots, z_{n_{rec}}$ for $i = 1, \ldots, n_{rec}$, with n_{rec} being the sample size of the recipient study. One could also refer to this validity level as 'hit rate' (Rässler 2002: 30).

2. Preservation of the joint distributions. Hence, the artificial distribution in the matched data file equals the true joint distribution, that is, $\tilde{f}(\mathbf{X},\mathbf{Y},\mathbf{Z}) = f(\mathbf{X},\mathbf{Y},\mathbf{Z})$.

3. Preservation of correlation structures. Accordingly, the correlations between the variables in the matched data file correspond to the true correlations, that is, $\widetilde{\text{cov}}(\mathbf{X},\mathbf{Y},\mathbf{Z}) = \text{cov}(\mathbf{X},\mathbf{Y},\mathbf{Z})$.

4. Preservation of marginal distributions. This implies that the marginal distributions $f(\mathbf{Z})$ and the joint distributions $f(\mathbf{X},\mathbf{Z})$ within the matched data file equals the distributions already observed in the donor data file, that is, $\tilde{f}(\mathbf{Z}) = f(\mathbf{Z})$ and $\tilde{f}(\mathbf{X},\mathbf{Z}) = f(\mathbf{X},\mathbf{Z})$.

First, it should be noted that the validation levels two and three are subject to the CIA. Hence, only if the CIA (at least approximately) holds, that is, $\mathbf{Y}$ and $\mathbf{Z}$ are conditionally independent given the common $\mathbf{X}$ variables, it can be expected that the joint distributions $f(\mathbf{Y},\mathbf{Z})$ and its correlations resemble reality using conventional data fusion approaches. This illustrates once again the relevance of the CIA to data fusion in practice, but also the relevance of investigating strategies that might be able to address these issues, as well as the need to investigate a wide range of data fusion algorithms in terms of their performance in different CIA-related scenarios. Moreover, only the validity level four, which reflects the preservation of the distributions already observed in the donor data file, can be evaluated in conventional empirical data fusions. In this respect, Meinfelder (2013) proposed a scatter plot that illustrates the correlations $\rho_{\mathbf{X}\tilde{\mathbf{Z}}}$ from the fused data file and $\rho_{\mathbf{XZ}}$ from the donor data, with the R^2 from a corresponding linear regression as quality measure for the fourth level. However, all other validity levels can only be evaluated within a simulation study or by means of auxiliary information, since $\mathbf{Y}$ and $\mathbf{Z}$ were typically not jointly observed (Rässler 2002: 31-32; Kiesl and Rässler 2006).

Furthermore, it should be noted that although the first level of validity, the preservation of individual values, might represent the maximum achievement of a data fusion, it is not considered an appropriate validation criterion. This is because minimising an individual error term does not give any hint on whether the distributions $f(\mathbf{Y},\mathbf{Z})$ and $f(\mathbf{Z})$ are adequately reproduced in the matched data file, which we will discuss in more detail in Section 4.5. However, the preservation of the distributions and its associations is essential for subsequent statistical analyses, for example for computing correlations or regression models, as such analyses based on $f(\mathbf{Y},\mathbf{Z})$ are the typical objective of a data fusion (see e.g. Kiesl and Rässler 2005). Therefore, of interest for subsequent statistical analyses are especially the second and the third validity level. The second validation level is only considered realistic if the common $\mathbf{X}$ variables are extremely

high correlated with the specific $\mathbf{Y}$ and $\mathbf{Z}$ variables, yielding tight Fréchet-Hoeffding bounds (see Kiesl and Rässler 2006). If the second validation level holds, then any statistical analysis based on $f(\mathbf{Y},\mathbf{Z})$ yields valid results. The third validation level implicates that the artificial distribution of $f(\mathbf{Y},\mathbf{Z})$ in the matched data file was generated from a population with at least the same moments and correlation structures as the population of interest (Kiesl and Rässler 2006). Thus, also the third validation level ensures valid results for typical analysis objectives like regression. The fourth validation level can instead be considered as a kind of minimum requirement to data fusions.

As already mentioned, many publications, especially in the context of official statistics, only consider one aspect of the fourth validity level for data fusion (see e.g. Webber and Tonkin 2013; Serafino and Tonkin 2017), namely the preservation of the univariate, marginal distributions of $f(\mathbf{Z})$. However, this is not sufficient in order to evaluate the performance of potential data fusion methods with regard to preserve joint distributions. If other validity levels are considered, this is typically done using the third level (see e.g. D'Orazio 2019; Endres et al. 2019), as the preservation of the correlation structures is essential for subsequent statistical analyses. Consequently, we also basically concentrate on the third validity level as evaluation criteria for the upcoming simulations, but also partly refer to the second and fourth validity levels.

2.4 Selected Data Fusion Scenarios in Official Statistics

The validity levels just presented are particularly relevant for the evaluation of data fusion methods. Beyond that, however, the claim of this work is also to evaluate a variety of concrete fusion algorithms in different data constellations. To our knowledge, this has been neglected in the scientific discussion so far. Therefore, in this section we discuss different data and imputation constellations to ensure a scientific framework for evaluating data fusion methods in different data and imputation contexts. We refer to these different data and imputation constellations as *scenarios*. The scenarios discussed in this work and presented in this section are particularly tailored to data fusion applications in official statistics, especially household surveys.

With regard to these scenarios, we distinguish three types of scenarios: (1) explicit scenarios, (2) implicit scenarios and (3) imputation scenarios. The first two types of scenarios refer to the specific data constellation, while imputation scenarios aim at the concrete implementation of the data fusion. Explicit scenarios describe data constellations that could directly influence the performance of different data fusion methods. Implicit scenarios, on the other hand, are

data constellations that, due to their nature, suggest or exclude the use of certain data fusion methods. Imputation scenarios, in turn, are conditioned by the data fusion method and its concrete implementation. We identify different scenarios based on selected areas of conflict that are inherent for data fusion applications. All three scenario types are defined in the following, presented in detail and discussed using selected scenarios derived from specific areas of conflict.

2.4.1 Explicit Scenarios

We define explicit scenarios as follows: *explicit scenarios are data constellations that have a direct, that is, explicit, impact on the performance of different data fusion methods based on theoretical considerations*. Consequently, these are of central importance in the practical implementation of a data fusion and are of particular focus in the upcoming simulations and evaluations in this work. Two areas of conflict are likely to be of particular relevance, as they give rise to different explicit scenarios and are likely to directly condition the performance of certain data fusion procedures: (1) the fulfilment or violation of the Conditional Independence Assumption (CIA) and (2) the sampling ratio between the donor and recipient data files, which can be high or low. Both areas of conflict and the resulting explicit scenarios will now be discussed, starting with the fulfilment or violation of the CIA.

Compliance or Violation of the CIA

In Section 2.2 we have already discussed the problems and shortcomings of the CIA. Here it also became clear that the fulfilment of the CIA is often unrealistic, but conventional data fusion procedures produce conditional independence in the fused dataset, which is why this assumption is implicitly made when performing a data fusion (see Sims 1972). Conversely, the violation of the CIA suggests that certain data fusion algorithms may experience performance problems and could therefore not be able to adequately reproduce the joint distribution between $\mathbf{Y}$ and $\mathbf{Z}$. In this respect, the fulfilment or violation of the CIA is likely to have a direct, that is, explicit, influence on the performance of certain data fusion methods. Many studies have dealt with such CIA-related problems (see e.g. Rodgers 1984; Barry 1988; Endres et al. 2019). However, these refer in particular to classical imputation algorithms and especially to covariate-based nearest neighbour methods. A competitive comparison of a concrete plethora of different data fusion algorithms, encompassing both classical imputation methods and alternative statistical learning approaches, remains open in the data fusion research, also with regard to their performance

under the approximate fulfilment or violation of the CIA. Accordingly, the first two scenarios derived from the conflict area of the CIA consist of (i) the approximate fulfilment or (ii) violation of the CIA.

Donor-Recipient Ratio

Another area of conflict involving explicit scenarios that should have a direct impact on the performance of data fusion algorithms is the donor-recipient ratio, meaning the sampling ratio between the donor and recipient datasets. Typically, the larger dataset in terms of the absolute number of observation units is used as donor study, while the smaller data source serves as recipient study. This has its justified rationale in order to ensure a sufficiently large donor pool for the imputation process, which is useful for classical imputation methods (see e.g. Andridge and Little 2010; Kleinke 2017) as well as for statistical learning approaches since larger training data tend to produce at least less biased predictions (see e.g. de Mello and Ponti 2018: ch. 2). In turn, however, in the case of a small donor study compared to the recipient dataset, different data fusion methods could suffer performance losses.

Therefore, for methodological reasons, it always seems to make sense to use the larger dataset as the donor data file and the smaller study as the recipient dataset. Consequently, this is often done in empirical data fusions (see e.g. Serafino and Tonkin 2017). In addition to such methodological considerations, however, there may also be content-based motivations, especially in official statistics, to consider the larger dataset as the recipient file. For example, official statistics also contain high-quality administrative or register data in addition to survey data. Such register data have some appealing properties, as they typically cover complete information on the entire distribution of some characteristics of interest on a small-scale level, for example on the taxable income (see Ch. 6). However, such administrative data cover only a very limited range of topics and thus typically lack information on other variables of interest, such as socio-demographic information or information on (political) attitudes or (social) behaviour, which in turn are typically part of surveys. If an analysis of the administrative data is required, taking into account such non-existent (but survey-based) information, then it seems beneficial to match the survey variables to the register data in order to exploit the aforementioned desirable properties of full samples. In this case, however, we consequently have to use the larger sample as recipient study and the smaller sample as donor study, although methodological data fusion aspects favour the larger dataset as donor study.

Therefore, the donor-recipient ratio is another area of conflict that needs to be considered in the

practical implementation of data fusion and will be examined in the subsequent simulations and evaluations. The resulting explicit scenarios are a (i) high donor ratio and a (ii) low donor ratio compared to the number of recipient units.

2.4.2 Implicit Scenarios

We define implicit scenarios as follows: *implicit scenarios are data constellations that suggest or exclude the use of specific data fusion methods due to the nature of the data.* These should have no explicit, that is, direct, influence on the performance of data fusion methods, but indirectly, that is, implicitly, restrict or exclude the availability and applicability of specific data fusion algorithms. Two areas of conflict that give rise to implicit scenarios are particularly relevant here: The (1) size of the underlying datasets to be fused and (2) the scale level of the specific $\mathbf{Z}$ variables to be matched. Both areas of conflict and the resulting implicit scenarios will be discussed in this section, starting with the sample size of the datasets.

Sample Size of the Datasets

This practical conflict between larger and smaller datasets is mainly due to computational aspects. With respect to samples with conventional sample sizes of survey data and thus with a compact number of observation units, we face almost no computational constraints regarding the data fusion of two (or more) survey data, as all available programme implementations, for example within the statistical software `R`, can cope with fusing such data sources. This is satisfactory for most empirical data fusion applications in social and economic sciences. From the perspective of official statistics, however, official data sources include not only conventional surveys with usual sample sizes but also quite extensive surveys with large sample sizes or even full surveys such as administrative or register data. The German Microcensus, for example, is based on a large sample size and covers 1 % of the households in Germany and thus roughly 800,000 observation units, while the Tax Statistics is a full register sample of all taxpayers in Germany and comprises around 40,000,000 observations. It is therefore obvious that some programme routines might encounter problems with computing capacity when fusing such large data sources. Hence, the summed size of the datasets and the associated computing capacities represent a further area of conflict in the practical implementation of data fusions, since not every programme implementation can process such large datasets with conventional computing capacities. Such possible computational limitations may restrict the available data fusion algo-

rithms.

Therefore, the resulting implicit scenarios derived from the conflict area of the sample size of the datasets are the data fusion of samples with conventional sample sizes and of quite large samples. How can we define 'quite large samples'? Logically, this cannot be determined in general, as it depends on the available server and computing capacities, which vary from institution to institution. By 'quite large samples' we mean that sample size, added up from the recipient and donor files, above which the first of the methods presented in Chapters 3 and 4 can no longer be implemented with common programme routines in `R` under the given computing capacities of the underlying `R` server, which comprise 76 cores and 756 gigabyte RAM. In a brief simulation in Section 3.4 we will see that this summed sample size, added from the recipient and donor data files, is about 170,000. Accordingly, we define 'quite large samples' as samples to be fused whose summed sample size is larger than 170,000. The implicit scenarios can thus be made concrete with the data fusion of (i) samples with a summed sample size of no more than 170,000 and (ii) samples with a summed sample size of more than 170,000.

Scale Level of Z

One further area of conflict when conducting a data fusion that also might suggest or restrict the available data fusion methods is the scale level of the specific **Z** variables to be fused. Recall that, in line with Figure 2.1, we assume to have two datasets to be matched, A and B. Data source A comprises the variables **X** and **Y**, while B covers information on **X** and **Z**. We seek to enhance the recipient study A with information on the **Z** variables from the donor study B. However, the potential data fusion methods to fuse studies A and B are also affected on whether the specific **Z** variables to be matched have (ordered) categorical or metric scale level. Most potential data fusion methods work for both, categorical and metric variables, while still a few algorithms originally target only one scale level. Therefore, some potential benefits of a particular data fusion algorithm cannot be exploited if the scale level of the **Z** variables to be fused differs from the scale level required for the method. The resulting implicit scenarios are therefore the imputation of (i) categorical or (ii) metric variables.

2.4.3 Imputation Scenarios

We define imputation scenarios as follows: *imputation scenarios are scenarios that concern the concrete imputation implementation at the meta level*. These scenarios, in contrast to explicit

and implicit scenarios, refer to methodological and technical aspects of the concrete fusion algorithm and can thus be directly governed by the choice of a respective data fusion method. Clearly, each of the fusion methods presented in this work has its own properties and thus implies an independent imputation scenario. Nevertheless, with regard to the concrete fusion implementation, certain overarching properties of a method can be identified that provide appealing properties of the fused dataset in practice. Therefore, the above definition of imputation scenarios refers to the methodological meta level of a data fusion algorithm, that is, to superordinate differences of the fusion algorithms. Imputation scenarios are generally strongly influenced by the purpose of the fused microdata file and the subsequent analyses. Different purposes require different imputation scenarios. This can be illustrated by the areas of conflict discussed in this section, which induce different imputation scenarios. Two areas of conflict are of particular relevance in this respect: (1) the univariate or multivariate imputation solution of the $\mathbf{Z}$ variables in case of $p_{don} > 1$ and (2) the imputation of real and previously observed values or the imputation of artificially generated values. Both areas of conflict will now be discussed on the basis of the resulting scenarios, starting with univariate or multivariate imputation.

Univariate or Multivariate Imputation

If more than one specific $\mathbf{Z}$ variable is to be imputed within the recipient dataset ($p_{don} > 1$), the question arises whether each variable is imputed sequentially in a univariate way, or whether a multivariate imputation solution, that is, the simultaneous imputation of all $\mathbf{Z}$ variables, is aimed for. The former, univariate imputation, has the consequence that the row vector $\mathbf{z}_i$ could remain inconsistent over all specific variables $\mathbf{Z}$. This is because, depending on the imputation model, different donors could be selected for each specific Z_r variable and thus different values imputed. However, this may induce biased associations between the imputed $\mathbf{Z}$ variables within the recipient data file (Little 1988; Meinfelder 2013). In contrast, a multivariate imputation solution ensures that the row vector $\mathbf{z}_i$ remains consistent by identifying an overall donor observation across all $\mathbf{Z}$ variables.

Multivariate imputation is particularly desirable if the subsequent analyses carried out on the fused data file consider multivariate correlations between the $\mathbf{Z}$ variables, for example in the course of multivariate regression analyses. In this respect, it can make sense and be desirable for the subsequent analyses to implement a multivariate imputation solution in the case of $p_{don} > 1$. This aspect can be directly governed by the selection of a certain, suitable fusion method, provided that the corresponding fusion algorithm proves to be promising with regard to the

reproduction of the joint distribution of **Y** and **Z**. Accordingly, the imputation scenarios derived from this consist of (i) univariate or (ii) multivariate imputation in the case of $p_{don} > 1$.

Imputation of Observed or Artificial Values

Another area of conflict that implies different imputation scenarios is the imputation of already observed or artificially generated values for **Z**. This area of conflict exists for metric **Z** variables, while categorical **Z** variables are always imputed with an already observed category in all potential fusion methods. For metric variables, on the other hand, values already observed in the donor data file or artificially generated values are imputed for **Z**, depending on the data fusion algorithm. Artificially generated values typically do not correspond to previously observed values. For continuous **Z** variables, the point probability that the artificially generated imputed value corresponds to a value previously observed in the donor data file would be zero anyway. However, the imputation of already observed values ensures that actually plausible values are complemented in the recipient dataset. This could reduce problems of model misspecification in some imputation variants (Koller-Meinfelder 2009: 32) and could also be helpful in covering outlier observations or extreme values.

Besides, there are also pragmatic and communication-related reasons for imputing real observed values instead of artificial values, especially for the purposes of official statistics and their public mandate. This is because the imputation of real and plausible values induces a more close-to-reality appearance of the fused microdata file. Data that appear to be more close to reality can be communicated more easily in social and political discourse as well as in the media. And one central target of official statistics is to complement and support political and social discourse with appropriately reliable data. Consequently, official statistics in particular are interested in data that comprises plausible values, which is why a fused data file that has previously observed values of the donor dataset is preferable to a matched dataset imputed by means of artificial values. This aspect can also be directly influenced by the choice of an appropriate fusion method, provided that the corresponding data fusion method meets the actual aim of a data fusion, the preservation of joint distributions between the specific variables **Y** and **Z**. The resulting imputation scenarios are thus the imputation of (i) already observed or (ii) artificial values in the case of metric **Z** variables to be fused.

2.4.4 Discussion

The aforementioned data fusion scenarios are summarised and presented compactly in Table 2.1. The corresponding scenarios now allow a targeted differentiation of the performance of data fusion methods with regard to different data and imputation situations. Based on the scenarios described, all data fusion algorithms can now be classified and specified. This is of central importance because different data fusion applications are based on different data and imputation situations from use case to use case. In this respect, the explicit data fusion scenarios in particular supplement the validity levels according to Rässler (2002: 29-32) with concrete criteria and conditions under which certain data fusion algorithms can be applied and yield adequate results with regard to the validity levels of interest. Likewise, for the upcoming comprehensive evaluations of a concrete plethora of data fusion procedures, an evaluation can be conducted along the described scenarios. While the data fusion methods presented in the next two chapters are to be classified directly with regard to the implicit and imputation scenarios, comprehensive evaluations under different data situations are necessary for the explicit scenarios. The simulations and evaluations in Chapters 5 and 6 serve this purpose.

Table 2.1: Overview of Selected Data Fusion Scenarios

Scenario Type	**Conflict Area**	**Scenario**
Explicit scenarios	Conditional Independence Assumption (CIA)	(Appr.) fulfilled
		Violated
	Donor-recipient ratio	High
		Low
Implicit scenarios	Summed sample size of the datasets	$\leq 170,000$
		$> 170,000$
	Scale level of $\mathbf{Z}$	Metric
		Categorical
Imputation scenarios	Imputation solution (in case of $p_{don} > 1$)	Univariate
		Multivariate
	Imputed values (in case of metric $\mathbf{Z}$ variables)	Observed values
		Artificial values

Except for the scenarios of compliance or violation of the CIA, all previously described scenarios shown in Table 2.1 can be identified and determined directly by the concrete use case and by

the purpose of the intended data fusion. For example, with regard to the use cases considered in this work, the data fusion of EU-SILC and HBS as well as of Tax Statistics and Microcensus, the present donor ratio can be determined directly by considering the sample sizes of the respective datasets. Whether, on the other hand, a scenario of compliance or violation of the CIA exists is not directly testable on the basis of the data situation, as was already made clear in Section 2.2. However, corresponding indications could be estimated by means of auxiliary information or a (small) additional study in which $\mathbf{Y}$ and $\mathbf{Z}$ are collected together. Similarly, the correlation structure between $\mathbf{X}$ and the specific $\mathbf{Y}$ and $\mathbf{Z}$ variables provides at least rough evidence regarding a possible fulfilment or violation of the CIA due to the Fréchet-Hoeffding bounds (Kiesl and Rässler 2006). If variables are present in both studies to be matched that are closely related in content to $\mathbf{Y}$ and $\mathbf{Z}$ (Donatiello et al. 2016), the CIA can be tested at least approximately by means of substitute variables (Meinfelder 2013). However, in simulation studies where complete information on the joint distribution of all relevant variable blocks $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ is available, the compliance or violation of the CIA can be assessed directly, which is carried out accordingly in Chapters 5 and 6.

Generally, it should be noted that the scenarios presented are only *selected* data fusion scenarios. Further explicit scenarios could, for example, represent different sampling designs of the data to be fused. If the respective sampling procedures of the underlying data sources differ substantially, this could restrict the comparability of the studies and negatively affect the fusion result achieved by the potential methods. The presence of certain distributional properties, such as a normal distribution or linear relationships between variables of interest, could also be an explicit scenario that could have a direct and differential impact on the performance of fusion methods. Different missing data mechanisms are also likely to have a direct impact on the performance of the fusion algorithms and could therefore be an explicit scenario as well. However, with the CIA, as mentioned in Section 2.1, MAR is implicitly assumed. Thus, the CIA-related scenarios already implicitly include missing data mechanisms. Further imputation scenarios could be, for example, the application of single or multiple imputation. This again emphasises the close connection of imputation scenarios with the analysis objective of the fused microdata file, since multiple imputation is useful when inferential statistical analyses are of interest. In this work, we concentrate on the selected scenarios mentioned above, as these are assumed to be of central importance for data fusion implementations in official statistics.

Furthermore, it should be noted that the scenarios described can be applied specifically to data fusion problems of two independently collected data sources, especially household samples. Further scenarios could be derived from integrated household surveys (see Kamgar et al. 2020).

These already aim a priori to provide comprehensive and reliable data while reducing the response burden on respondents through subsampling and a hierarchical survey structure. In this respect, official statistics in Germany, for example, pursue the integration of official social surveys as a subsample in the Microcensus. Kamgar et al. (2020) investigate potential scenarios for integrated household samples, which include not only the specific missing data pattern of a data fusion, but also scenarios of overlap at the observation level (horizontal) as well as at the characteristic level (vertical). Such scenarios can be directly incorporated into the survey process of the integrated household survey system to increase data quality. In conventional data fusion problems, on the other hand, the incorporation of such scenarios is typically not possible because the studies to be fused were created independently of each other and originally served different purposes. Vertical overlaps are at best possible by chance, while horizontal overlaps are extremely unlikely in samples. Only the data fusion of extremely large datasets or of register data with large samples (see Ch. 6) imply horizontal overlap, although these again cannot be verified due to lack of unique identifiers across the data. Therefore, scenarios of integrated household surveys are not subject of this work. Rather, the scenarios examined here focus on data fusion problems that can be derived from household surveys that were collected independently of each other.

In conclusion, it became clear that data fusions represent a specific missing data pattern. With the Conditional Independence Assumption (CIA), data fusions are based on an assumption that is difficult to verify and often unrealistic. The result of a data fusion can be evaluated using the validity levels according to Rässler (2002: 29-32), whereby special attention will be paid to the third level. In addition, various selected scenarios were introduced and discussed in the context of data fusion. In particular, the explicit scenarios are to be incorporated into the simulations in order to be able to examine corresponding effects in different data constellations. In this respect, each of the data fusion methods presented in the following Chapters 3 and 4 will be classified with regard to the corresponding scenarios and evaluated on the basis of the simulations in Chapters 5 and 6, particularly with regard to the explicit scenarios.

Chapter 3

Classical Imputation Approaches

This chapter begins to set out the methodological framework of this work, initially focusing on classical approaches from imputation research. These classical imputation techniques to complement the recipient file A with the relevant **Z** information can be roughly divided into three types of methods, namely non-parametric, parametric or semi-parametric approaches. Within these three types there are different methodological implementations and thus also a variety of possible concrete fusion algorithms (for an overview see D'Orazio et al. 2006b). To meet the objective in this work of evaluating a concrete plethora of potential data fusion algorithms, one representative archetype for each of the three types is selected and presented in this chapter. This comprise the *Distance Hot Deck* (DHD) method as most general representative for the non-parametric approaches, the *Regression Model* (RM) as archetype of the parametric methods and *Predictive Mean Matching* (PMM) as representative for the semi-parametric algorithms. All three approaches are presented in the following and discussed in Section 3.4. Although the respective methods reflect general imputation approaches for various missing data problems, we discuss them in the specific missing-by-design context of data fusion.

3.1 Distance Hot Deck (DHD)

A quite traditional approach for data fusion is *Distance Hot Deck* (DHD) (see D'Orazio et al. 2006b: ch. 2.4.3), which represents an archetype of the covariate-based and non-parametric nearest neighbour methods. The general idea of DHD is to fuse observation units by means of a minimum distance between the common **X** variables, that is, observations are matched that are maximally similar and minimally different according to the **X** variables observed in

both the recipient and the donor study (see e.g. Rodgers 1984; van der Putten et al. 2002). In the simplest case with $p = 1$ where only a single X variable with continuous scale level is available, the respective distance between an observation i from the recipient file A and a donor observation j from the donor file B is, in accordance with D'Orazio et al. (2006b: 41), given by

$$D_{i,j} = |x_i - x_j|. \tag{3.1}$$

After computing the distances in (3.1), for each recipient observation i its maximally similar donor unit j can be identified, which is defined by the minimum distance according to (3.1). Subsequently, the missing values for each specific variable Z_r from $\mathbf{Z}$ (with $r = 1, \ldots, p_{don}$) are imputed by the real observed values of the closest donor observation. Hence, in case of multivariate $\mathbf{Z}$ variables with $p_{don} > 1$, all missing $\mathbf{Z}$ values of the recipient observation are simultaneously imputed by the respective values for $\mathbf{Z}$ stemming from the closest donor observation.

Note that the distance in (3.1) equals the City-Block or Manhattan distance, which is a variant of the general Minkowski distances (see e.g. Singh et al. 2013). Theoretically, other distance concepts, for example further metrics derived from the general Minkowski distance or the Mahalanobis distance, could also be defined as underlying distance measures.

In practice, however, we usually deal with $p > 1$ common variables, while these common $\mathbf{X}$ variables typically have different scale levels and are both categorical or metric. This also applies to both data fusion use cases considered in this work, the fusion of EU-SILC and HBS on the one hand and of TS and MC on the other. Hence, for computing distances between the recipient and donor observations for $p > 1$ common $\mathbf{X}$ variables, we rely on the distance proposed by Gower (1971) who introduced the following dissimilarity coefficient:

$$D_{i,j} = \frac{\sum_{t=1}^{p} \delta_{ijt} d_{ijt}}{\sum_{t=1}^{p} \delta_{ijt}}. \tag{3.2}$$

δ_{ijt} indicates if comparisons between the values x_{it} from the recipient file and x_{jt} from the donor data are possible for the t-th $\mathbf{X}$ variable, that is, $\delta_{ijt} = 1$ if x_{it} and x_{jt} are both non-missing and $\delta_{ijt} = 0$ if x_{it} or x_{jt} or both are missing. Depending on the scale level of the common variables $X_1, \ldots, X_p$, the following distances d_{ijt} for the tth variable are applied (see also D'Orazio 2021, 2022):

Binary:

$$\begin{aligned} &d_{ijt} = 0 \text{ if } x_{it} = x_{jt} \text{ and} \\ &d_{ijt} = 1 \text{ otherwise;} \\ &d_{ijt} = 1 \text{ if } x_{it} \text{ or } x_{jt} \text{ or both values are missing.} \end{aligned} \tag{3.3}$$

Unordered Categorical:

$$\begin{aligned} &d_{ijt} = 0 \text{ if } x_{it} = x_{jt}; \\ &d_{ijt} = 1 \text{ if } x_{it} \neq x_{jt}; \\ &d_{ijt} = 1 \text{ if } x_{it} \text{ or } x_{jt} \text{ or both values are missing.} \end{aligned} \tag{3.4}$$

Ordered categorical:

For both the recipient and the donor data file the ordered variable X_t is substituted by a position variable O_t representing their natural order (with $1 \leq O_t \leq K$ for K categories). Subsequently, a new variable U_t is defined with $u_{it} = \frac{o_{it}-1}{\max(o_{it})-1}$ and $u_{jt} = \frac{o_{jt}-1}{\max(o_{jt})-1}$, respectively. The new variable U_t is then treated as a metric variable and the final distance is computed analogously to the metric case (Kaufman and Rousseeuw 1990: 29-31, 35-36). (3.5)

Metric:

$d_{ijt} = \frac{|x_{it}-x_{jt}|}{R_t}$ with $R_t = \max(X_t) - \min(X_t)$ reflecting the range of X_t. (3.6)

Note that the original Gower distance does not comprise the computation for ordered categories, but the respective extension has been provided by Kaufman and Rousseeuw (1990: 29-31, 35-36).

Besides the DHD method presented here, there exist further covariate-based nearest neighbour approaches with deviating distance processing (see D'Orazio et al. 2006b: ch. 2.4). One is the Random Hot Deck (RHD) method (D'Orazio et al. 2006b: ch. 2.4.1), which was proposed by Eurostat to match EU-SILC and HBS (Lamarche et al. 2020). The RHD procedure from Eurostat according to Lamarche et al. (2020) is based on categorising all common $\mathbf{X}$ variables in order to define different matching classes, while recipient units are randomly assigned to a donor unit within the same matching class (Lamarche et al. 2020). Thus, the datasets are fused by alleged exact matches with zero distances (due to categorisation). However, this categorisation step yields information losses compared to considering the $\mathbf{X}$ variables at their original scale level. Meinfelder and Schaller (2022) investigated the performance of the proposed RHD approach with respect to preserve associations between the specific $\mathbf{Y}$ and $\mathbf{Z}$ variables. Their primary finding was that Predictive Mean Matching (PMM) (see Sec. 3.3) outperforms RHD in terms of preserving associations between the variables of interest. Another finding was that refining the categorisation of the common $\mathbf{X}$ variables (for example by using 14 age categories instead of 8 age categories) improves the RHD results and, in addition, using DHD via the

Gower distance (where each common variable can remain at its original scale level) brings further improvements (Meinfelder and Schaller 2022). Therefore, we focus on the DHD method with the Gower distance as a proper representative for the traditional covariate-based nearest neighbour methods due to its more general and precise distance processing.

3.2 Regression Model (RM)

Another classical method from the field of imputation research is the *Regression Model* (RM). In contrast to the DHD method, this model-based procedure is subject to distributional assumptions and is therefore also referred to as fully parametric approach. For each specific variable Z_r (with $r = 1, \ldots, p_{don}$) from $\mathbf{Z}$, the idea in the data fusion context is to estimate a regression model of Z_r on the common $\mathbf{X}$ variables within the donor data file and then predict the missing Z_r values within the recipient file using the estimated regression.[1] Depending on the scale level of Z_r, either linear regression for metric Z_r variables or logistic regression for categorical Z_r variables is implemented. Both regression methods are well-known in statistical applications, yet the basics are briefly outlined in this section.

Multiple linear regression, which is used when $p > 1$ variables $X_1, X_2, \ldots, X_p$ are to be included in the model, estimates a set of parameters $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)^T$, which represent the regression coefficients and are typically estimated via the Ordinary Least Square (OLS) method. $\hat{\beta}_0$ reflects the intercept and $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$ serve as slope parameter for the variables $X_1, \ldots, X_p$, indicating their mean impact on the variable Z_r controlling for the remaining independent variables. Based on the estimations for the parameters, it is straightforward to predict the values for Z_r within the recipient data file, which yields the following imputation for each recipient unit:

$$\hat{z}_{ir} = \mathbf{x}_i' \hat{\beta}. \tag{3.7}$$

Here, $\mathbf{x}_i$ reflects the row vector of a recipient unit i over all $\mathbf{X}$ variables and $\hat{\beta}$ represents the vector for the coefficients. Thus, for a given observation unit i from the recipient data, we simply insert its corresponding values of the $\mathbf{x}_i$ characteristics into (3.7) to obtain the imputation $\hat{z}_{ir}$. This approach, however, explicitly assumes a linear relationship between the specific Z_r variable to be matched and the common $\mathbf{X}$ variables. This strongly simplifies the estimation process by just computing different β parameters, as it is far more complicated to model any arbitrary

[1] In the regression context, note that Z_r is also referred to as *dependent variable* and is typically denoted as Y, while the $\mathbf{X}$ variables are also referred to as *independent variables*.

functional form (James et al. 2021: 21-22).

In case of a categorical Z_r variable with K classes $k = 1, 2, \ldots, K$ rather than a metric Z_r variable to be fused, we rely on (multinomial) logistic regression. Instead of modelling the Z_r variable directly as in linear regression, logistic regression is based on the logistic transformation and estimates probabilities $\hat{P}(z_{ir} = k \,|\, \mathbf{x}_i)$ to belong to a certain category k (see Nelder and Wedderburn 1972). This transformation yields

$$\log\left(\frac{\hat{P}(z_{ir} = k \,|\, \mathbf{x}_i)}{\hat{P}(z_{ir} \neq k \,|\, \mathbf{x}_i)}\right) = \mathbf{x}_i'\hat{\boldsymbol{\beta}}, \tag{3.8}$$

which can be rewritten as

$$\hat{P}(z_{ir} = k \,|\, \mathbf{x}_i) = \frac{e^{\mathbf{x}_i'\hat{\boldsymbol{\beta}}}}{1 + e^{\mathbf{x}_i'\hat{\boldsymbol{\beta}}}}. \tag{3.9}$$

In order to obtain probabilities that sum up to one for each category, the multinomial logistic regression following James et al. (2021: 140) involves the model

$$\hat{P}(z_{ir} = k \,|\, \mathbf{x}_i) = \frac{e^{\hat{\beta}_{k0} + \hat{\beta}_{k1}x_{i1} + \hat{\beta}_{k2}x_{i2} + \cdots + \hat{\beta}_{kp}x_{ip}}}{1 + \sum_{l=1}^{K-1} e^{\hat{\beta}_{l0} + \hat{\beta}_{l1}x_{i1} + \hat{\beta}_{l2}x_{i2} + \cdots + \hat{\beta}_{lp}x_{ip}}}. \tag{3.10}$$

For binary Z_r variables with $K = 2$, a single binary logistic regression model is estimated. In the multinomial case with $K > 2$, $K - 1$ logistic regression models are implemented, with the left-out category, typically the Kth class, as baseline. Hence, for each $K - 1$ models, regression coefficients $\hat{\boldsymbol{\beta}}_k = (\hat{\beta}_{k0}, \hat{\beta}_{k1}, \hat{\beta}_{k2}, \ldots, \hat{\beta}_{kp})^T$ are estimated that now reflect the impact on the probabilities $\hat{P}(z_{ir} = k \,|\, \mathbf{x}_i)$ controlling for the remaining independent variables. The estimation process for the parameters in this case corresponds to a Maximum Likelihood (ML) estimation. However, since the ML estimation yields no closed solution in this case, in practice the regression coefficients are estimated iteratively with the Newton-Raphson algorithm via the Taylor approximation (see e.g. Greene 2020: ch. 14).

Based on the $K - 1$ logistic regressions estimated within the donor data file, it is straightforward to predict probabilities for each recipient observation to belong to class $k = 1, \ldots, K - 1$ via Equation (3.10) and to class K via

$$\hat{P}(z_{ir} = K \,|\, \mathbf{x}_i) = \frac{1}{1 + \sum_{l=1}^{K-1} e^{\hat{\beta}_{l0} + \hat{\beta}_{l1}x_{i1} + \hat{\beta}_{l2}x_{i2} + \cdots + \hat{\beta}_{lp}x_{ip}}}. \tag{3.11}$$

As with linear regression, for a given observation i from the recipient data file, we insert its

values for $\mathbf{x}_i$ into the models from (3.10) and (3.11) to obtain probabilities that lay between 0 and 1 for each of the k classes $1,\ldots,K$ of the specific variable Z_r. Based on the resulting probabilities for each category, the respective class k, typically the one with the highest probability, is imputed for the missing Z_r information within the recipient data. Logistic regression predictions do not explicitly assume linearity between the categorical variable Z_r and the common $\mathbf{X}$ variables, but assume linearity between the log odds and the common $\mathbf{X}$ variables. This is already apparent from Equation (3.8) where the left-hand side reflects the log odds (James et al. 2021: 140).

3.3 Predictive Mean Matching (PMM)

This section partly draws from Meinfelder and Schaller (2022). Predictive Mean Matching (PMM) combines the idea of non-parametric nearest neighbour matching like DHD and parametric imputation from regressions. While it is a popular and widely used imputation method in general, and moreover the standard method for metric scaled variables in the `R` package `mice` (van Buuren 2022), PMM has not frequently been discussed as a dedicated data fusion method in previous research. The PMM approach was first introduced by Rubin (1986), while Little (1988) elaborated an extension for the simultaneous imputation of continuous variables. The basic idea is that for each missing value its 'predictive mean' (Little 1988: 291), which is based on linear regression, is compared with the predictive means of all observed values, and the observation with the most similar predictive mean serves as donor record, whose actually observed values are imputed.

Therefore, the PMM algorithm for any specific variable Z_r (with $r = 1,\ldots,p_{don}$) from $\mathbf{Z}$ to be imputed within the recipient data file is as follows: First, an OLS regression of Z_r on $\mathbf{X}$ is calculated based on the donor data file. Subsequently, the predictive means are computed for each observation in the recipient and the donor data file using the previously estimated regression model. The final distance between recipient and donor observations is obtained by

$$D_{i,j} = |\hat{\mu}_i - \hat{\mu}_j| \tag{3.12}$$

with $i = 1,\ldots,n_{rec}$ and $j = 1,\ldots,n_{don}$. $\hat{\mu}_i$ represents the predictive mean of the i-th observation from the recipient file and $\hat{\mu}_j$ reflects the predictive mean of the j-th observation from the donor file. The missing Z_r observation within the recipient data file is then imputed by the real observed Z_r value of the closest donor observation according to (3.12), that is, $\hat{z}_i = z_j$ if

$D_{ij} \leq D_{il} \; \forall \; l = 1, \ldots, n_{don}$ (Little 1988).

For multivariate specific variables $\mathbf{Z} = (Z_1, \ldots, Z_{p_{don}})$ with $p_{don} > 1$, Little (1988) suggests to determine the closest donor observation to a recipient unit via the Mahalanobis distance function:

$$\mathbf{D}_{i,j} = (\hat{\boldsymbol{\mu}}_i - \hat{\boldsymbol{\mu}}_j)^T \mathbf{S}^{-1}_{\mathbf{Z}|\mathbf{X}} (\hat{\boldsymbol{\mu}}_i - \hat{\boldsymbol{\mu}}_j). \tag{3.13}$$

Here, $\mathbf{S}^{-1}_{\mathbf{Z}|\mathbf{X}}$ represents the $p \times p$-dimensional inverse variance-covariance matrix of the residuals from the regression of $\mathbf{Z}$ on $\mathbf{X}$ within the donor data file. This serves as weight matrix, in that distances between recipients and donors are penalised more severely the lower the explanatory power of the common $\mathbf{X}$ variables is with respect to the specific $\mathbf{Z}$ variables. Accordingly, those Z_r specific predictive means are weighted more strongly for the total distance that can be well explained by the common $\mathbf{X}$ variables in the context of the preceding regression model (Koller-Meinfelder 2009: 33-34). Based on the distance in (3.13), an overall closest donor observation for the set of multivariate $\mathbf{Z}$ variables to be fused is determined for each recipient observation. Subsequently, the missing values for $Z_1, \ldots, Z_{p_{don}}$ for each recipient unit are simultaneously imputed by the real observed values from the closest donor observation (Little 1988).

Generally, PMM can be considered as mixed method between the non-parametric and the fully parametric approaches, where the (non-parametric) distance computations are based on previously computed (parametric) predictive means using an underlying linear model. An alternative semi-parametric (or mixed) approach to PMM is for example Rank Hot Deck, which was frequently discussed in publications in official statistics, especially from Eurostat (see Webber and Tonkin 2013; Serafino and Tonkin 2017). However, this approach deviates from PMM in that ranks are used as distance measure instead of the overall distances, thus indicating a less precise distance processing. Similar to the less precise distance processing of Random Hot Deck compared to DHD, which yields a weaker performance (Meinfelder and Schaller 2022), PMM also appears to be a more suitable and promising representative of the semi-parametric methods compared to Rank Hot Deck.

3.4 Discussion

Non-parametric, covariate-based nearest neighbour approaches such as DHD are not subject to any distributional assumptions and are, therefore, robust to different distributional aspects in

the data. This is logically not the case for the parametric regression and the semi-parametric PMM approach. However, their advantage compared to DHD is that the information on the joint distribution between the common **X** variables and the specific **Z** variables to be matched is exploited and represented within the linear model, and therefore also accounted for in the imputation process. Hence, PMM and RM tend to preserve the joint associations between **X** and **Z**, which optimises the fourth validity level. However, if distributional assumptions are violated, for example through a strong non-linear relationship, then PMM and RM might exhibit performance problems. Yet, PMM has been proven robust to model misspecification (see e.g. Koller-Meinfelder 2009: 32; van Buuren 2018: ch. 3.4.1), and Landerman et al. (1997) argues that PMM also yields acceptable imputation results for income as a typically skewed variable.

Comparative studies of classical imputation methods in the context of data fusion are underrepresented in the literature and, if available, typically limited to certain types of methods and to evaluations based on the marginal distributions of **Z**. Studies from Eurostat, for example, investigated the performance of all three types of methods, that is, the non-parametric, semi-parametric and parametric approaches based on marginal distributions (Webber and Tonkin 2013; Serafino and Tonkin 2017). Findings from Webber and Tonkin (2013) and Serafino and Tonkin (2017) suggest that all three types of methods yield basically similar results, while Webber and Tonkin (2013) observed slight performance advantages for the semi-parametric approach. The results from Serafino and Tonkin (2017), on the other hand, indicated slightly improved results for both the non-parametric and the semi-parametric approach compared to the parametric, regression-based method. These studies based on the marginal distributions conclude that the data fusion results were rather satisfying. However, Meinfelder and Schaller (2022) showed that this is not necessarily the case for preserving joint associations between the **Y** and **Z** variables originally not jointly observed, which corresponds to the actual objective of a data fusion. While the results from Meinfelder and Schaller (2022) indicate performance advantages for PMM compared to Random Hot Deck as a non-parametric nearest neighbour approach in terms of joint distributions, their comparative analysis were in turn restricted to PMM and Random Hot Deck only. Meinfelder and Schaller (2022) also partly investigated effects of the explicit scenarios of different donor-recipient ratios, but the results proved to be hardly sensitive in this respect and other scenarios were not considered. However, comprehensive comparative studies covering evaluations on the joint distributions of **Y** and **Z** as well as different scenarios and data situations have so far been neglected in the literature, but are targeted throughout this work.

While the explicit scenarios require corresponding evaluations, the classical imputation meth-

ods just presented can already be classified along the implicit and imputation scenarios. Table 3.1 illustrates the applicability of the data fusion methods discussed in this chapter to the respective implicit and imputation scenarios. With regard to the implicit scenarios, it became clear in Section 3.3 that PMM is an imputation method for metric **Z** variables only, since linear regressions build the basis for nearest neighbour imputation. Yet, it is technically possible to apply PMM to categorical **Z** variables if we treat a categorical **Z** variable as a metric variable within the programme routines. However, if an (ordinal) categorical variable is falsely considered metric, the underlying linear regression model might suffer, justifying methodological reservations against using PMM with (ordinal) categorical variables. In this respect, at least caution is advised when using PMM with categorical variables (pronounced by the exclamation mark in Tab. 3.1). In the simulations in Chapter 6 we will also investigate the performance of PMM for ordered categorical **Z** variables and will then be able to better assess whether the caveats remain. The DHD method as well as the regression approach, in contrast, are applicable to both categorical and metric **Z** variables.

Concerning the sample size of the studies to be fused, the DHD method and its conventional programme implementation in `R`, which is the `StatMatch` package (D'Orazio 2022), indicate restrictions that arise in terms of computational capacities and the size of the datasets. The `StatMatch` package was implemented to fuse different survey data and therefore induces no computational problems in terms of matching data of conventional sample size. In official statistics, however, larger or even full samples are also available for data fusion, and the produced distance matrix required to identify nearest neighbours grows strongly the larger the underlying data sources are. One data fusion use case considered in this work, the data fusion of the Tax Statistics and the Microcensus (see Ch. 6), covers large samples, namely a full sample (TS) as well as a 1 % sample from the population (MC). In this case, the DHD method with the Gower distance via the `NND.hotdeck()` function of `StatMatch` is computationally infeasible with modern server capacities. With the sample sizes of the prepared TS and MC data of $n_{TS} = 12,757,629$ and $n_{MC} = 162,575$ (see Ch. 6), this induces a distance vector with an approximate size of 16 terabytes. However, this is significantly above the usual size of modern servers in the year 2022 (which capture of about three terabytes). Therefore, one restriction to DHD is the size of the datasets to be fused. In additional simulations based on an `R` server with 76 cores and 756 gigabyte RAM, we successively increased the sample size of the studies to be fused, in this case of the Tax Statistics and the Microcensus. The findings indicate that the computing capacities are exceeded from an overall sample size of about $n_{rec} + n_{don} \approx 170,000$ and thus lead to an error message. Note that this is irrespective of the number of common **X** vari-

ables, since the error message also occur with only one X variable, while the required distance vector would again be about 16 terabytes. Therefore, a fusion of larger data by means of the DHD method and the Gower distance using the `NND.hotdeck()` function from the `StatMatch` package is no longer possible. However, due to the ever-increasing demand for data, future server upgrades could mitigate such computational problems.

With regard to the imputation scenarios, a practical advantage of DHD and PMM compared to the Regression Model is the possibility to impute all p_{don} variables simultaneously in case of multivariate $\mathbf{Z}$ variables with $p_{don} > 1$. Thus, the row vector for $\mathbf{z}_i$ stays consistent, which benefits the preservation of the multivariate distributions of all $\mathbf{Z}$ variables and its associations. This is particularly desirable for multivariate analyses, for example for regressions including several imputed $\mathbf{Z}$ variables. The multivariate imputation is implicitly presumed for DHD since the distance is purely covariate-based without considering any form of associations between $\mathbf{X}$ and $\mathbf{Z}$. Therefore, recipient and donor units are matched irrespective of the number of $\mathbf{Z}$ variables to be fused. For PMM, the respective extension to the multivariate imputation has been elaborated by Little (1988), but the separate, univariate imputation of the $\mathbf{Z}$ variables is still possible. In the case of metric scale $\mathbf{Z}$ variables, a further advantage of both DHD and PMM compared to the regression approach is that real observed values from the donor file are imputed within the recipient data instead of imputing artificial values. This has the advantage that it tends to impute a wider range of the entire value range of a given metric Z_r variable. As stated in the previous chapter, note that other potential imputation scenarios like single or multiple imputation are not considered in the scope of this work. However, for example, Meinfelder and Schaller (2022) discussed possible extensions of PMM as a semi-parametric and Random Hot Deck as a non-parametric method to multiple imputation.

The majority of data fusions in practice seem to be based on some form of non-parametric, covariate-based nearest neighbour approaches (see e.g. Koschnick 1995; van der Putten et al. 2002). Hence, the DHD method and other covariate-based imputation variants appear to be the traditional data fusion approach in empirical applications. In the overview of Table 3.1, it is apparent that DHD has some desirable properties in practice, since DHD is flexible with regard to the scale level of $\mathbf{Z}$, involves a multivariate imputation solution and imputes real observed values. In addition to these practical advantages of DHD, Meinfelder and Schaller (2022) mention two other possible reasons for their popularity. On the one hand, covariate-based nearest neighbour methods like DHD could be considered as a softened or 'fuzzy' record linkage. For the logic behind matching observations using minimum distance measures is to fuse 'statistical twins' that are as similar as possible according to the common $\mathbf{X}$ variables

observed in the datasets to be matched. Especially in official statistics, where direct record linkage is sometimes sought but often cannot be applied due to the restrictions mentioned in Section 1.1, the consideration of a method that comes close to the rationale behind direct record linkage seems obvious. On the other hand, the synonymous term 'statistical matching' already insinuates that maximally similar observation units are to be considered as a 'statistical match' for the data fusion process, which in turn makes matching-based nearest neighbour methods appear to be the logical fusion method (Meinfelder and Schaller 2022). However, because there are doubts about such traditional methods with regard to their potential of preserving joint distributions (Meinfelder and Schaller 2022), it seems useful to present and evaluate a concrete plethora of possible methods.

Table 3.1: Implicit and Imputation Scenarios of Classical Approaches

			DHD	RM	PMM
Implicit Scenarios	**Summed sample size**	$\leq 170,000$	✅	✅	✅
		$> 170,000$	⛔	✅	✅
	Scale level of Z	Metric	✅	✅	✅
		Categorical	✅	✅	⚠
Imputation Scenarios	**Imputation solution**	Univariate	⛔	✅	✅
		Multivariate	✅	⛔	✅
	Imputed metric values	Observed values	✅	⛔	✅
		Artificial values	⛔	✅	⛔

To conclude, we now introduced three classical imputation methods that can be applied in the data fusion context. The non-parametric Distance Hot Deck is a prominent form of the covariate-based nearest neighbour methods that appears to be the traditional algorithm for data fusion in practical applications. The regression approach as well as PMM are, however, prominent imputation methods, but seem underrepresented in data fusion implementations. DHD has some desirable properties with regard to the implicit and imputation scenarios, but faces problems with large datasets due to computational restrictions. Comparative studies on the performance of the data fusion procedures are underrepresented and mainly based on the preservation of marginal distributions, which yield no indication on whether the joint distributions are ade-

quately preserved. The next chapter continues with an introduction of the alternative statistical learning approaches for data fusion purposes.

Chapter 4

Statistical Learning Approaches

In this chapter, we first provide a brief overview of statistical learning (SL) rationales and the connection between statistical learning and data fusion. Subsequently, two standard approaches from the field of statistical learning will be introduced: Decision Trees (DT) and Random Forest (RF). We choose these methods as two prominent and widely used archetypes for SL. In addition, SL methods in general and also tree-based methods in particular seem to be of continuously growing consideration in various statistical applications, for example in the field of medical statistics (see e.g. Yang et al. 2009), in social (see e.g. Montgomery and Olivella 2018) or economic sciences (see e.g. Tofan 2015) or also in imputation research (see e.g. Tang and Ishwaran 2017). This underlines the necessity to evaluate the performance of SL methods for conventional data fusion objectives in general and in official statistics in particular. We present the concrete algorithms of Decision Trees and Random Forest in a data fusion framework. Additionally, we introduce a new general statistical learning-based nearest neighbour imputation approach which we call *Predictive Value Matching* (PVM). Finally, the methods presented are discussed, especially with regard to their implications for data fusion purposes and the corresponding scenarios. A first simulation study will also be carried out as part of the discussion.

4.1 Application to Data Fusion

Supervised statistical learning aims at predicting or estimating a certain *output* variable Z (also referred to as *response* or *dependent* variable)[1] by means of *input* variables $\mathbf{X}$ (also denoted as

[1]Note that the common notation for the output or dependent variable is Y. However, we refer to the output variable as Z to already conform to the concept of data fusion and the target of imputing the missing Z variables within the recipient data file.

predictors or *independent* variables). To meet the prediction or estimation target of supervised statistical learning, the crucial task is generally to specify an appropriate functional form f to model the response variable Z based on input variables $\mathbf{X}$ (see Vapnik 1999). Assuming some associations between the output variable Z and the input variables $\mathbf{X}$, this motivation can generally be written as $Z = f(\mathbf{X}) + \varepsilon$, where ε reflects some error term. Decision Trees and Random Forest represent two possible methods to estimate f. Thus, predictions or forecasts for Z on new data with previously unseen observations can be obtained by $\hat{Z} = \hat{f}(\mathbf{X})$, where $\hat{Z}$ reflects the predictions for Z and $\hat{f}$ represents the estimated functional form of f obtained, for example, from a Decision Tree or a Random Forest. The process to estimate f, that is, the concrete computation of a certain statistical learning approach, is referred to as *fitting* or, more in line with the SL wording, *training* (see e.g. James et al. 2021: 16-17, 21, 30). Therefore, to train for example a Random Forest means to implement this method based on a predefined dataset containing the specified output variable Z and some input variables $\mathbf{X}$. This predefined dataset is typically referred to as *training data*.

Note that the regression approach introduced in Section 3.2 can also be considered as a SL method where $f(\mathbf{X})$ is assumed to be linear. However, since the regression approach is a common imputation strategy and, moreover, is associated with general statistical modelling rather than a method of statistical learning, subsuming the regression approach under the classical imputation methods seems more appropriate.

Since supervised statistical learning methods like Decision Trees or Random Forest specify a model to obtain predictions for previously unseen observations, the statistical learning rationale could be useful for data fusion purposes. In the data fusion context, we use the donor dataset as training data to fit a supervised SL model for any specific variable Z_r to be imputed within the recipient dataset, while the common $\mathbf{X}$ variables observed in both datasets to be matched serve as input in the SL framework. Subsequently, the trained Decision Tree or Random Forest is used to impute the missing Z_r information of the recipient data file with the obtained predictions.

It is to be noted, however, that the general target of supervised statistical learning approaches is to minimise a certain error rate of the predictions, thus focusing on the reproduction of individual values rather than on the reproduction of distributions. However, in statistical applications in general, and in data fusion scenarios in particular, the primary aim is to adequately represent (joint) distributions from a population, rather than, for example, optimising the individual prediction accuracy of a certain observation unit i. Therefore, within the simulation studies and the respective evaluations, we particularly focus on the performance of SL methods with regard

to preserve joint distributions of **Y** and **Z** according to the typical data fusion objective. The upcoming sections introduce the respective SL methods considered for data fusion purposes in this work, starting with Decision Trees.

4.2 Decision Trees (DT)

Early implementations of a Decision Tree traces back to Morgan and Sonquist (1963). While their implementations is in the context of social science, Breiman et al. (1984) extensively discussed different general algorithms to fit a Decision Tree both for classification and regression problems in statistics. By now, there exist different routines to build a Decision Tree. We focus in particular on the *Classification And Regression Tree (CART)*[2] algorithms as introduced in Breiman et al. (1984). These are based on *recursive binary partitioning* and are also the basis for common programme routines and `R` packages. In order to introduce the CART algorithm, we start with regression trees and then continue with classification problems before moving on to tree pruning. We present the CART and pruning routines according to Breiman et al. (1984), but in terms of notation we basically follow Hastie et al. (2009) and James et al. (2021).

4.2.1 CART I: Regression Trees

For any specific metric-scaled variable Z_r (with $r = 1, \ldots, p_{don}$) from **Z** to be imputed within the recipient data file, we aim to specify a regression tree to obtain predictions for Z_r. Therefore, in line with the logic of supervised SL, based on the donor data file we seek to specify a regression tree with Z_r representing the output variable and the common **X** characteristics reflecting the input variables, the latter also referred to as *feature space* in the context of Decision Trees. Generally, the CART algorithm aims to automatically partition the feature space of **X** into M distinct and non-overlapping regions $R_1, R_2, \ldots, R_M$ by relevant splitting variables and suited split points (Breiman et al. 1984: ch. 8.4; James et al. 2021: 330). The corresponding procedure to fit a regression tree within the donor data file is explained in detail below.

First, an estimation value $\hat{z}_{R_m}$ for the output variable is required for each region. Since typically the residual sum of squares (RSS), that is, $\sum_{j \in R_m} (z_j - z_{R_m})^2$, is to be minimised, the mean of z_j in a certain region R_m with N_{R_m} observations serves as estimate for z_{R_m} (Breiman et al. 1984:

[2]Note that the main CART competitor is the Iterative Dichotomiser 3 (ID3) (see Quinlan 1986) and its successors C4.5 and C5.0 (see Quinlan 1986), while C5.0 with newer implementations is meanwhile quite similar to CART (Hastie et al. 2009: 312). For a comparison between CART, ID3 and C4.5, see e.g. Singh and Gupta (2014).

230; Hastie et al. 2009: 307):

$$\hat{z}_{R_m} = \frac{1}{N_{R_m}} \sum_{j \in R_m} z_j \ . \tag{4.1}$$

Note that we use the index j because j reflects an observation unit from the donor data file. CART attempts to construct the M regions in order to optimally partition the feature space of $\mathbf{X}$. To find the optimal regions, we need to specify a relevant criterion here as well, and again we rely on the RSS and seek to minimise the sum of RSS across all regions (Breiman et al. 1984: 230; James et al. 2021: 330):

$$\sum_{m=1}^{M} \sum_{j \in R_m} (z_j - \hat{z}_{R_m})^2. \tag{4.2}$$

With regard to the splitting procedure, the formation of the partitioned regions could theoretically take any shape. However, for simplicity and ease of interpretation, CART relies on binary splits, that is, the feature space is divided into high-dimensional rectangles or boxes. Additionally, the problem with multiple splits is that they fragment the feature space too quickly, which can lead to insufficient data on the next split level. Moreover, multiple splits can also be achieved by a series of binary splits, which is why this approach is generally preferred (Hastie et al. 2009: 311).

To construct the M regions, we therefore seek to find the best binary partition with regard to minimise the RSS in (4.2). However, this is computationally infeasible, which is why a *top-down*, *greedy* algorithm known as *recursive binary partitioning* is applied. The starting point of the algorithm is therefore that all observations form a single region and thus represent the top of the tree. Subsequently, the feature space is successively split (*top-down*). At each step of the tree-building process, the algorithm selects the best split at that particular step, rather than considering future splits that lead to a better tree at some subsequent steps (*greedy*) (Breiman et al. 1984: ch. 2; James et al. 2021: 330).

For conducting the recursive binary splitting, we first need to specify a variable X_t of the feature space $\mathbf{X}$ and a corresponding split point s where splitting into the regions $\{\mathbf{X} \,|\, X_t < s\}$ and $\{\mathbf{X} \,|\, X_t \geq s\}$ induces the largest possible reduction of the RSS. Hence, all variables $X_1, X_2, \ldots, X_p$ of the feature space and all possible values of the split point s for each input variable are considered, and for the split of the feature space we select the variable X_t at the split point s where we obtain the lowest RSS. More precisely, for any t and s, a pair of half-planes is defined as

follows (Hastie et al. 2009: 307; James et al. 2021: 331):

$$R_1(t,s) = \{\mathbf{X} | X_t < s\} \quad \text{and} \quad R_2(t,s) = \{\mathbf{X} | X_t \geq s\}. \tag{4.3}$$

Subsequently, we seek the respective value for the splitting variable t and the split point s based on the following condition (Hastie et al. 2009: 307; James et al. 2021: 331):

$$\min_{t,\, s} \left[\sum_{j:\, x_j \in R_1(t,s)} (z_j - \hat{z}_{R_1})^2 + \sum_{j:\, x_j \in R_2(t,s)} (z_j - \hat{z}_{R_2})^2 \right]. \tag{4.4}$$

According to (4.1), $\hat{z}_{R_1}$ and $\hat{z}_{R_2}$ are the mean response for the observations in $R_1(t,s)$ and $R_2(t,s)$, respectively. Thus, the feature space, which initially consisted of one region to which all observations belong, was now divided into two regions R_1 and R_2 by a suited variable X_t at an optimal split point s. This process is then repeated on the two resulting regions, where further splits are conducted according to a suited variable and an optimal split point that minimises the RSS. However, we only split one of the two resulting regions according to minimise the RSS and then get three regions. Again, we further split only one of the three identified regions, which results in four regions. This process successively continues until a stopping criterion is reached, for example when no further substantial reduction of the RSS is possible. Remaining true to the tree concept, we also refer to the final regions as *terminal nodes* or *end nodes* and to regions where further splits are conducted as *decision nodes* or *splitting nodes* (Breiman et al. 1984: ch. 2; James et al. 2021: 331).

4.2.2 CART II: Classification Trees

The CART algorithm for a classification tree in case of a categorical Z_r variable to be imputed is quite similar to that of regression trees. We only need to adjust the estimation $\hat{z}_{R_m}$ for each region as well as the splitting criterion to specify optimal regions. For the former, that is, for $\hat{z}_{R_m}$, instead of using the mean, we now use the mode of z_j in every region R_m with N_{R_m} observations, which means that $\hat{z}_{R_m}$ is simply the response value that occurs with the highest frequency in the resulting region R_m. More precisely, we assume to have a categorical response variable Z_r with k classes, where the proportion of the class k in region R_m with N_{R_m} observations can be written as

$$\hat{p}_{mk} = \frac{1}{N_{R_m}} \sum_{j \in R_m} I(z_j = k), \tag{4.5}$$

and the majority class, that is

$$\hat{z}_{R_m} = \arg\max_k (\hat{p}_{mk}), \tag{4.6}$$

serves as response estimate (Hastie et al. 2009: 309).

With regard to the splitting criteria to specify the M regions, a criterion comparable to the RSS would be the *classification error rate*, which represents the fraction of observations in a certain region that do not belong to the mode class. However, focusing on the classification error rate as relevant criterion for splitting the feature space in optimal regions is not sufficiently sensitive to node purity, which is why in practice two other criteria are preferred. One is the *Gini Index*

$$G = \sum_{k=1}^{K} \hat{p}_{mk} (1 - \hat{p}_{mk}) \tag{4.7}$$

that represents a measure of the total variance across the K classes (Breiman et al. 1984: ch. 4.3.1; James et al. 2021: 336). The other is the *entropy* (James et al. 2021: 336):

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}. \tag{4.8}$$

With regard to the latter, if $0 \leq \hat{p}_{mk} \leq 1$, it follows that $0 \leq -\hat{p}_{mk} \log \hat{p}_{mk}$. Both measures, the Gini index as well as the entropy, are numerically quite similar and, moreover, differentiable, which makes them more suitable for numerical optimisation. Both take smaller values the more the values for $\hat{p}_{mk}$ are closer to 0 or 1, which is why they serve as a measure of node purity. Therefore, as we want to obtain regions that are as pure as possible and thus predominantly contain the same class k, we seek to minimise the sum of the Gini index or the entropy, weighted by the size of the regions (Hastie et al. 2009: 309-310; James et al. 2021: 336).

4.2.3 Tree Pruning

A crucial question when adapting a classification or regression tree is how large and thus complex the tree should be. A tree that is too large and complex may perform well on the predefined training dataset on which the model is fitted, that is, the donor data in the data fusion context, but poor on new data with previously unseen observations, in our case the recipient data. This might lead to *overfitting* to the donor data. To avoid such overfitting, we could adopt a stopping criteria, for example by allowing further splits only if the reduction of RSS (or the Gini or

entropy in the classification case) exceeds a certain threshold. However, a split that results in a seemingly worthless reduction in RSS could be followed by a very good split at the next stage below. Therefore, the preferred approach is to grow a large tree T_0 where the splitting process is stopped when each terminal node has fewer observations than a predefined minimum node size (for example less than five observations per terminal node), and then applying *cost complexity pruning*, also known as *weakest link pruning*, to prune the tree back to a certain *subtree* (Breiman et al. 1984: ch. 3; Hastie et al. 2009: 308). This section briefly describes the principle of cost complexity pruning.

In order to prune back a tree T_0 resulting from recursive binary partition, we first define a subtree $T \subset T_0$ that could represent any pruned tree of T_0. $|T|$ indicates the number of terminal nodes of the tree T. However, there exist an extremely large number of possible subtrees and, thus, considering every possible subtree is too cumbersome. Therefore, we only take a sequence of trees into account that are indexed by a non-negative tuning parameter $\alpha \geq 0$. For each value of α there exists a subtree $T \subset T_0$ such that

$$T_\alpha \;=\; \arg\min \left[\sum_{m=1}^{|T|} \sum_{j: x_j \in R_m} (z_j - \hat{z}_{R_m})^2 + \alpha\,|\,T\,| \right], \tag{4.9}$$

with R_m representing the region of the mth terminal node and $\hat{z}_{R_m}$ reflecting the prediction, that is, the mean of the output values observed in R_m (Breiman et al. 1984: 63, 66; James et al. 2021: 332). For each α, we seek to find a subtree $T_\alpha \subset T_0$ that minimises the expression in (4.9). The tuning parameter α controls the trade-off between the size of the tree and thus its complexity, and its goodness of fit to the donor data. With $\alpha = 0$, the subtree T equals the full tree T_0. With increasing α we get smaller trees T_α and vice versa (Breiman et al. 1984: ch. 3.3; Hastie et al. 2009: 308).

To select an appropriate value of α, K-fold cross-validation (for details see e.g. Fushiki 2011) is applied, typically with $K = 10$ (which is also the default in common programme routines). Thus, the observations of the donor data are randomly divided into K groups, called folds. For each fold $k = 1, \ldots, K$, a large Decision Tree is grown on all but the held-out kth fold with an implemented stopping criteria, typically the minimum node size. Subsequently, we prune the tree by cost complexity pruning in order to obtain a sequence of best subtrees in all but the kth fold, as a function of α, and evaluate the mean squared prediction error on the data in the left-out kth fold, also as a function of α. For each value of α, we average the results of the prediction error and choose the respective α value that leads to the smallest cross-validated error. This α

value serves as the estimate $\hat{\alpha}$, and our final and pruned tree is $T_{\hat{\alpha}}$ (Breiman et al. 1984: ch. 3; James et al. 2021: 333).

The pruned classification or regression tree can now be used to obtain predictions for Z_r within the recipient data file. Hence, we impute the missing Z_r information in the recipient file by means of the resulting predictions $\hat{Z}_r$ from the grown and subsequently pruned Decision Tree.

While pruned Decision Trees are now less prone to overfitting, they still face the problem of the tendency to suffer from high variance, since slightly different donor data could induce very different splits and thus quite varying predictions. Moreover, due to the hierarchical structure of a Decision Tree, there is a high dependence on the accuracy of the first split and if the first split is poor, then the whole tree suffers (see e.g. Hastie et al. 2009: 312). As a high variance is not necessarily desirable for a statistical learning method, research has been conducted to establish tree-based methods that are capable to this variance issue, such as Random Forest, which is presented in the next section.

4.3 Random Forest (RF)

The Random Forest (RF) method was first introduced by Breiman (2001) and is, besides the Decision Tree just presented, a further and popular method from the field of statistical learning. Random Forests are strongly based on Decision Trees, but instead of considering only a single tree, Random Forest involves considering multiple trees. Again, we aim to impute the missing Z_r information within the recipient data file, for which a Random Forest is to be trained based on the donor data with Z_r as output variable and the common $\mathbf{X}$ variables as input. This section introduces the Random Forest algorithm in detail.

The basis for Random Forest is *bagging* (see Breiman 1996), derived from the words *bootstrap aggregation*. Bagging is a general procedure to reduce the variance of a statistical learning method by training a certain method multiple times and averaging its results. This is particularly useful in the context of Decision Trees as they tend to suffer from high variance. The idea of bagging is to draw B bootstrap samples (see Efron 1979; Efron and Tibshirani 1994; Efron 2003) from the donor data, that is, the samples are drawn with replacement. Thus, an observation unit is allowed to occur more than once in a bootstrap sample. The sample size is typically and by default in common programme routines equal to the number of training observations. Within each bootstrap sample, the respective statistical learning method, Decision Tree in our case, is trained. The classification or regression tree within each bootstrap sample is grown deep, that

is, without conducting any tree pruning. For the bth bootstrap sample we then obtain $\hat{f}^b(\mathbf{x}_i)$ as prediction for a recipient unit i. In case of a metric Z_r variable to be fused, averaging all B predictions for each recipient observation yields the final bagging prediction (Breiman 1996; Hastie et al. 2009: 282):

$$\hat{z}_{ir} = \hat{f}_{\text{bag}}(\mathbf{x}_i) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^b(\mathbf{x}_i). \tag{4.10}$$

If Z_r is categorical rather than metric, the majority vote is typically used as the final bagging prediction, that is, the predicted category is the class that occurs most frequently over the B predictions (Breiman 1996).

However, the trees obtained from bagging could in some cases be quite similar. Suppose there is one common variable X_t from $\mathbf{X}$ that is highly correlated with Z_r, while all other common variables are only moderately correlated with Z_r. In this instance, the classification or regression trees of the different bootstrap samples would mostly or always select the highly correlated variable as first split candidate, resulting in quite similar bagged trees. Thus, the bagged trees produce highly correlated predictions, which in turn yield no substantial variance reduction compared to averaging quantities that are less highly correlated (James et al. 2021: 344).

Therefore, motivated by previous works of Ho (1995, 1998), Breiman (2001) proposed only to consider a random subset m of the p input variables $\mathbf{X}$ as split candidates for each time a split is conducted within the bagged trees. Typically, m is chosen to be the (rounded down) square root of the available common $\mathbf{X}$ variables, that is, $m \approx \sqrt{p}$ (James et al. 2021: 343). To consider only a substantially smaller subset of m input variables from the feature space seems counterintuitive at first glance. However, the clever rationale behind is that the bagged trees are less prone to be highly correlated to each other. Hence, Random Forest induces a *decorrelation* of the resulting trees and thus reduces the variance (Hastie et al. 2009: ch. 15).

Analogously to Decision Trees, for data fusion purposes the Random Forest is trained on the donor data and is then used to impute the missing Z_r information within the recipient data file by means of the resulting predictions.

4.4 Predictive Value Matching (PVM)

In the previous sections, Decision Trees and Random Forest have been introduced as pure prediction methods with the intention of using their forecasts to impute the missing $\mathbf{Z}$ variables

within the recipient data file. However, as already mentioned, this might indicate undesirable results in terms of distributional aspects. Furthermore, with regard to the imputation scenarios, it is apparent that DT and RF only involve a univariate imputation solution and impute artificial values for metric $\mathbf{Z}$ variables. Therefore, it seems useful to extend the statistical learning approaches to a more general nearest neighbour imputation method where multivariate imputations are possible and real observed values are imputed based on, for example, Decision Trees and Random Forests. In Section 3.3 we introduced the PMM approach where regression predictions build the basis for nearest neighbour imputation. In this section, the PMM idea of Rubin (1986) and Little (1988) is adopted and transferred into a general, statistical learning-based nearest neighbour approach. We call this approach *Predictive Value Matching* (PVM) in reference to PMM. This section introduces and discusses the proposed PVM approach including an extension to the simultaneous imputation of multivariate $\mathbf{Z}$ variables.

For each specific variable Z_r (with $r = 1, \ldots, p_{don}$) of $\mathbf{Z}$, the PVM algorithm is as follows: First, a statistical learning method, Decision Tree or Random Forest in our case, is trained on the donor data file with Z_r as output variable and the common $\mathbf{X}$ variables as input. Subsequently, predictive values are computed for each observation in both the recipient file and the donor study using the previously trained SL method. Analogously to PMM and Equation (3.12), we obtain the final distance via

$$D_{i,j}^{(PVM)} = |\hat{z}_i^{SL} - \hat{z}_j^{SL}| \tag{4.11}$$

with $\hat{z}_i^{SL}$ representing the predictive value of the i-th observation from the recipient file and $\hat{z}_j^{SL}$ reflecting the predictive value of the j-th observation from the donor study. Again, the missing Z_r value for each observation in the recipient data is then imputed by the real observed value of the closest donor observation according to (4.11), that is, $\hat{z}_i = z_j$ if $D_{ij}^{(PVM)} \leq D_{il}^{(PVM)} \; \forall \; l = 1, \ldots, n_{don}$. If more than one donor has the smallest distance to a recipient unit according to (4.11), one of these donors is randomly selected.

As mentioned in Section 3.3, Little (1988) also proposed an extension of PMM to multivariate $\mathbf{Z} = (Z_1, \ldots, Z_{p_{don}})$ variables with $p_{don} > 1$ using the Mahalanobis distance, with the variance-covariance matrix of the residuals as distance weighting. However, this concept of considering the variance-covariance matrix of residuals is not consistently transferable to other prediction methods like Decision Trees or Random Forest. Therefore, in order to specify a general, overall PVM distance over multivariate $\mathbf{Z}$ variables, we propose to standardise the predicted values for each of the $\mathbf{Z}$ variables both in the recipient and the donor data file and then to sum up the

respective distances.

Hence, the standardisation of the predicted values of a specific variable Z_r for a recipient observation i is obtained by

$$\hat{z}_i^* = \frac{\hat{z}_i^{SL} - \mu}{\sigma}, \tag{4.12}$$

with $\mu = \frac{1}{n_{rec}} \sum_{i=1}^{n_{rec}} \hat{z}_i^{SL}$ reflecting the mean and $\sigma = \sqrt{\frac{1}{n_{rec}} \sum_{i=1}^{n_{rec}} (\hat{z}_i^{SL} - \mu)^2}$ the standard deviation of $\hat{z}_i^{SL}$ within the recipient file. Analogously, the standardisation of the predictions from the donor data for observation j is given by

$$\hat{z}_j^* = \frac{\hat{z}_j^{SL} - \mu}{\sigma}, \tag{4.13}$$

with $\mu = \frac{1}{n_{don}} \sum_{j=1}^{n_{don}} \hat{z}_j^{SL}$ and $\sigma = \sqrt{\frac{1}{n_{don}} \sum_{j=1}^{n_{don}} (\hat{z}_j^{SL} - \mu)^2}$.

The final multivariate PVM distance is subsequently be obtained by summing up the respective distances to obtain a distance over all specific $\mathbf{Z}$ variables:

$$D_{i,j}^{(PVM)} = \sum_{r=1}^{p_{don}} (|\hat{z}_{ir}^* - \hat{z}_{jr}^*|). \tag{4.14}$$

The missing values for $Z_1, \ldots, Z_{p_{don}}$ for each recipient unit are simultaneously imputed by the real observed values from the closest donor observation according to (4.14). As with PMM, the multivariate extension also ensures that the row vector of $\mathbf{z}_i$ stays consistent in the recipient file, which is desirable in some statistical applications. Furthermore, analogously to the univariate case, if more than one donor has the smallest distance to a recipient unit according to (4.14), one of these donors is randomly selected.

Note that if the distance in the univariate case according to (4.11) is zero, then the PVM procedure to impute missing values equals the implementation of Decision Trees and Random Forest in the `R` package `mice` (van Buuren and Groothuis-Oudshoorn 2011). Within `mice`, for each missing value the algorithm involves to identify in which end node a recipient unit will end, and then one donor is drawn randomly from this end node (Doove et al. 2014; van Buuren 2022). This procedure was also suggested in Burgette and Reiter (2010) and van Buuren (2018: ch. 3.5.1) with regard to Decision Trees. Therefore, the PVM procedure using Decision Trees or Random Forest as basis for the intermediate values equals the `mice` procedure in the case of zero distances. Zero distances naturally occur with categorical $\mathbf{Z}$ variables and partly also with

metric variables, namely whenever there are equal intermediate values between the recipient and donor data. For metric **Z** variables, this could be the case when the intermediate values are based on single Decision Trees with few splits, since these tend to frequently provide the same predictions for different observations, namely certain conditional means. However, zero distances for DT in the metric case are unlikely for the multivariate imputation procedure, since in this case the distances are build over more than one variable. The implementation of Decision Trees and Random Forest in `mice`, on the other hand, is only available for univariate imputations (van Buuren 2022), whereas the proposed PVM procedure also provides a multivariate imputation solution. Further note that the procedure of Spaziani et al. (2019), which was introduced for categorical **Y** and **Z** variables, is closely related to PVM. Their suggestion was to predict both the **Y** and **Z** variables within the recipient and the donor data and then to use the predicted categories for **Y** and **Z** as matching classes. However, the procedure of Spaziani et al. (2019) is restricted to the univariate imputation and, moreover, includes no solution for the cases in which one or both of the specific variables **Y** and **Z** are metric.

Thus it has already become clear that PVM, in contrast to PMM, can be applied not only to metric but also to categorical **Z** variables. This implies that the distances from (4.11) and (4.14) yield many zero distances, which is to be expected for categorical variables also in the multivariate case. However, in line with the rationale behind the implementation in `mice` (Doove et al. 2014; van Buuren 2022), this indicates a random draw from a respective end node where the probability for the imputed categorical value equals the relative share within this end node. Consider the following example: We only have one common variable, age in this case. The Decision Tree does only perform one split and partitions the feature space into observations under 50 and greater than or equal to 50. In this case, all observations under 50 form one matching class. However, instead of always imputing the mode category for all observations under 50, which is what a simple Decision Tree prediction would do, the PVM implementation ensures that, besides the mode, also other categories have a probability to be imputed. This is also the case for metric **Z** variables if the single Decision Tree has few splits and would thus, especially in the univariate case, predominantly impute certain conditional mean values (thus also indicating many zero distances). Hence, the imputation process includes a stochastic component and might be able to better map the data-generating process. Furthermore, in the case of many zero distances, the PVM approach can be considered as an extension to Random Hot Deck (D'Orazio et al. 2006b; Lamarche et al. 2020), where the matching classes are not based on the **X** variables only, but on the additional information of the association between **X** and **Z**, which is close to the idea of Spaziani et al. (2019).

An alternative to the proposed PVM procedure for categorical variables could be to use predicted probabilities resulting from a SL method and then to randomly draw a category with the respective probability for each recipient unit, as suggested in D'Orazio (2019). However, their results indicate no substantial improvements compared to the direct imputation of the **Z** variables using the predicted categories resulting from the SL methods (D'Orazio 2019).

We implement the proposed PVM method based on Decision Trees and Random Forests. However, PVM could generally be implemented with any supervised statistical learning approach. Furthermore, PVM is discussed in this work as a dedicated data fusion method, but PVM can also be applied to common imputation problems and thus serve as a general imputation method. The proposed PVM algorithm is applicable to single imputation, which is sufficient for our purposes and the prevailing analysis objectives in official statistics.

In this respect, an already parallelised `R` function for PVM-based data fusion using Decision Trees or Random Forest is provided both for the univariate and the multivariate cases. The parallelisations and fast computations are based on the packages `parallel` (R Core Team 2022a) and `Rfast` (Papadakis et al. 2022). The implementation of the Decision Tree and the Random Forest, on the other hand, are based on the packages `rpart` (Therneau et al. 2022) and `ranger` (Wright and Ziegler 2017; Wright et al. 2022), respectively, while the latter is a fast version of the package `randomForest` (Breiman et al. 2022). We refer to the PVM approach as *PVM-DT* or *PVM-RF* in the upcoming simulations, depending on whether PVM is based on a single Decision Tree or on a Random Forest.

4.5 Discussion

In contrast to the regression approach where a linear relationship between the specific **Z** variables to be fused and the common **X** characteristics is assumed, the presented SL methods are not subject to any distributional assumptions. Hence, these are non-parametric approaches, but unlike the DHD method presented in Section 3.1, they incorporate information on the association of the common **X** variables and the specific **Z** variables in the imputation process (and thus also optimise the fourth validity level). Consistently, Decision Trees should be superior to the regression approach if the relationship is strongly non-linear and inferior if the relationship between **X** and **Z** approximates to a linear form. The advantage of Random Forest is to implement a series of decorrelated Decision Trees and to average its results which typically implicates a superior performance of Random Forest over Decision Trees in terms of predictions, and should

further imply a superior performance over regressions if the distributional assumptions are violated.

However, for data fusion purposes and in statistical applications in general, we are rather interested in distributional aspects instead of pure prediction accuracy. In this respect, a practical disadvantage of Decision Trees is that with smaller trees and thus with a larger number of observations in a terminal node, Decision Trees could predominantly predict a small set of different group-related means or, for classification problems, primarily the mode class while ignoring some other classes with smaller proportions. This tends to the undesirable problem in statistics of producing unrealistic distributions for $\widetilde{\mathbf{Z}}$. The tendency towards smaller trees could, for example, be due to a small sample size, a small number of common $\mathbf{X}$ variables or to poor associations between $\mathbf{X}$ and $\mathbf{Z}$.

If sufficient information in the data nevertheless induces an adequately large and informative Decision Tree, it tends to yield 'peaked' distributions with lack of smoothness (see e.g. Hastie et al. 2009: 312). In case of metric $\mathbf{Z}$ variables, this is because the tree always 'jumps' from end node to end node and thus from mean to mean and therefore reveal the tendency of producing multimodal distributions with many peaks. Such distributions of the fused $\mathbf{Z}$ variables, however, tend to overestimate associations between $\mathbf{Y}$ and $\mathbf{Z}$ due to the overestimated amount of 'peaks' in the distribution. In case of categorical $\mathbf{Z}$ variables, classification trees face the vulnerability of overestimating the proportion of the mode class and underestimating the classes with smaller proportions. This in turn could equally lead to exaggerated gradations between the different class proportions for $\mathbf{Z}$, thus also indicating higher associations between $\mathbf{Y}$ and $\mathbf{Z}$ than appropriate. Since Random Forest averages many decorrelated bagged trees, such problems are less to be expected with Random Forest, but could still occur, albeit possibly in a mitigated manner. An advantage of PVM, on the contrast, could be that the DT- and RF-based predictions only serve as intermediate values for the distance calculation, while finally the real $\mathbf{Z}$ values are imputed. This should further mitigate such distribution-related disadvantages of statistical learning methods and lead, for example, to less overestimation of correlations.

Such potential effects can be investigated by a straightforward simulation study with $k = 1,000$ runs: Suppose we have two datasets A and B to be matched, each with two variables. Data file A contains the variables X and Y, while B comprises the characteristics X and Z (see Fig. 2.1). We thus assume univariate variable blocks. A data fusion is to be conducted by means of the X variable observed in both studies. Hence, in line with common data fusion applications, we aim to impute the missing Z information in the recipient study A with the Z information obtained

from study B, which yields an artificial distribution characterised by $\widetilde{Z}$ within the matched data file. To first consider the case where Z is a metric variable, in each simulation run we draw three variables (X,Y,Z) from a multivariate normal distribution with sample size $n = 1{,}000$, mean $\mu = (0\ 0\ 0)$ and the standardised covariance matrix

$$\boldsymbol{\Sigma} = \boldsymbol{\rho} = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} & \Sigma_{XZ} \\ \Sigma_{YX} & \Sigma_{YY} & \Sigma_{YZ} \\ \Sigma_{ZX} & \Sigma_{ZY} & \Sigma_{ZZ} \end{pmatrix} = \begin{pmatrix} 1 & 0.6 & 0.4 \\ 0.6 & 1 & 0.24 \\ 0.4 & 0.24 & 1 \end{pmatrix}. \tag{4.15}$$

In this case, the CIA is fulfilled as $\Sigma_{XY}\Sigma_{XX}^{-1}\Sigma_{XZ} = 0.6 \cdot 0.4 = 0.24$ (see Rässler 2002: 36). After the sampling step, we randomly split both datasets into a test and training data file, or, to stay in data fusion parlance, a recipient and donor data file. Two thirds of the observation units are randomly assigned to the donor study where the SL model is trained, and the remaining third of the observations are assigned to the recipient dataset on which the methods are evaluated. Hence, in each simulation run, we obtain a sample size of $n_A = 333$ for the recipient file and of $n_B = 667$ for the donor data. We then apply the data fusion procedures for DT, PVM-DT, RF, and PVM-RF as explained in the previous sections and calculate the correlations between Y and Z in the recipient data file. Hence, we obtain $k = 1{,}000$ correlation estimates resulting from each data fusion method. The left-hand side of Table 4.1 shows the Monte Carlo means of the $k = 1{,}000$ resulting correlations for each method as well as the difference to the benchmark value of 0.24.

In order to be able to estimate the respective correlation effects for a categorical Z variable, the above simulation is extended accordingly. Instead of drawing from a normal distribution for each simulation run, three random draws were first made from a multivariate normal distribution to create a suitable database. For this, the first draw was conducted with $\mu_1 = (1\ 1\ 1)$ and $N_1 = 5{,}000$, the second with $\mu_2 = (2\ 2\ 2)$ and $N_2 = 10{,}000$ and the third with $\mu_3 = (3\ 3\ 3)$ and $N_3 = 5{,}000$. We again used the standardised covariance matrix from (4.15) for each of the three random draws. Subsequently, all three draws are then row-binded in order to obtain a full database, which then comprises $N = N_1 + N_2 + N_3 = 20{,}000$ observations. Finally, instead of using the metric Z variable, this variable is transferred into three categories based on the quantiles of Z, thus resulting in a categorical Z variable. The $k = 1{,}000$ simulation draws are now taken from the database comprising $N = 20{,}000$ observations, while the sample size is $n = 1{,}000$ for each random draw. Again, $n_A = 333$ observation units are assigned to the

recipient data and $n_B = 667$ to the donor data. The described procedure is useful here, since it ensures a data situation that is comparable to the metric case. Thus, relying on eta-squared (η^2) as measure for the associations between Y and the categorical Z variable, we obtain a benchmark value of $\eta^2_{YZ} \approx 0.2$. Additionally, the described procedure ensures that the CIA at least approximately holds in order to be comparable to the metric case. The right-hand side of Table 4.1 shows the Monte Carlo means of the $k = 1,000$ resulting $\widehat{\eta}^2_{Y\tilde{Z}}$ estimates for each method as well as the difference to the benchmark value of approximately 0.2.

Table 4.1: Correlation Effects of Statistical Learning Approaches

	Metric Z Variable		**Categorical Z Variable**	
	mean($\widehat{\rho}_{Y\tilde{Z}}$) $= \frac{1}{k}\sum_{i=1}^{k}\widehat{\rho}_i$	Difference to $\rho_{YZ} = 0.24$	mean($\widehat{\eta}^2_{Y\tilde{Z}}$) $= \frac{1}{k}\sum_{i=1}^{k}\widehat{\eta}^2_i$	Difference to $\eta^2_{YZ} \approx 0.2$
DT	0.54	0.30	0.36	0.16
PVM-DT	0.22	0.02	0.09	0.11
RF	0.36	0.12	0.16	0.04
PVM-RF	0.28	0.04	0.16	0.04

With regard to the results for the metric Z variable, it is apparent in Table 4.1 that the mean DT correlation yields 0.54 and thus is on average more than two times higher than the original correlation of 0.24. RF leads to a mean correlation of 0.36 and thus mitigates the effect of exaggerating correlations. We further see that the resulting correlations of PVM-DT and PVM-RF yield additional mitigating effects for both SL approaches. This mitigating effect is particularly evident for Decision Trees, as the correlation is reduced by a substantial value of $0.54 - 0.22 = 0.32$, while the RF correlation decreases from 0.36 to 0.28 and thus comes on average also closer to the true correlation. The PVM procedures produce mean correlations of 0.22 (PVM-DT) and 0.28 (PVM-RF), respectively, and thus come on average quite close to the original correlation of 0.24. In this simulation, such effects are less apparent when a categorical Z variable is to be imputed within the recipient data file. While DT also exaggerates the associations and yields a mean value of 0.36 (left-hand side of Tab. 4.1), the PVM-DT approach underestimates the association, which reduces bias but still does not come quite close to the true relationship. RF again mitigates the exaggerating effects in a reduced manner compared to PVM-DT, but leads to satisfying results in this case, as the mean value of 0.16 is close to 0.2. PVM-RF yields identical results as RF, which could be due to the fact that only one X variable was used in this case and thus the end nodes of RF always include donor observations with a Z

value identical to the prediction.

Overall, it appears from this small simulation study that PVM for metric Z variables copes much better with the imputation process under CIA and produces significantly more unbiased estimates. Clearly, the higher the correlations, the less problems are to be expected for the tree-based methods to exaggerate correlations, as these are limited to the range of $-1 \leq \rho \leq 1$. If the true correlation is for example 0.95, then there is less potential of overestimating this value. With regard to a categorical Z variable, we see mitigating effects for PVM-DT, but these appear to be too strong and lead to underestimated correlations. Concerning RF and PVM-RF, additional simulations with two $\mathbf{X}$ variables also yield quite similar results for both approaches. One explanation could be that there is no mitigating effect for PVM-RF in the categorical case, another explanation could be that this simple example is too short-hand, thus indicating that simulations based on more realistic data is useful, which is the purpose of Chapters 5 and 6.

Beyond this initial simulation study, few studies have considered SL approaches for data fusion. D'Orazio (2019) investigates the performance of several SL procedures in comparison to traditional covariate-based nearest neighbour approaches such as Distance Hot Deck (DHD), also in terms of preserving joint distributions between $\mathbf{Y}$ and $\mathbf{Z}$. However, the research objective was restricted to imputing categorical $\mathbf{Z}$ variables. The findings suggest that no method was able to adequately reproduce the associations between $\mathbf{Y}$ and $\mathbf{Z}$ according to the simulations, which were based on the seventh round of the European Social Survey (ESS). However, the covariate-based nearest neighbour approaches yield better performance compared to Decision Trees or Random Forest with regard to preserve associations between $\mathbf{Y}$ and $\mathbf{Z}$. D'Orazio (2019) also considered the data fusion use case of matching EU-SILC and HBS, again for the scenario with a categorical Z variable, and in this case the covariate-based nearest neighbour methods yield similar results to the tree-based methods in terms of joint distributions (D'Orazio 2019).

As already pointed out, Spaziani et al. (2019) discussed an approach that is closely related to PVM, since the predictions of both specific variables $\mathbf{Y}$ and $\mathbf{Z}$ resulting from a SL method, such as Decision Tree or Random Forest, serve as matching classes. They also compared their data fusion procedure for different SL methods, including Decision Trees and Random Forest, with Random Hot Deck (RHD) based on a proper subset of the common $\mathbf{X}$ variables as matching classes. The findings indicate slight performance advantages when the matching classes are based on predictions from Decision Trees or Random Forest compared to purely covariate-based matching classes with regard to reproduce joint distributions (Spaziani et al. 2019). D'Ambrosio et al. (2012) introduced a tree-based procedure for data fusion with additional components of

Adaptive Boosting (see Freund and Schapire 1997), which was evaluated based on the marginal distributions and compared to the performance of single Decision Trees, the regression approach and a covariate-based nearest neighbour method. The results based on marginal distributions indicate a superior performance of the tree-based methods compared to the classical regression and nearest neighbour approaches, while the extension of trees with boosting further improves the results (D'Ambrosio et al. 2012).

The aforementioned studies do not consider different explicit scenarios that are relevant for the data fusion outcome. Furthermore, these are predominantly restricted to certain implicit scenarios such as the imputation of survey samples with a conventional sample size and of categorical variables, which underlines the need of comprehensive scenario-related evaluations. Concerning the relevant scenarios discussed throughout this work, we are already able to classify the proposed SL approaches with regard to the implicit and imputation scenarios, as was done for the classical imputation methods in Chapter 3. Table 4.2 summarises the respective classifications and also includes the already presented overview for the classical imputation methods.

Table 4.2: Implicit and Imputation Scenarios of Classical and Statistical Learning Approaches

			DHD	**RM**	**PMM**	**DT**	**PVM (DT)**	**RF**	**PVM (RF)**
Implicit Scenarios	**Summed sample size**	$\leq 170,000$	✅	✅	✅	✅	✅	✅	✅
		$> 170,000$	⛔	✅	✅	✅	✅	✅	✅
	Scale level of Z	Metric	✅	✅	✅	✅	✅	✅	✅
		Categorical	✅	✅	⚠	✅	✅	✅	✅
Imputation Scenarios	**Imputation solution**	Univariate	⛔	✅	✅	✅	✅	✅	✅
		Multivariate	✅	⛔	✅	⛔	✅	⛔	✅
	Imputed metric values	Observed values	✅	⛔	✅	⛔	✅	⛔	✅
		Artificial values	⛔	✅	⛔	✅	⛔	✅	⛔

In contrast to the classical approaches, we obtain no restrictions of the SL methods with regard to the implicit scenarios. For the imputation scenarios, it is apparent that DT and RF, as pure prediction methods, only allow univariate imputation and also impute artificial values instead of previously observed values for metric **Z** variables. The advantage of PVM, on the other hand,

is that real and previously observed values are imputed for metric **Z** characteristics and both univariate and multivariate imputation is possible.

To conclude, we have now presented the methodological framework of this thesis. In this chapter, the statistical learning methods DT and RF were introduced and a learning-based nearest neighbour method, PVM, was developed. Thus, the concrete plethora of data fusion approaches investigated in this thesis comprise DHD, RM and PMM as classical imputation algorithms and DT, PVM-DT, RF and PVM-RF as statistical learning approaches. Profound evaluations, especially with regard to the explicit scenarios, will be carried out in the next two chapters in the context of two concrete data fusion use cases in official statistics.

Chapter 5

Data Fusion of EU-SILC and HBS

So far, we discussed the methodological framework of data fusions in general and introduced specific methods to be investigated in different data fusion scenarios in official statistics. In this chapter, we start our simulations and evaluations based on the data fusion objective of statistically matching EU-SILC with the HBS. First, the motivation and the concrete scenario of this specific data fusion use case are outlined. With regard to the simulation design, a description of the underlying database, the Monte Carlo simulation carried out and the manipulations to cover the explicit scenarios in particular follows. Subsequently, we present and discuss the simulation results.

5.1 Motivation and Data Fusion Scenario

This current data fusion use case in official statistics is, following Meinfelder and Schaller (2022), motivated by the 2009 report of the Stiglitz-Sen-Fitoussi 'Commission on the Measurement of Economic Performance and Social Progress' (Stiglitz et al. 2009). The commission recommends that the components 'income', 'consumption' and 'wealth' (ICW) are to be jointly considered in order to acquire new and more in-depth insights on the socio-economic well-being of private households in the European Union and its member states (Stiglitz et al. 2009). However, no official statistics data source exist that covers all three relevant components to an adequate extent. As a first step, Eurostat and several National Statistical Institutes (NSIs) within the EU therefore aim to provide an integrated database containing common information on income and consumption expenditures of private households. This led to various studies on the data fusion of EU-SILC (European Union Statistics on Income and Living Conditions) and

HBS (Household Budget Survey) of Eurostat and many European NSIs (see e.g. D'Orazio et al. 2018; Dalla Chiara et al. 2019; Lamarche et al. 2020; Meinfelder and Schaller 2022).

In this context, EU-SILC provides comprehensive income details (**Y**), while HBS covers extensive information on the consumption expenditures (**Z**) of private households. Hence, by means of fusing EU-SILC and HBS, joint information on income from EU-SILC and consumption expenditures from HBS could thereby be obtained. For this purpose, Eurostat and many NSIs within the EU pursue to enhance the EU-SILC data source with consumption information from HBS (see e.g. Donatiello et al. 2014; Albayrak and Masterson 2017). Accordingly, the aim of this chapter is to evaluate the potential data fusion methods in the data fusion context of matching EU-SILC with HBS, and thus with regard to their potentials to allow for a joint analysis of the income variables from EU-SILC and the consumption expenditures from HBS.

Concerning the concrete data fusion scenario, it is first apparent that EU-SILC represents the recipient study while HBS serves as donor data. Furthermore, depending on the participating countries,[1] HBS contains of about five times more observation units than EU-SILC due to the sample size of both studies (see Eurostat 2015, 2016). Consequently, with regard to the explicit scenarios, the larger sample is used as donor data and the smaller sample as recipient data. This has its rationale from a methodological point of view in order to ensure a sufficiently large donor pool for data fusion purposes. Due to the donor-recipient ratio and the fact that surveys with feasible sample sizes are to be matched, this data fusion use case of EU-SILC and HBS can be considered as a classical data fusion scenario. With regard to the specific **Z** variables to be matched, it should be noted that these have a metric scale level as the aim is to impute consumption expenditures within the EU-SILC data file. Figure 5.1 displays this respective data fusion constellation. Here, the **Y** variables correspond to the income information from EU-SILC and **Z** to the consumption characteristics from HBS.

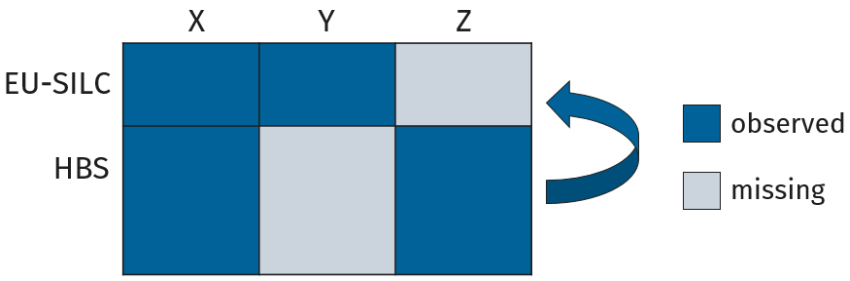


Figure 5.1: Data Fusion Scenario of EU-SILC and HBS

[1]Note that EU-SILC and the HBS are carried out by the NSIs of the EU and coordinated by Eurostat. In Germany, the HBS equals the *Einkommens- und Verbrauchsstichprobe (EVS)*.

Concerning the respective methods investigated so far for the data fusion of EU-SILC and HBS, the concrete data fusion algorithm proposed by Eurostat to match EU-SILC and HBS equals a Random Hot Deck (RHD) approach (Lamarche et al. 2020). In Section 3.4, we already argued why such traditional matching methods might frequently be considered as default method for data fusion in practice. In addition, the consideration of RHD as promising data fusion method is presumably also due to the original focus of their respective studies, which were based on preserving marginal distributions of the $\mathbf{Z}$ variables to be fused (Webber and Tonkin 2013; Serafino and Tonkin 2017). In this respect, we extend the analysis objective to the preservation of joint distributions and evaluate possible data fusion methods based on their potential to reproduce joint associations between different variables.

As the data fusion scenario consists of matching conventional samples where metric $\mathbf{Z}$ variables are to be imputed within the recipient data file, any of the data fusion methods presented in Chapters 3 and 4 can be used as a possible data fusion algorithm and evaluated in the upcoming simulations. In the following section, we provide details on the simulation database and set up the simulation design.

5.2 Simulation Design

In order to evaluate the corresponding data fusion methods in the fusion context of EU-SILC and HBS under different scenarios, we perform a Monte Carlo simulation. To set up the simulation design, we start with an overview on the database and then turn to details on the concrete Monte Carlo study. Since our aim is to draw random samples repeatedly from a real dataset that serves as surrogate population, our study equals a design-based simulation study. The simulation design and the corresponding descriptions in this section closely follow Meinfelder and Schaller (2022). However, the simulation design is further expanded by incorporating different scenarios.

5.2.1 Database

The Monte Carlo (MC) study (see Morris et al. 2019) is based on Scientific Use Files (SUFs) of EU-SILC from the year 2015. This serves the purpose of practical relevance, as official statistics in the European Union, namely Eurostat and the NSIs, also focus on the fusion of the 2015 data files of EU-SILC and HBS. To ensure a sufficiently large data file to draw simulation sam-

ples, EU-SILC data for Germany ($N_{DE} = 12,861$) and France ($N_{FR} = 11,384$) are combined. This dataset serves as surrogate population with a total number of $N = N_{DE} + N_{FR} = 24,245$ observations, and we draw $k = 1,000$ random samples from the respective database. These simulation samples are subsequently split into two data files to obtain substitutes for EU-SILC as the recipient file and HBS as the donor data. Hence, within the simulations, all data are based on EU-SILC in order to assess the 'true' joint distribution of $\mathbf{Y}$ and $\mathbf{Z}$ and their associations. This would not be possible with real data. However, it would be possible to draw random samples from known distributions instead, such as the normal distribution. But it is essential to ensure that the data-generating process is 'neutral' and not based on known distribution families, such as the normal distribution, as this could favour or hinder some data fusion methods and thus prevent a fair evaluation. For example, random samples from a multivariate normal distribution are based on some linear relationships, which could favour the regression approach or PMM. Hence, a simulation study based on empirical data seems more useful (Meinfelder and Schaller 2022).

From the respective database, EU-SILC SUFs for Germany and France from 2015, seven common $\mathbf{X}$ variables are selected that reflect those common variables Eurostat had chosen for the data fusion of EU-SILC and HBS (Leulescu and Agafitei 2013; Lamarche et al. 2020). Thus, we stay as close as possible to the practical application of Eurostat and the NSIs. Note that for data preparations we essentially rely on previous work of Meinfelder and Schaller (2022) and on a `R` code from Eurostat, which Eurostat kindly provided us with. Table 5.1 shows an overview of the respective $p = 7$ common variables $X_1, \ldots, X_7$ that will be considered in the upcoming simulation study, including information on the value range and the measurement level (Meinfelder and Schaller 2022).

The *activity status* (X_1) reflects information on the types of employment (self-employed or non-self-employed, pensioner, unemployed, etc.) (Eurostat 2016: 285). The *population density level* (X_3) yields information on the population density of the residential area (Eurostat 2016: 173), while the *dwelling type* (X_4) reflects the type of accommodation (residential building, flat, etc.) (Eurostat 2016: 173). However, both variables (X_3 and X_4) are empty for Germany (presumably due to confidentiality reasons), which is why we imputed the respective values using the `mice` package (van Buuren 2022) with single imputation. The *tenure status* (X_5) represents combined information on the ownership status of the housing unit (sole owner, tenant, etc.) and on (classified) rental costs incurred (Eurostat 2016: 174, 181). The binary variable *main source of income* (X_6) contains information on (1) income from self-employment or non-self-employment, property, ownership and assets and (2) income from pensions, social benefits and other transfers

(Eurostat 2013: 20, 27-28; Eurostat 2016: 7, 313-316, 322-336). Note that the details described with regard to the common **X** variables closely follow Meinfelder and Schaller (2022).

Table 5.1: Overview of Relevant Variables for Simulations, SILC/ HBS

	Variables	**Range / Scale Level**
X: **Common Variables**	X_1: Activity Status of RP[a]	1 to 5 / categorical
	X_2: Age of RP[a]	acc. X_2 / metric
	X_3: Population Density Level	1 to 3 / categorical
	X_4: Dwelling Type[b]	1 to 4 / categorical
	X_5: Tenure Status	1 to 5 / categorical
	X_6: Main Source of Income[c]	1 to 2 / categorical
	X_7: Income	acc. X_7 / metric
Y: **SILC Variables**	Y_1: Total disposable household income before social transfers including old-age and survivor's benefits	acc. Y_1 / metric
	Y_2: Interest, dividends, profit from capital investments in unincorporated business	acc. Y_2 / metric
Z: **Sub. HBS Variables**	Z_1: Total household gross income	acc. Z_1 / metric
	Z_2: Total disposable household income before social transfers other than old-age and survivor's benefits	acc. Z_2 / metric

[a] RP: 'Reference person' (interviewed person of the household);
[b] Actual range 1 to 5, category 5 is empty;
[c] Here, the missing values also form a category (coded as 9).

Source: EU-SILC SUF DE (2015); EU-SILC SUF FR (2015).

In addition to the common **X** variables just presented, we also need to select specific variables for **Y** and **Z** to (approximately) represent the income variables from EU-SILC (**Y**) and the consumption characteristics from HBS (**Z**) that are originally not jointly observed. In this respect, we select $p_{silc} = p_{hbs} = 2$ substitutes each, that is, $\mathbf{Y} = (Y_1, Y_2)$ and $\mathbf{Z} = (Z_1, Z_2)$, from the underlying database. This is useful due to the fact that besides the univariate imputation with $p_{don} = 1$, some data fusion methods also allow for the simultaneous imputation in a multivariate framework (with $p_{don} > 1$). Thus, for some methods with multivariate extensions, we are able to present simulation results both for the univariate and the multivariate cases (Meinfelder and Schaller 2022).

It is apparent that an exact coverage of the income variables **Y** and the specific consumption characteristics **Z** is only applicable for the income information from EU-SILC, as the database consists of EU-SILC. Due to the fact that both statistical and methodological conclusions are of

interest, however, it is important to ensure a comparable level of measurement for the income and consumption substitute variables. In this respect, it is essential to select metric variables from the underlying database. Hence, for the specific income variables $\mathbf{Y} = (Y_1, Y_2)$ we choose for Y_1 the characteristic 'total disposable household income before social transfers including old-age and survivor's benefits' (Eurostat 2016: 209), while for Y_2 the variable 'interest, dividends, profit from capital investments in unincorporated business' (Eurostat 2016: 214) is selected. For the specific consumption information from the HBS, the variables 'total household gross income' (Eurostat 2016: 207) and 'total disposable household income before social transfers other than old-age and survivor's benefits' (Eurostat 2016: 209) are chosen as substitute variables Z_1 and Z_2. In addition to the common $\mathbf{X}$ variables used in the simulation study, Table 5.1 also displays an overview of the specific variables $\mathbf{Y}$ and $\mathbf{Z}$ (Meinfelder and Schaller 2022).

It is worth noting that the variables X_7, Y_1, Z_1 and Z_2 all reflect household income variables, whereas Y_2 represents capital gains. Detailed information on the corresponding income concepts can be found in Eurostat (2016: 207-211, 214-215). The frequent use of the income variables is due to the data situation of EU-SILC, which lacks information on the consumption expenditures of private households, the imputation of which is the motivation for the intended data fusion of EU-SILC and HBS. However, as already discussed in Section 2.3, an evaluation of the preservation of joint distributions with respect to the specific variables $\mathbf{Y}$ and $\mathbf{Z}$ is typically not possible in real data fusion applications because the joint distribution of $\mathbf{Y}$ and $\mathbf{Z}$ is unknown. Hence, all simulations are based on EU-SILC only where different metric variables are specified as proxies for income and consumption in order to obtain (artificially generated) information on the joint distribution of $\mathbf{Y}$ and $\mathbf{Z}$. The chosen (substitute) variables for $\mathbf{Y}$ and $\mathbf{Z}$ reflect those metric variables where information losses are as small as possible, since many metric variables in the EU-SILC SUFs comprise a high proportion of missing values, are completely blank or have a high proportion of zeros (Meinfelder and Schaller 2022).

As already suggested in Section 2.2, the associations between the common $\mathbf{X}$ variables and the specific $\mathbf{Z}$ variables to be matched could give an indication of the extent to which the CIA might be fulfilled. In this respect, a look at this correlation structure seems to be of interest. Table 5.2 shows the respective matrix of associations. For the associations between the unordered categorical $\mathbf{X}$ variables and $\mathbf{Z}$, we applied the eta-squared (η^2) measure, while for the associations between the metric $\mathbf{X}$ variables and $\mathbf{Z}$, Pearson's correlation coefficient (ρ) is computed. Here we can see that X_7 has an extremely high correlation with the specific variables Z_1 and Z_2, while all other common variables $X_1, \ldots, X_6$ induce low associations. However, since one variable, X_7, is extremely highly correlated with the specific $\mathbf{Z}$ variables, one would expect tighter

Fréchet-Hoeffding bounds and thus a smaller range for the possible correlations between **Y** and **Z**. This indicates a stronger degree of fulfilment of the CIA, which is later also reflected in the benchmark values of the simulations.

Table 5.2: Associations Between **X** and **Z**, SILC/ HBS

	Z_1	Z_2
X_1: Activity Status [a]	0.0895	0.0770
X_2: Age [b]	−0.1081	−0.0208
X_3: Population Density Level [a]	0.0048	0.0024
X_4: Dwelling Type [a]	0.0221	0.0329
X_5: Tenure Status [a]	0.0976	0.1191
X_6: Main Source of Income [a]	0.0913	0.0594
X_7: Income [b]	0.9699	0.9737

[a] Eta-squared (η^2);
[b] Pearson's correlation coefficient (ρ).

Source: EU-SILC SUF DE (2015); EU-SILC SUF FR (2015).

5.2.2 Monte Carlo Study

The Monte Carlo study that is based on the data just described is as follows: First, $k = 1,000$ random samples without replacement (Jackknife) are drawn from the database. For each random draw, the specific missing-by-design pattern of a data fusion (see Fig. 2.1) is generated and the missing information for Z_1 and Z_2 are imputed by means of the concrete plethora of data fusion methods from Chapters 3 and 4 (Meinfelder and Schaller 2022).

Therefore, for each random draw, we create a simulated dataset reflecting EU-SILC consisting of the variables $\mathbf{X} = (X_1, X_2, X_3, X_4, X_5, X_6, X_7)$ and $\mathbf{Y} = (Y_1, Y_2)$ without information on the **Z** variables, and a simulated dataset, representing HBS, consisting of the observed variables $\mathbf{X} = (X_1, X_2, X_3, X_4, X_5, X_6, X_7)$ and $\mathbf{Z} = (Z_1, Z_2)$, which in turn lacks information about the **Y** variables. If we 'stack' both data files, we obtain the specific missing data pattern of the data fusion scenario of EU-SILC and HBS illustrated in Figure 5.1. Within the simulated EU-SILC data file, the missing values for Z_1 and Z_2 are imputed using the proposed data fusion methods in order to obtain an artificial distribution $\widetilde{\mathbf{Z}} = (\widetilde{Z}_1, \widetilde{Z}_2)$. After imputation, the correlations between **Y** and $\widetilde{\mathbf{Z}}$ are computed and compared to the original correlation $\rho_{\mathbf{YZ}}$ resulting from the surrogate population comprising $N = 24,245$ individuals as described in the previous section.

Hence, as already indicated, we focus particularly on the third validity level. However, since the fourth validity level, that is, the preservation of the distributions already observed in the donor data file, is a kind of minimum requirement for data fusion, this aspect is also briefly evaluated. Therefore, we also calculate the correlations between the metric $\mathbf{X}$ variables (X_2 and X_7) and $\widetilde{\mathbf{Z}}$ within the $k = 1,000$ simulated and imputed datasets and compare them with the real correlations $\rho_{\mathbf{XZ}}$ from the database. As indicated in Section 2.2 we apply single imputation ($M = 1$) for all resulting correlations, since we presume that NSIs are predominantly interested in descriptive analyses. The entire process, from sampling to the imputation step to the calculation of the correlations, is performed with $k = 1,000$ simulation runs (Meinfelder and Schaller 2022).

Furthermore, the aim of this work is to investigate the performance of various data fusion methods under different scenarios. In this context, the data fusion of EU-SILC and HBS forms an example of a classic fusion scenario of household surveys with samples and a high donor ratio. Additionally, according to the associations between $\mathbf{X}$ and $\mathbf{Z}$ presented in Table 5.2, it is expected that the CIA could be increasingly fulfilled for this data fusion use case. In order to do justice to the claim of evaluating the data fusion methods in detail in different scenarios, which has been neglected in the literature so far, the incorporation of further scenarios into the simulation is indicated. In this respect, the effects of the fulfilment or violation of the CIA as well as different donor-recipient ratios, that is, the explicit scenarios (see Sec. 2.4), are particularly relevant for the performance of the presented data fusion methods. The corresponding effects of the explicit scenarios are also to be examined within the simulation.

Therefore, on the one hand, we run the simulations both including the X_7 variable and excluding X_7 as a common variable. The rationale is that if we exclude the common variable X_7 from the data fusion process, then we would expect an increased violation of the CIA. In addition, on the other hand, we vary the donor-recipient ratio within the simulations. As mentioned above, the HBS as donor study includes about five times more observations than EU-SILC, depending on the country, that is, $\frac{n_{hbs}}{n_{silc}} \approx 5$. However, to more precisely examine the effects of high and low donor ratios, we use more distinctive sampling ratios of $n_1 = \frac{n_{hbs}}{n_{silc}} = 10$ and $n_2 = \frac{n_{hbs}}{n_{silc}} = 0.1$. We draw the data from the underlying database with an overall sample size of $n = 4,400$, and assign $n_{1_{silc}} = 400$ observations to EU-SILC and $n_{1_{hbs}} = 4,000$ observations to HBS for n_1. Thus, for n_1 the donor pool is ten times higher than the available recipients ($n_{hbs} >> n_{silc}$). For n_2, we also draw an overall sample of $n = 4,400$, but now assign $n_{2_{silc}} = 4,000$ observations to EU-SILC and $n_{2_{hbs}} = 400$ to HBS, which means that the donor pool comprises only 10 % of the number of recipient observations and, therefore, indicates a low donor ratio ($n_{hbs} << n_{silc}$).

With regard to the imputation scenario of univariate or simultaneous, multivariate imputation for $p_{don} > 1$, some of the proposed algorithms, namely DHD (only multivariate), PMM and PVM, comprise solutions for the simultaneous imputation of multivariate $\mathbf{Z}$ variables. We basically present the results for the multivariate PMM and PVM solutions, since we presume that these are preferred in practice due to the desirable fact that the row vector $\mathbf{z}_i$ stays consistent over all variables to be matched. However, we also conduct the simulations with the univariate imputation solutions for PMM and PVM, and the corresponding results are presented in Appendix B.1. In conclusion, we thus perform the described simulation study under the fulfilment and violation of the CIA, with a high and low donor ratio, and for PMM and PVM under the multivariate and univariate imputation solution.

The MC simulation is conducted using `R` (R Core Team 2022b) (version 4.0.2). With regard to the packages, we use `StatMatch` (D'Orazio 2022) for DHD, `rpart` (Therneau et al. 2022) for DT and `ranger` (Wright et al. 2022) for RF. For the Regression Model (RM), the `lm()` and `predict.lm()` functions are used, while for PVM-DT and PVM-RF the created `R` function is applied. For PMM, we rely on `BaBooN` (Meinfelder and Schnapp 2015), since this package already contains a multivariate PMM solution.

5.3 Results

In this section, we present the results of the MC study just described. First, simulation results are presented for the correlations between $\mathbf{Y}$ and $\mathbf{Z}$ and between $\mathbf{X}$ and $\mathbf{Z}$, both for the case where X_7 is included and thus the CIA approximately holds, and for different sample sizes of the recipient and donor studies to reflect different donor-recipient ratios. Subsequently, we discuss results from a simulation study where X_7 is excluded and the CIA is thus strongly violated, also for varying sample sizes. All relevant exact values from this section can be found in Appendix A.1. The results presented in this section and in Appendix A.1 refer to the multivariate PMM and PVM imputation, while Appendix B.1 includes the results for univariate PMM and PVM. Note that the conducted simulations are partly based on previous work by Meinfelder and Schaller (2022). However, Meinfelder and Schaller (2022) only considered the methods RHD (a variant of covariate-based nearest neighbour methods) and PMM, and their simulations comprise scenarios of equal and high donor-recipient ratios. Hence, we extend the simulations to various potential data fusion methods, to CIA-related scenarios and to more distinctive donor-recipient ratios.

5.3.1 CIA Compliance

Correlations Between Y and Z

To evaluate the performance of the data fusion algorithms with regard to preserve correlations between the specific variables **Y** and **Z**, which equals the third validity level of a data fusion, we consider the resulting 'true' correlations from the underlying database as benchmark. In order to assess effects of the CIA, we additionally consider the respective correlations that would result if the CIA holds, that is, the theoretical correlations between **Y** and **Z** if conditionally independent given **X**. Following Rässler (2002: 36), the artificial covariance matrix $\widetilde{\Sigma}_{\mathbf{YZ}}$ for multivariate normal data assuming that the CIA holds is given by

$$\widetilde{\Sigma}_{\mathbf{YZ}} = \widetilde{\text{cov}}(\mathbf{Y},\mathbf{Z}) = \Sigma_{\mathbf{YX}}\Sigma_{\mathbf{XX}}^{-1}\Sigma_{\mathbf{XZ}}. \tag{5.1}$$

Note that in our case Equation (5.1) is subject to the assumption of partial uncorrelation due to the absence of a normal distribution for the variables of interest. Hence, we assume that $\Sigma_{\mathbf{YZ}} - \Sigma_{\mathbf{YX}}\Sigma_{\mathbf{XX}}^{-1}\Sigma_{\mathbf{XZ}} = 0$ holds. However, there is no trivial way to obtain correlations under CIA for non-normal data, which is why we use Equation (5.1) under the assumption of partial uncorrelation as proxy for the correlations under CIA. The benchmark correlations between $\mathbf{Y} = (Y_1, Y_2)$ and $\mathbf{Z} = (Z_1, Z_2)$ and the correlations assuming conditional independence obtained by Equation (5.1) are displayed in Table 5.3. Here we see that the correlations between Y_1 and Z_1 as well as between Y_1 and Z_2 are quite high (0.87 and 0.85), whereas medium correlations are observed between Y_2 and Z_1 as well as between Y_2 and Z_2 (0.44 and 0.48). It is further apparent that the correlations under CIA are quite close to the respective benchmark values of the true correlations. This indicates that the CIA approximately holds, which was already obvious when looking at the correlation structure between **X** and **Z** (see Tab. 5.2).

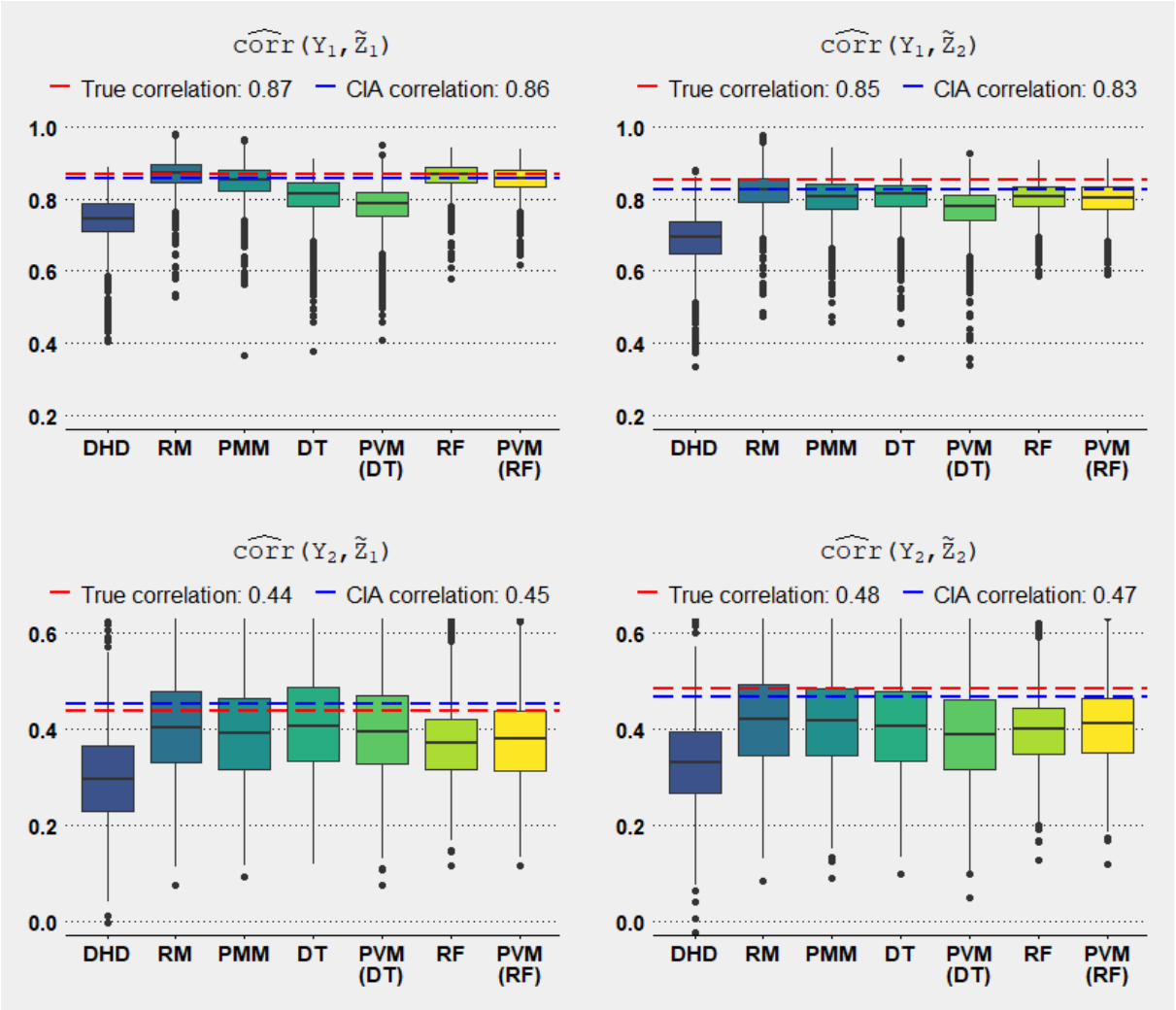
Table 5.3: Benchmark Parameters for $\rho_{\mathbf{YZ}}$ under CIA Compliance

	$\text{corr}(Y_1, Z_1)$	$\text{corr}(Y_1, Z_2)$	$\text{corr}(Y_2, Z_1)$	$\text{corr}(Y_2, Z_2)$
True correlation	0.8678	0.8536	0.4361	0.4831
CIA correlation	0.8576	0.8266	0.4528	0.4651

Source: EU-SILC SUF DE (2015); EU-SILC SUF FR (2015).

It is now of crucial interest to what extent the respective fusion algorithms can reproduce the original correlations from Table 5.3. In this respect, Figure 5.2 illustrates the Monte Carlo

distributions for all $k = 1{,}000$ resulting correlations for each method under n_1, that is, with a high proportion of available donor observations ($n_{silc} = 400$ and $n_{hbs} = 4{,}000$). In addition, the red line corresponds to the original correlation as correlation benchmark, whereas the blue line marks the theoretical CIA correlation.



Source: EU-SILC SUF DE (2015); EU-SILC SUF FR (2015).

Figure 5.2: MC distributions for $\widehat{\rho}_{\mathbf{Y}\widetilde{\mathbf{Z}}}$ with n_1 under CIA Compliance

For high original correlations between $\mathbf{Y}$ and $\mathbf{Z}$ under n_1 (upper part of Fig. 5.2), it is first apparent that the most conventional and traditional data fusion method, the covariate-based and non-parametric DHD method, is not able to adequately reproduce the original correlations of 0.87 and 0.85, respectively. The mean correlations for each method resulting from all $k = 1{,}000$ simulation runs are displayed in Table A.1. Here, it is apparent that DHD produce mean correlations of 0.74 for $\widehat{\text{corr}}(Y_1, \widetilde{Z}_1)$ and 0.69 for $\widehat{\text{corr}}(Y_1, \widetilde{Z}_2)$, which deviate by $0.87 - 0.74 = 0.13$ and $0.85 - 0.69 = 0.16$ from the original correlations. All other methods tend to yield acceptable results in the case of high original correlations, while the best performance can be

observed for RM, RF and PVM-RF, which basically yield quite similar results. We consider the Root Mean Squared Error (RMSE)[2] as a suitable diagnostic of the overall performance, as it combines the concepts of bias and variance and thus takes into account the fact that neither low bias and high variance, nor high bias and low variance are desirable. The RMSE values for n_1 are illustrated in Figure 5.3, while the exact values are displayed in Table A.2. Here, it is also apparent that RM, RF and PVM-RF yields the best overall performance according to the RMSE for n_1 and high original correlations.

Source: EU-SILC SUF DE (2015); EU-SILC SUF FR (2015).

Figure 5.3: RMSE of $\widehat{\rho}_{\mathbf{Y}\tilde{\mathbf{Z}}}$ with n_1 under CIA Compliance

DT slightly underestimates the high association between Y_1 and Z_1. This could be due to the fact that DT predominantly predicts a small number of conditional means for Z_1, with the number of different means in the underlying simulations often being in the single digits. As a result, a large number of recipient observations getting the same imputation. Hence, for extremely

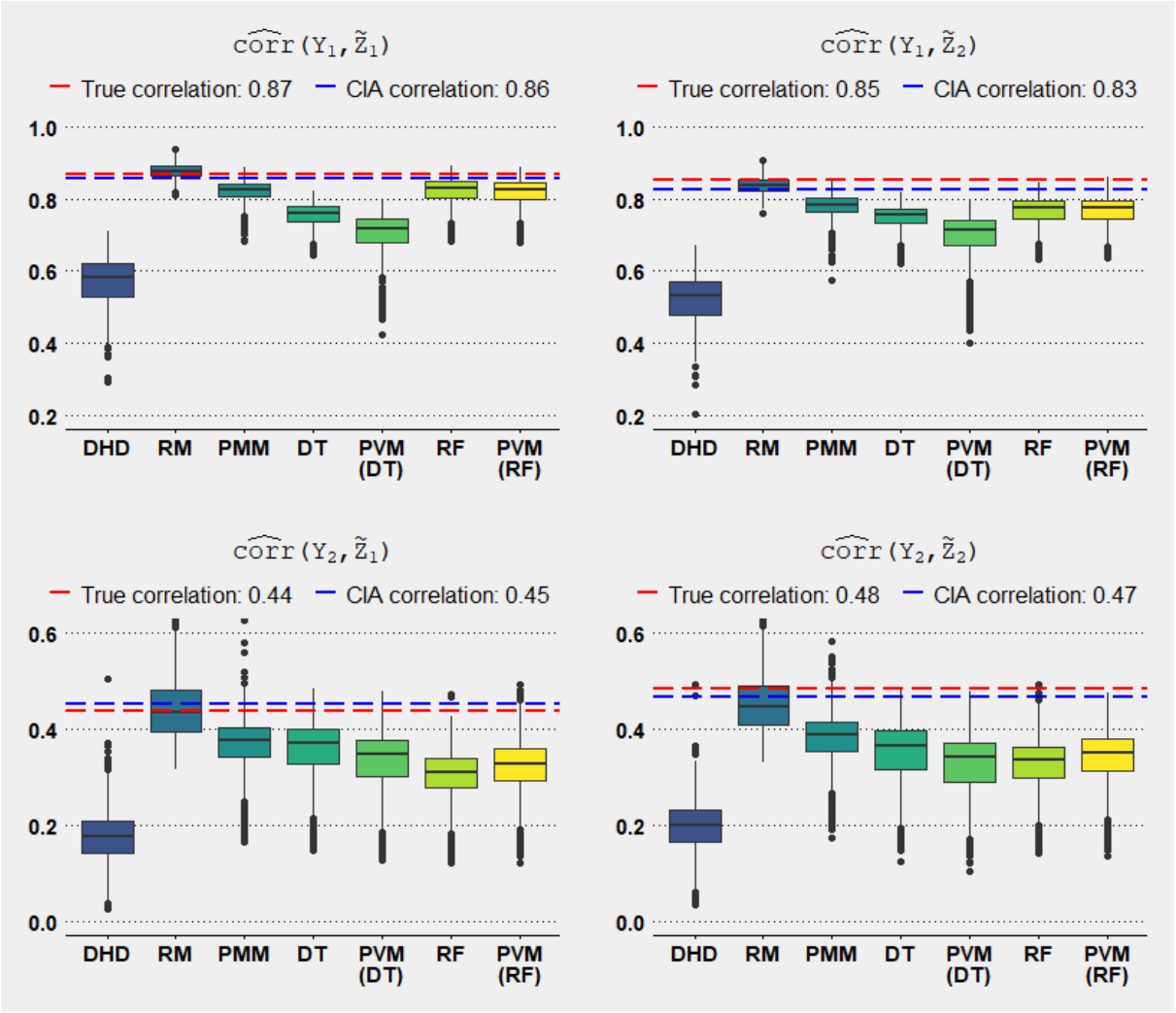
[2] $\text{RMSE}(\widehat{\rho}) = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(\widehat{\rho}_i - \rho)^2}$

high linear correlations between at least one X variable and $\mathbf{Z}$ as well as between $\mathbf{Y}$ and $\mathbf{Z}$, the DT exaggeration effects originally assumed and discussed in Section 4.5 seem to have slight opposite effects. This is because if the imputed $\mathbf{Z}$ variable only comprises a small number of different mean values and the original correlations are quite high, then part of the association is lost after imputation, which leads to slightly lower correlations after imputation. Such potential problems do not occur for RF, since RF produces a series of (decorrelated) trees and averages the results. Hence, it is not expected that RF imputes the same $\mathbf{Z}$ value for more than one recipient unit, and additional evaluations of the simulation show that in almost all simulation runs, RF yields different $\mathbf{Z}$ imputations for all recipient observations (EU-SILC SUF DE 2015; EU-SILC SUF FR 2015).

In addition, PVM-DT yields further underestimations for both $\widehat{\text{corr}}(Y_1, \widetilde{Z}_1)$ and $\widehat{\text{corr}}(Y_1, \widetilde{Z}_2)$ and thus higher RMSEs compared to RM, PMM, DT, RF and PVM-RF. This is likely due to the fact that the few conditional means of DT imply too undifferentiated matching classes and thus imply more imprecise links between recipient and donor observations. However, such matching classes guided by the predicted values appear to be more accurate than identifying nearest neighbours based solely on the common $\mathbf{X}$ variables, as in DHD. Consequently, PVM-DT gives better results than DHD, but the performance of PVM-DT here is still somewhat inferior to that of RM, PMM, DT, RF and PVM-RF.

For medium original correlations of 0.44 and 0.48, it is again apparent that DHD is not able to reproduce the benchmark associations (lower parts of Fig. 5.2 and Fig. 5.3), since the vast majority of the resulting DHD correlations for $\widehat{\text{corr}}(Y_2, \widetilde{Z}_1)$ and $\widehat{\text{corr}}(Y_2, \widetilde{Z}_2)$ do not cover the immediate area around the benchmark values of 0.44 and 0.48. The mean DHD correlations for $\widehat{\text{corr}}(Y_2, \widetilde{Z}_1)$ and $\widehat{\text{corr}}(Y_2, \widetilde{Z}_2)$ reported in Table A.1 are 0.31 and 0.34, deviating from the true values by $0.44 - 0.31 = 0.13$ and $0.48 - 0.34 = 0.14$ on average. All other methods yield basically similar and acceptable results. The RMSE values illustrated in Figure 5.3 and Table A.2 for $\widehat{\text{corr}}(Y_2, \widetilde{Z}_1)$ and $\widehat{\text{corr}}(Y_2, \widetilde{Z}_2)$ are the lowest for RF and PVM-RF, while those for RM, PMM, DT and PVM-DT are slightly higher. However, such small differences in simulation studies could be random and should therefore not be overinterpreted. The RMSE values indicate an acceptable performance for all algorithms except of DHD, with slight performance disadvantages of PVM-DT for $\widehat{\text{corr}}(Y_2, \widetilde{Z}_2)$. Despite the slight performance disadvantages observed in some cases with PVM-DT, it can be concluded for the CIA compliance and high donor ratio scenarios that all methods except DHD yield acceptable results.

An important aspect in this thesis and in examining different data fusion approaches is not only

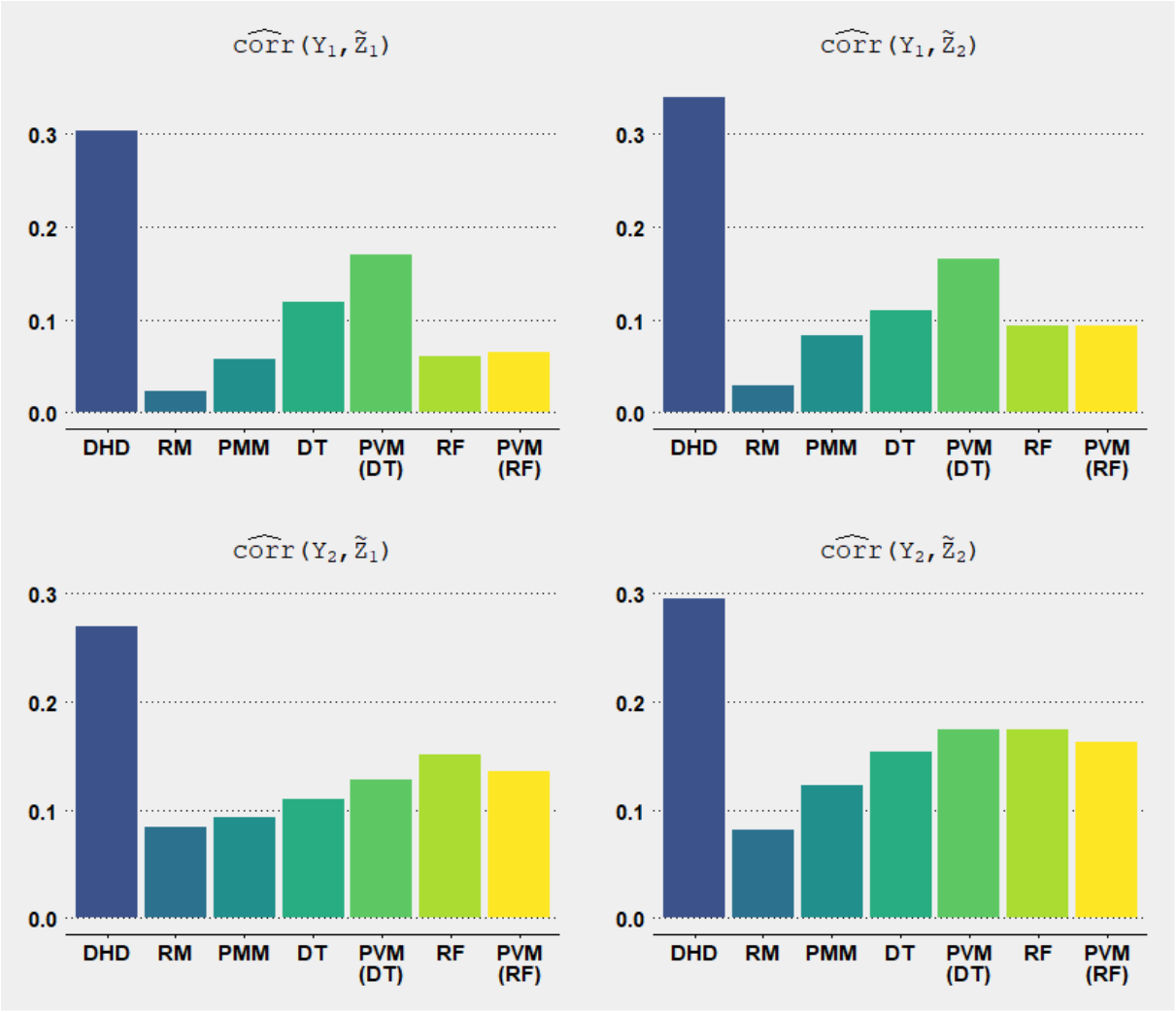


Source: EU-SILC SUF DE (2015); EU-SILC SUF FR (2015).

Figure 5.4: MC distributions for $\widehat{\rho}_{\mathbf{Y}\tilde{\mathbf{Z}}}$ with n_2 under CIA Compliance

to consider a single but various scenarios. We therefore provide in Figure 5.4 the distributions for all $k = 1{,}000$ resulting correlations for n_2, that is, for the scenario of a low donor ratio ($n_{silc} = 4{,}000$ and $n_{hbs} = 400$). Compared to the high donor ratio, it is apparent that DHD yields worse results compared to n_1 and therefore seems to be quite sensitive to the low donor pool. Accordingly, the mean DHD correlations for $\widehat{\text{corr}}(Y_1, \tilde{Z}_1)$ and $\widehat{\text{corr}}(Y_1, \tilde{Z}_2)$ decreases from 0.74 and 0.69 under n_1 to 0.57 and 0.52 under n_2, as can be seen in Table A.1. Thus, the DHD correlations are even more biased under the low donor ratio scenario for high original correlations. This is also the case for medium associations, where the mean DHD correlations for $\widehat{\text{corr}}(Y_2, \tilde{Z}_1)$ and $\widehat{\text{corr}}(Y_2, \tilde{Z}_2)$ decreases from 0.31 and 0.34 under n_1 to 0.17 and 0.19 under n_2. This sensitivity to a small donor pool can also be observed, albeit to a lesser extent, for PVM-DT for all four correlations considered, and for PVM-RF only for medium original correlations with regard to $\widehat{\text{corr}}(Y_2, \tilde{Z}_1)$ and $\widehat{\text{corr}}(Y_2, \tilde{Z}_2)$. The matching classes for PVM-DT, resulting from

a large number of minimal distances, are likely to be even less precise due to the smaller donor pool, which further biases the correlations downwards compared to n_1.



Source: EU-SILC SUF DE (2015); EU-SILC SUF FR (2015).

Figure 5.5: RMSE of $\widehat{\rho}_{\mathbf{Y}\tilde{\mathbf{Z}}}$ with n_2 under CIA Compliance

PMM, on the other hand, which is also a nearest neighbour method, seems to be less sensitive to the small donor pool in this case. In the underlying simulation, this is probably related to the fact that the regression model on which PMM is based in this case represents a very good approximation of the high linear correlation between one X variable and $\mathbf{Z}$, and can thus also reproduce the linear correlations between $\mathbf{Y}$ and $\mathbf{Z}$ quite precisely due to the approximate CIA fulfilment. Since PMM is based on regressions, PMM can reproduce the original correlations somewhat better here than PVM-DT and PVM-RF. RM, however, shows the best performance under n_2 as well as the lowest sensitivity to the small donor pool across all four correlations considered, which is also evident from the respective RMSEs in Figure 5.5, each of which is lowest for RM. In this respect, RM even seems to benefit slightly from the small donor pool,

since the RMSEs for RM are lower than for n_1 for all four correlations (see Tab. A.2). For DT and RF, however, the small donor pool means that the original correlations cannot be perfectly reproduced. Yet, as indicated in Table A.2, the DT and RF correlations are still less biased for n_2 compared to DHD, even when compared to the DHD results under n_1.

Overall, RM and PMM show an extremely low sensitivity to a small donor pool, which is probably mainly due to the high linear correlations. The sensitivity of the statistical learning methods is somewhat higher, yet the RMSE values for DT, PVM-DT, RF and PVM-RF increase only moderately under a low donor ratio. Thus, the difference between the RMSE values of the SL methods under n_1 and n_2 is mostly between about 0.02 and 0.05. The largest RMSE difference among the SL methods is observed for PVM-DT and is $0.17 - 0.11 = 0.06$ for $\widehat{\text{corr}}(Y_1, \widetilde{Z}_1)$ and $\widehat{\text{corr}}(Y_1, \widetilde{Z}_2)$, respectively (see Tab A.2). The performance losses of DHD, on the other hand, are much higher, since the RMSE values for DHD under n_2 are always at least 0.1 higher than under n_1. The differences in the DHD RMSEs for $\widehat{\text{corr}}(Y_1, \widetilde{Z}_1)$ and $\widehat{\text{corr}}(Y_1, \widetilde{Z}_2)$, that is, for high original correlations, are even $0.30 - 0.15 = 0.15$ and $0.34 - 0.18 = 0.16$, respectively.

Overall, the first interim conclusion is that the traditional DHD method, which has been frequently used for data fusions in practice, is clearly inferior to the other approaches, even if the basic prerequisite for a data fusion is favourable due to the approximate CIA fulfilment. Nevertheless, DHD cannot reproduce the original correlations between $\mathbf{Y}$ and $\mathbf{Z}$ under either a high donor ratio or a low donor ratio scenario. All other methods, on the other hand, can reproduce the correlations between $\mathbf{Y}$ and $\mathbf{Z}$ to an acceptable extent, with small losses in performance observed in particular for PVM-DT. The sensitivity to a low donor ratio is lowest for RM and PMM, moderate for DT, PVM-DT, RF and PVM-RF and relatively high for DHD.

Correlations Between X and Z

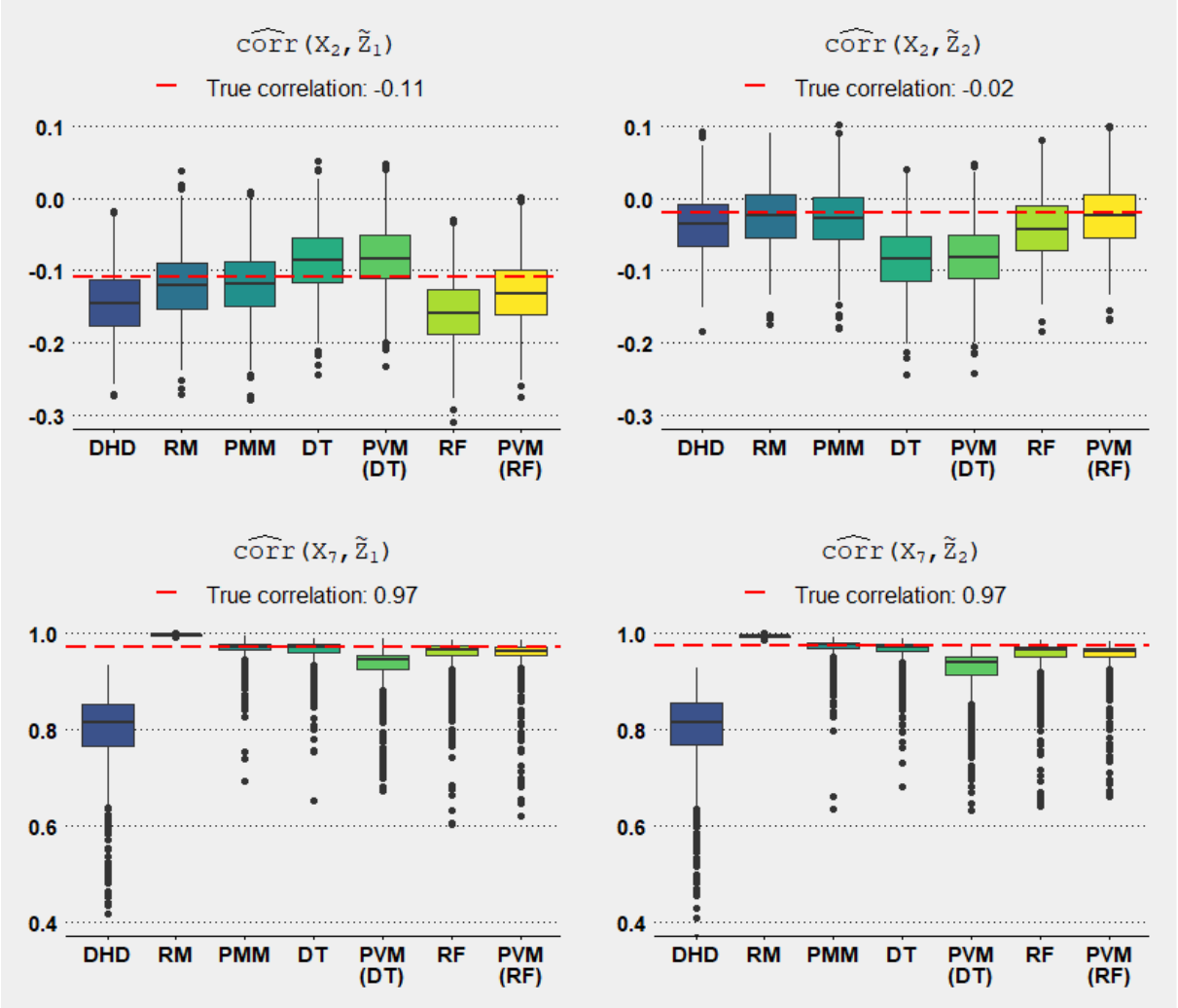
The preservation of the joint distribution $f(\mathbf{X}, \mathbf{Z})$ already observed in the donor data file is a kind of minimum requirement for data fusions and corresponds to the fourth validity level. In the context of the simulations discussed in this chapter, it is useful to briefly examine how the potential data fusion algorithms are able to meet this requirement. Hence, we additionally evaluate the correlations between the metric common variables (X_2 and X_7) and the specific $\mathbf{Z}$ variables to be fused. Table 5.4 contains the respective benchmark correlations $\rho_{\mathbf{XZ}}$ observed from the database described in Section 5.2.1, and it is apparent that the correlations between X_2 and $\mathbf{Z} = (Z_1, Z_2)$ are relatively low with -0.11 and -0.02, respectively, whereas quite high correlations are obtained between X_7 and $\mathbf{Z} = (Z_1, Z_2)$ of 0.97 each. We are still in the CIA

compliance scenario, since **Y** and **Z** given **X** are approximately conditionally independent (see Tab. 5.3). However, note that the CIA is irrelevant here, since the joint distribution of **X** and **Z** is known from the donor data file and thus no identifying assumptions are required.

Table 5.4: Benchmark Parameters for $\rho_{\mathbf{XZ}}$

$\text{corr}(X_2, Z_1)$	$\text{corr}(X_2, Z_2)$	$\text{corr}(X_7, Z_1)$	$\text{corr}(X_7, Z_2)$
−0.1081	−0.0208	0.9699	0.9737

Source: EU-SILC SUF DE (2015); EU-SILC SUF FR (2015).

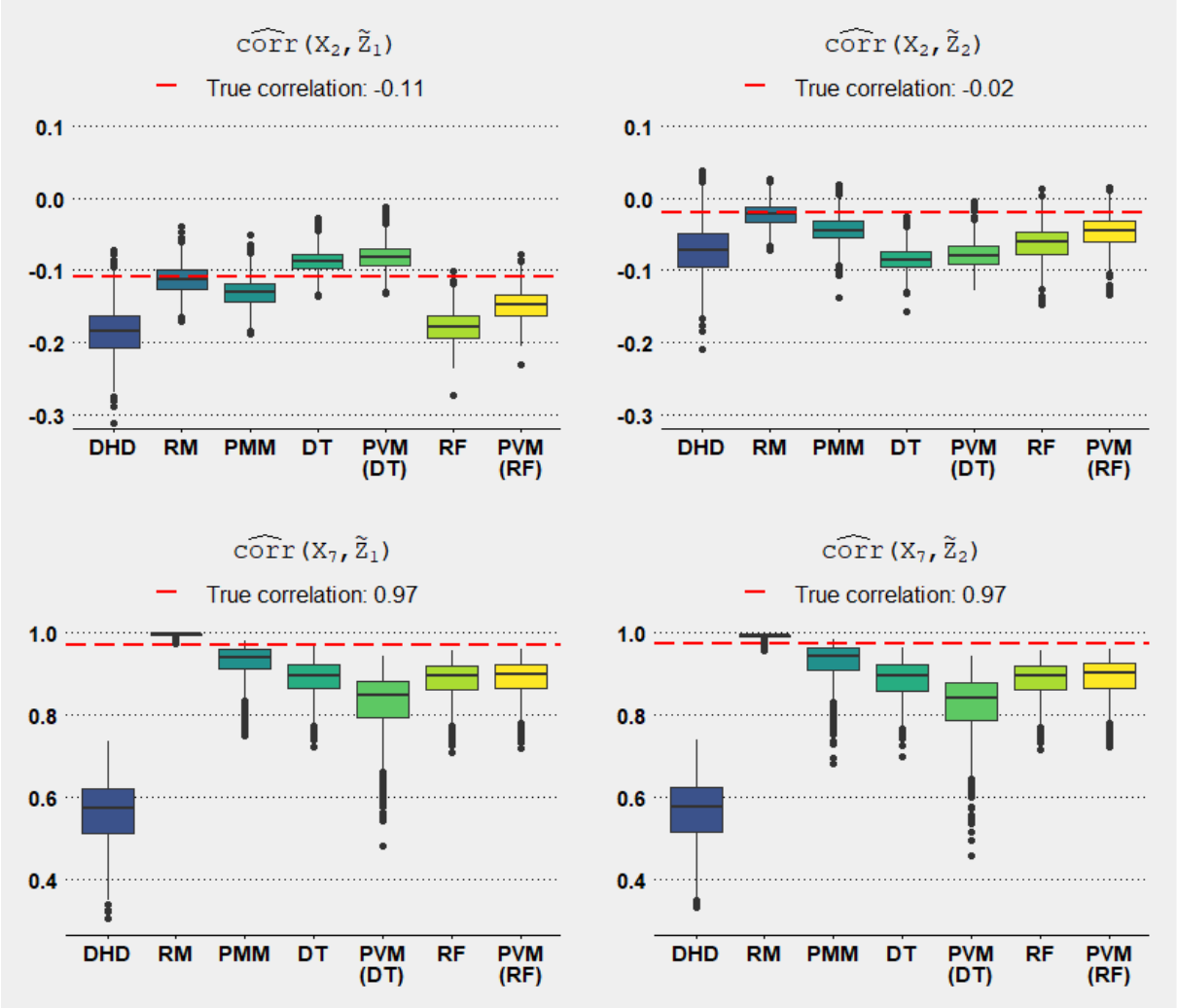


Source: EU-SILC SUF DE (2015); EU-SILC SUF FR (2015).

Figure 5.6: MC distributions for $\widehat{\rho}_{\mathbf{X}\tilde{\mathbf{Z}}}$ with n_1 under CIA Compliance

Figure 5.6 illustrates the Monte Carlo distributions of all $k = 1{,}000$ resulting correlations for each method under n_1, that is, a ten times higher number of donors compared to recipient units. Here, it is apparent that no striking differences can be observed between the different data fusion

methods for low original correlations between **X** and **Z** (upper part of Fig. 5.6). One explanation could be that all methods except DHD include strategies to account for different associations between the common **X** variables and the specific **Z** variables. Since the correlations between X_2 and **Z** are low, the variable X_2 should rarely have high influence in the imputation models of the respective methods (especially since with X_7 a highly correlated variable is present). Accordingly, correlations close to 0 after imputation are not surprising for the corresponding methods. Note that DHD, in contrast, does not use information about the correlations between **X** and **Z** and thus gives equal weight to all **X** variables. Figure 5.7 illustrates the results for the low donor ratio scenario under n_2. Here, no substantial differences are observed for $\widehat{\text{corr}}(X_2, \widetilde{Z}_1)$ and $\widehat{\text{corr}}(X_2, \widetilde{Z}_2)$ compared to n_1, but the boxes (whose boundaries correspond to the lower and upper quartiles) are narrower, suggesting lower variance under n_2.



Source: EU-SILC SUF DE (2015); EU-SILC SUF FR (2015).

Figure 5.7: MC distributions for $\widehat{\rho}_{\mathbf{X}\widetilde{\mathbf{Z}}}$ with n_2 under CIA Compliance

Regarding the high original correlations between X_7 and **Z**, it is evident that DHD cannot repro-

duce them adequately. For DHD, the mean correlations for $\widehat{\text{corr}}(X_7, \widetilde{Z}_1)$ and $\widehat{\text{corr}}(X_7, \widetilde{Z}_2)$ under n_1 are 0.80 each, and thus on average $0.97 - 0.80 = 0.17$ lower than the benchmark correlations (see Tab. A.3). Again, DHD reveals relatively high sensitivity to a low donor ratio. Under n_2, the mean correlations of DHD are 0.56 for $\widehat{\text{corr}}(X_7, \widetilde{Z}_1)$ and 0.57 for $\widehat{\text{corr}}(X_7, \widetilde{Z}_2)$, biasing them on average by $0.97 - 0.56 = 0.41$ and $0.97 - 0.57 = 0.40$, respectively. Similarly, the RMSE values for DHD under n_2 are more than two times higher than under n_1 (see Tab. A.4). The performance of the other methods for $\widehat{\text{corr}}(X_7, \widetilde{Z}_1)$ and $\widehat{\text{corr}}(X_7, \widetilde{Z}_2)$ is acceptable, whereas for RM it has to be stated that here the correlations are rather overestimated and biased towards 1. This is probably due to the strong linear correlation between X_7 and $\mathbf{Z}$, which is already assumed by the linear regression. Nevertheless, due to the low variance of the RM correlations, very low RMSE values result for RM. PMM, DT, RF and PVM-RF lead to relatively similar results for n_1, whereby the small donor pool under n_2 leads to a slight underestimation of the correlations in each case. The next largest RMSE after DHD can be observed under n_2 for PVM-DT, which is 0.15 for $\widehat{\text{corr}}(X_7, \widetilde{Z}_1)$ and 0.17 for $\widehat{\text{corr}}(X_7, \widetilde{Z}_2)$. In this respect, PVM-DT nevertheless performs significantly better than DHD, but performance losses are again observed for PVM-DT, similar to $\rho_{\mathbf{YZ}}$, compared to RM, PMM, DT, RF and PVM-RF. Thus, the results for the correlations between $\mathbf{X}$ and $\mathbf{Z}$ are relatively similar to those previously discussed for the correlations between $\mathbf{Y}$ and $\mathbf{Z}$.

5.3.2 CIA Violation

Correlations Between Y and Z

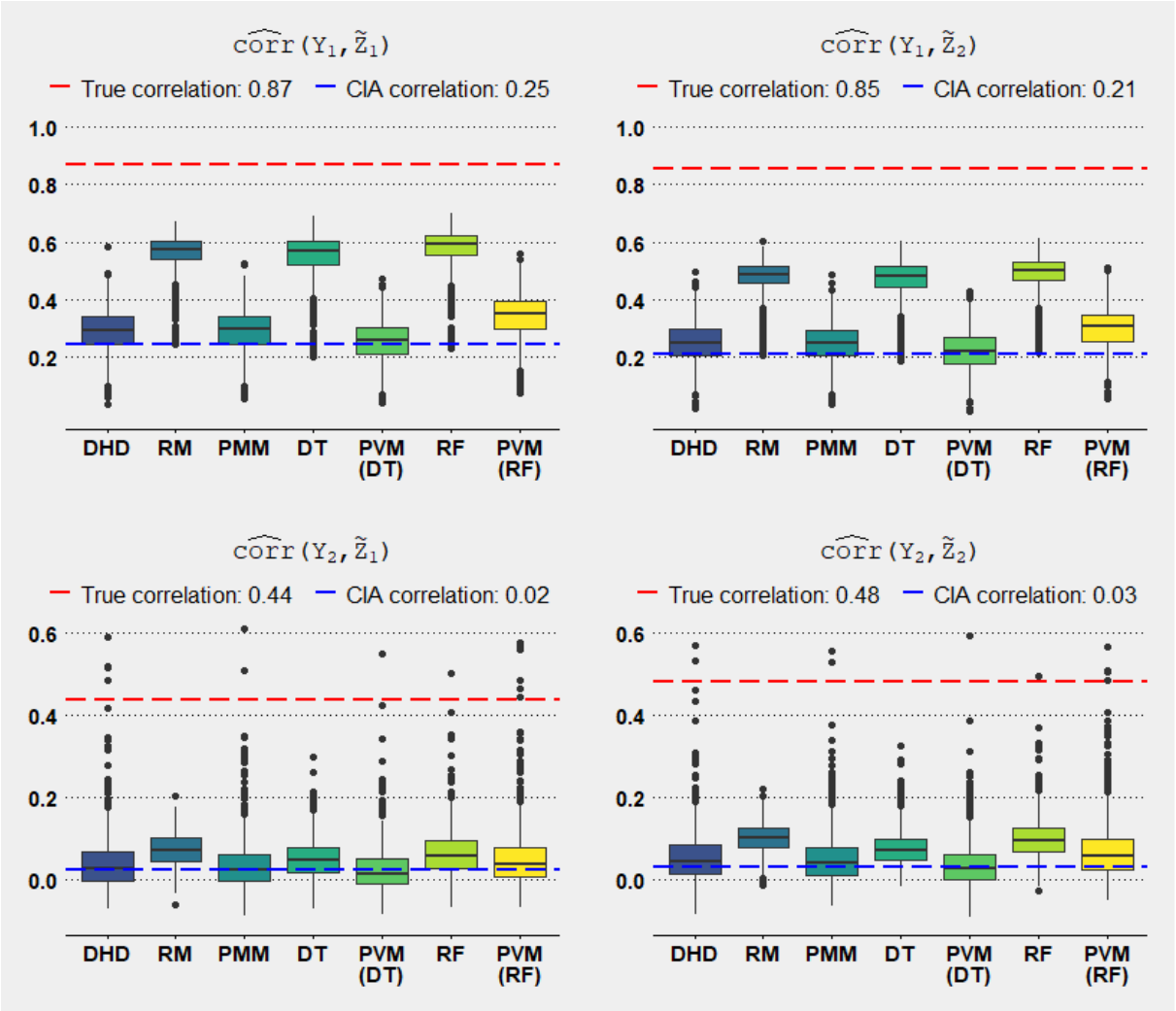
Besides the explicit scenarios of a high or low donor ratio, compliance or violation of the CIA is another scenario that is expected to have a direct impact on the data fusion result of each considered data fusion algorithm. In this respect, we now introduce the results concerning the correlations between the specific variables $\mathbf{Y}$ and $\mathbf{Z}$ under CIA violation, also for varying donor-recipient ratios, represented by n_1 and n_2. Table 5.5 shows the benchmark correlations from the database with regard to the specific $\mathbf{Y}$ and $\mathbf{Z}$ variables originally not jointly observed, with the 'true' correlations being equivalent to those of Table 5.3. Again, the CIA correlation is obtained using Equation (5.1). However, since we now exclude the highly correlated X_7 variable as a common characteristic, it can be seen that the CIA correlations are now much lower than the original correlations, indicating an increased violation of the CIA. While CIA correlations of 0.25 and 0.21 are observed for $\widehat{\text{corr}}(Y_1, \widetilde{Z}_1)$ and $\widehat{\text{corr}}(Y_1, \widetilde{Z}_2)$, the correlations under CIA for

$\widehat{\text{corr}}(Y_2, \widetilde{Z}_1)$ and $\widehat{\text{corr}}(Y_2, \widetilde{Z}_2)$ are close to 0 and amount to 0.02 and 0.03, respectively.

Table 5.5: Benchmark Parameters for $\rho_{\mathbf{YZ}}$ under CIA Violation

	$\text{corr}(Y_1, Z_1)$	$\text{corr}(Y_1, Z_2)$	$\text{corr}(Y_2, Z_1)$	$\text{corr}(Y_2, Z_2)$
True correlation	0.8678	0.8536	0.4361	0.4831
CIA correlation	0.2466	0.2108	0.0230	0.0319

Source: EU-SILC SUF DE (2015); EU-SILC SUF FR (2015).

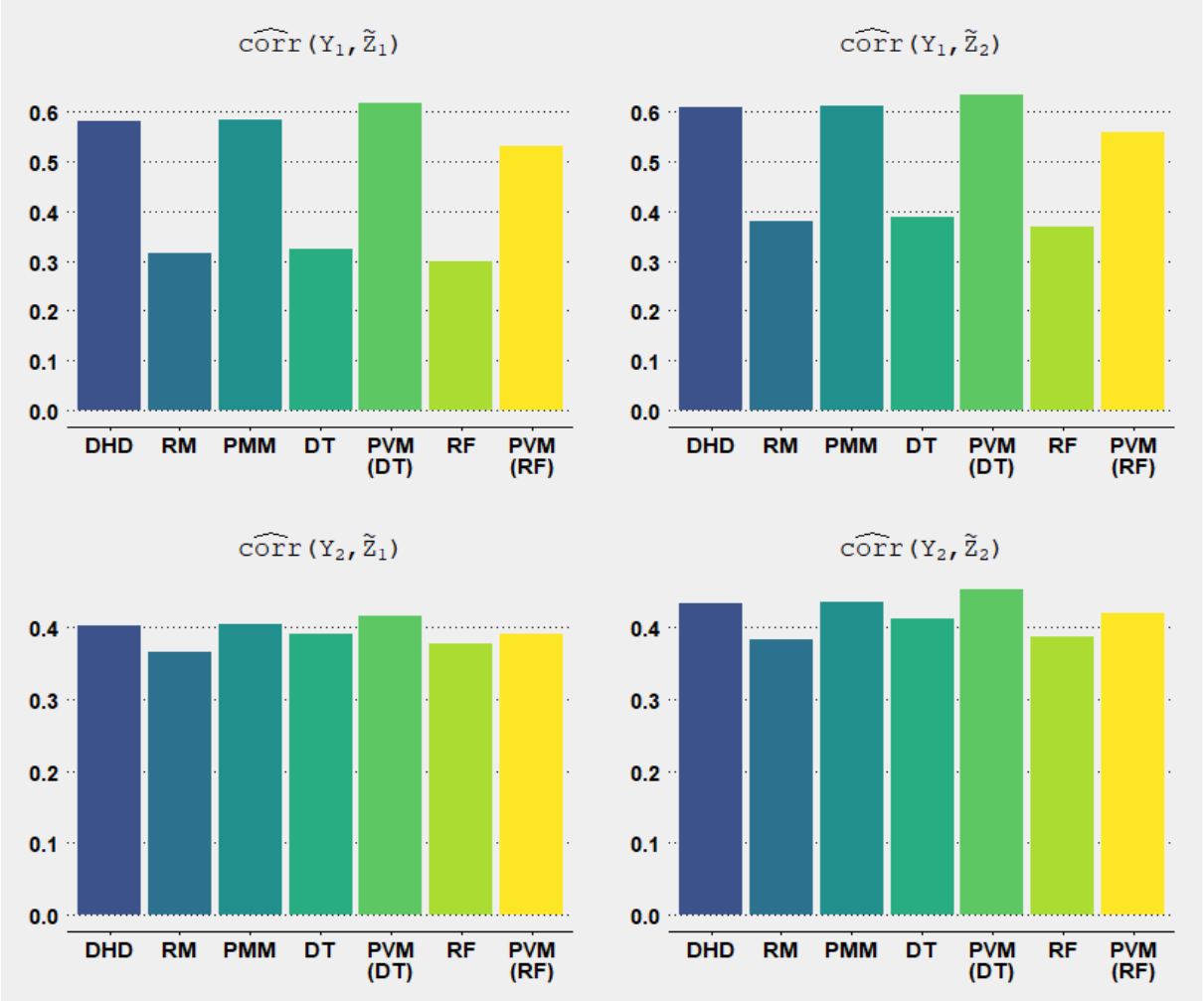


Source: EU-SILC SUF DE (2015); EU-SILC SUF FR (2015).

Figure 5.8: MC distributions for $\widehat{\rho}_{\mathbf{Y}\widetilde{\mathbf{Z}}}$ with n_1 under CIA Violation

Figure 5.8 illustrates the respective Monte Carlo distributions of the $k = 1,000$ resulting correlations for each method under a high donor ratio (n_1). Quite interesting phenomena arise here. DHD, PMM, PVM-DT and PVM-RF predominantly produce the CIA correlations and thus fairly underestimate the true correlations. Especially for high original correlations, how-

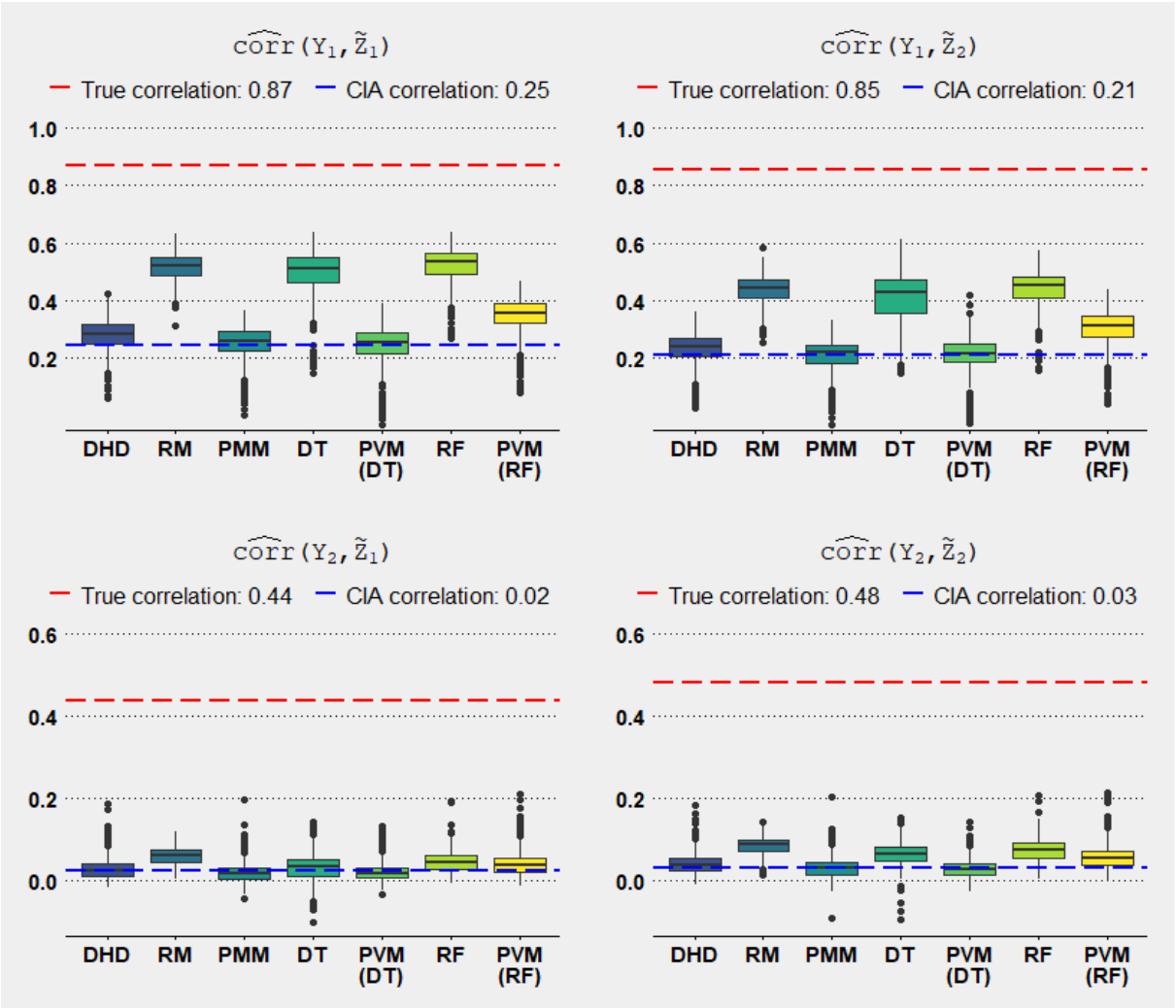
ever, it can be seen that the three pure prediction methods, RM, DT and RF, come closer to the true associations than all other algorithms. The mean correlations resulting from RM, DT and RF for $\widehat{\text{corr}}(Y_1, \widetilde{Z}_1)$ and $\widehat{\text{corr}}(Y_1, \widetilde{Z}_2)$ lay between 0.47 and 0.58 under n_1 (see Tab. A.5). The highest mean correlation is observed for RF and is 0.58 for $\widehat{\text{corr}}(Y_1, \widetilde{Z}_1)$. Thus, the high original correlations are nevertheless biased by at least $0.87 - 0.58 = 0.29$ when applying one of the prediction methods RM, DT and RF. Although this does not seem immediately satisfactory, it should also be noted that all other methods, DHD, PMM, PVM-DT and PVM-RF on average only produce correlations between 0.22 and 0.34 for $\widehat{\text{corr}}(Y_1, \widetilde{Z}_1)$ and $\widehat{\text{corr}}(Y_1, \widetilde{Z}_2)$, which are thus biased by at least $0.87 - 0.34 = 0.53$. Consequently, the RMSE values illustrated in Figure 5.9 indicate for high original correlations much lower RMSEs for RM, DT and RF compared to the nearest neighbour methods DHD, PMM, PVM-DT and PVM-RF. Among the nearest neighbour methods, PVM-RF reveals a slightly better performance than DHD, PMM and PVM-DT, but is still considerably inferior to the prediction methods.



Source: EU-SILC SUF DE (2015); EU-SILC SUF FR (2015).

Figure 5.9: RMSE of $\widehat{\rho}_{\mathbf{Y}\widetilde{\mathbf{Z}}}$ with n_1 under CIA Violation

The effects described for high original correlations are not observable for medium original correlations and a CIA correlation close to 0 and can only be recognised to some extent. The lower boxplots in Figure 5.8 illustrate that all methods merely produce correlations around the theoretical CIA correlation, except of few outliers. The observed potential of RM, DT and RF to come closer to the original correlations despite the CIA violation is only rudimentarily recognisable. With regard to the medium correlations, the boxes for RM, DT and RF are slightly higher than those of the other methods. This is also true for the mean values, while their RMSE values are slightly lower compared to DHD, PMM, PVM-DT and PVM-RF, as apparent in Figure 5.9 and Table A.6. The striking effects of RM, DT and RF to at least get closer to the original correlations are thus not evident here.



Source: EU-SILC SUF DE (2015); EU-SILC SUF FR (2015).

Figure 5.10: MC distributions for $\widehat{\rho}_{\mathbf{Y}\widetilde{\mathbf{Z}}}$ with n_2 under CIA Violation

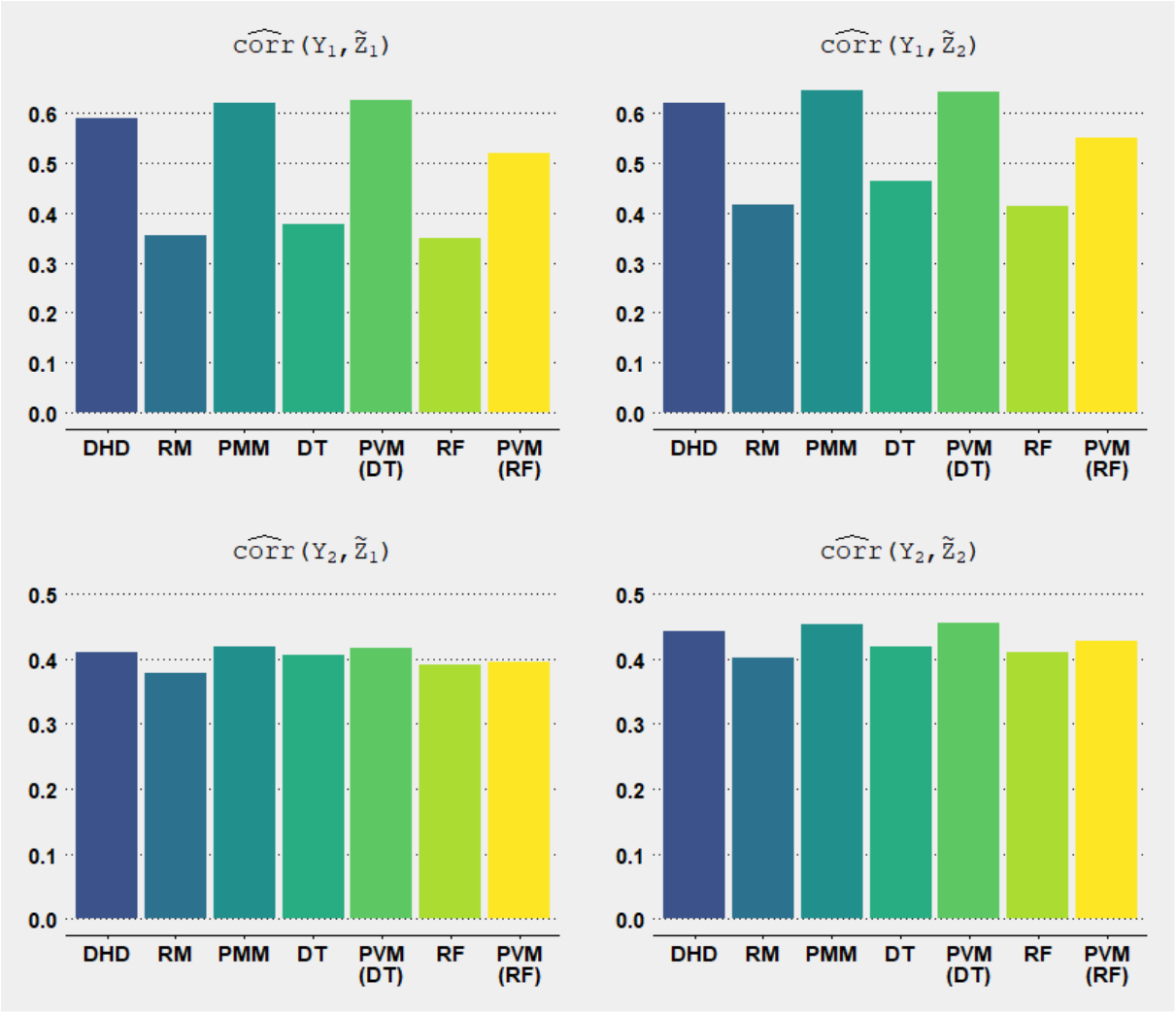
The results under a low donor ratio are illustrated in Figures 5.10 and 5.11. Here, we see for high original correlations that the boxes of RM, DT and RF are a little further down compared

to those under n_1. Consequently, slightly lower means and slightly higher RMSE values are observed for RM, DT and RF under n_2 compared to n_1 (see Tab. A.5 and A.6). In contrast to the CIA compliance scenario, in which RM was found to be robust to a low donor ratio, under the CIA violation RM now appears to be slightly sensitive to a low donor ratio. Similarly, for high original correlations, the PMM approach (which is based on linear regressions) also performs slightly worse under n_2 compared to n_1. Apart from that, the results under n_2 for the high correlations between Y_1 and Z_1 and between Y_1 and Z_2 are relatively similar to those under n_1. Once again it can be seen that the nearest neighbour procedures DHD, PMM, PVM-DT and PVM-RF predominantly produce the CIA correlation, whereby PVM-RF always slightly overestimates the CIA correlation and thus performs slightly better than DHD, PMM and PVM-DT. The correlations for $\widehat{\text{corr}}(Y_1, \widetilde{Z}_1)$ and $\widehat{\text{corr}}(Y_1, \widetilde{Z}_2)$ resulting from the prediction methods RM, DT and RF are again mostly between the CIA correlation and the high original correlation. For medium original correlations, no significant change is observed under n_2 compared to n_1. However, it is evident that the variances for $\widehat{\text{corr}}(Y_2, \widetilde{Z}_1)$ and $\widehat{\text{corr}}(Y_2, \widetilde{Z}_2)$ decrease somewhat under a low donor ratio.

Thus, it can be stated that the CIA violation has a significant impact on the performance of the data fusion methods. However, the respective impacts strongly vary, especially between nearest neighbour methods and prediction approaches. Generally, none of the fusion methods presented is able to reproduce the original correlations in our case. DHD, PMM, PVM-DT and PVM-RF predominantly produce the theoretical CIA correlation. RM, DT and RF come closer to the benchmark values at least with high original correlations and medium underlying CIA correlations of 0.25 and 0.21, respectively, although they are still biased by at least 0.29.

Correlations Between X and Z

Analogous to the scenario of CIA compliance, the extent to which the respective data fusion methods can reproduce the correlations between **X** and **Z** should also be briefly discussed here. The benchmark values can be found in Table 5.4 and remain unchanged. However, a consistent evaluation of the correlations already observed in the donor data file is only possible here for the correlation between X_2 (age) and **Z** (upper part of Fig. 5.12 and 5.13). For the sake of completeness, the results of the correlations between X_7 and **Z** are nevertheless presented here (lower part of Fig. 5.12 and 5.13). It should be noted here, however, that the consideration of the correlations between X_7 and **Z** corresponds to an evaluation along the third validity level, since X_7 does not represent a common variable in the fusion process and can thus be regarded

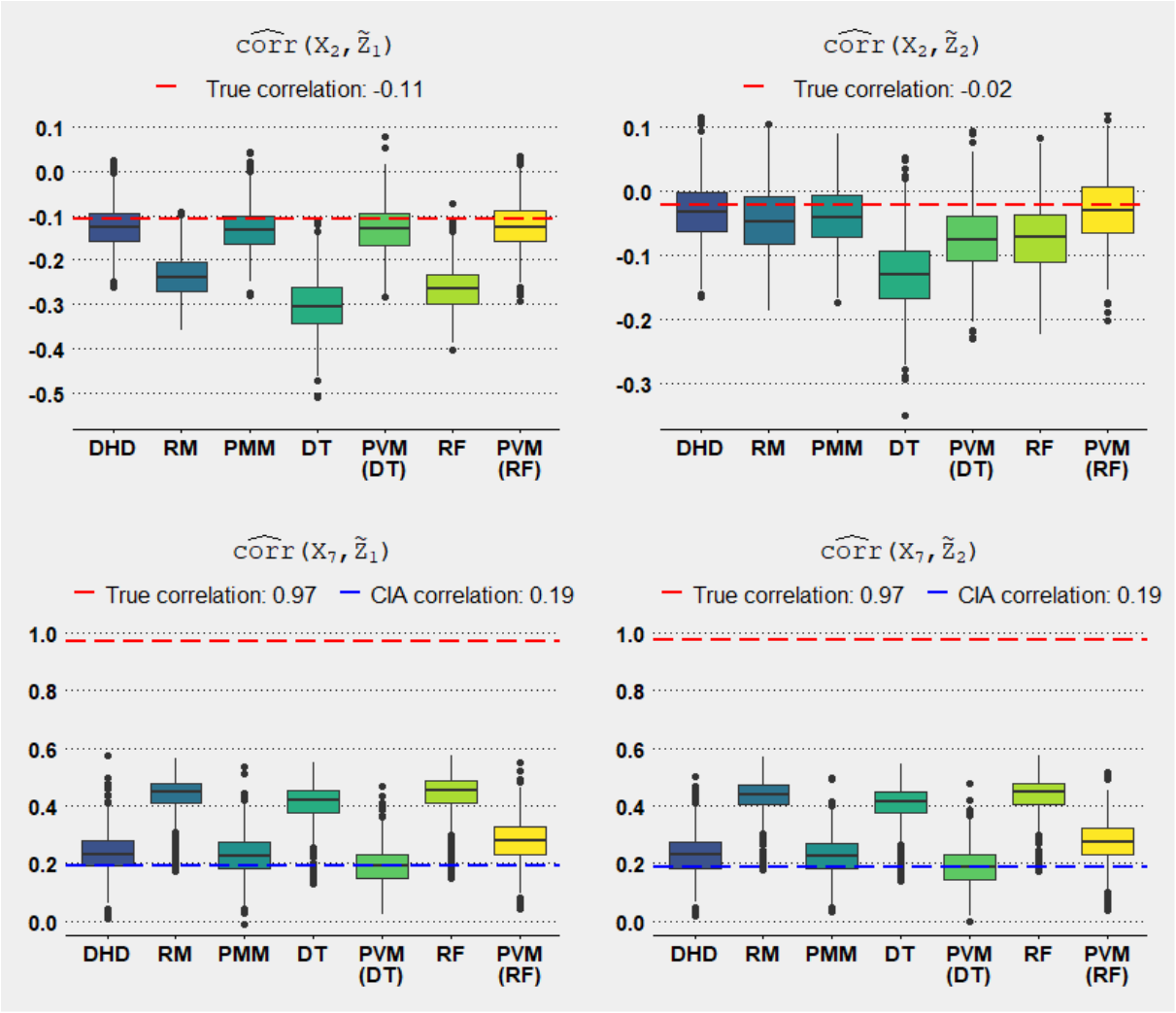


Source: EU-SILC SUF DE (2015); EU-SILC SUF FR (2015).

Figure 5.11: RMSE of $\widehat{\rho}_{\mathbf{Y}\widetilde{\mathbf{Z}}}$ with n_2 under CIA Violation

as a specific Y or Z variable.

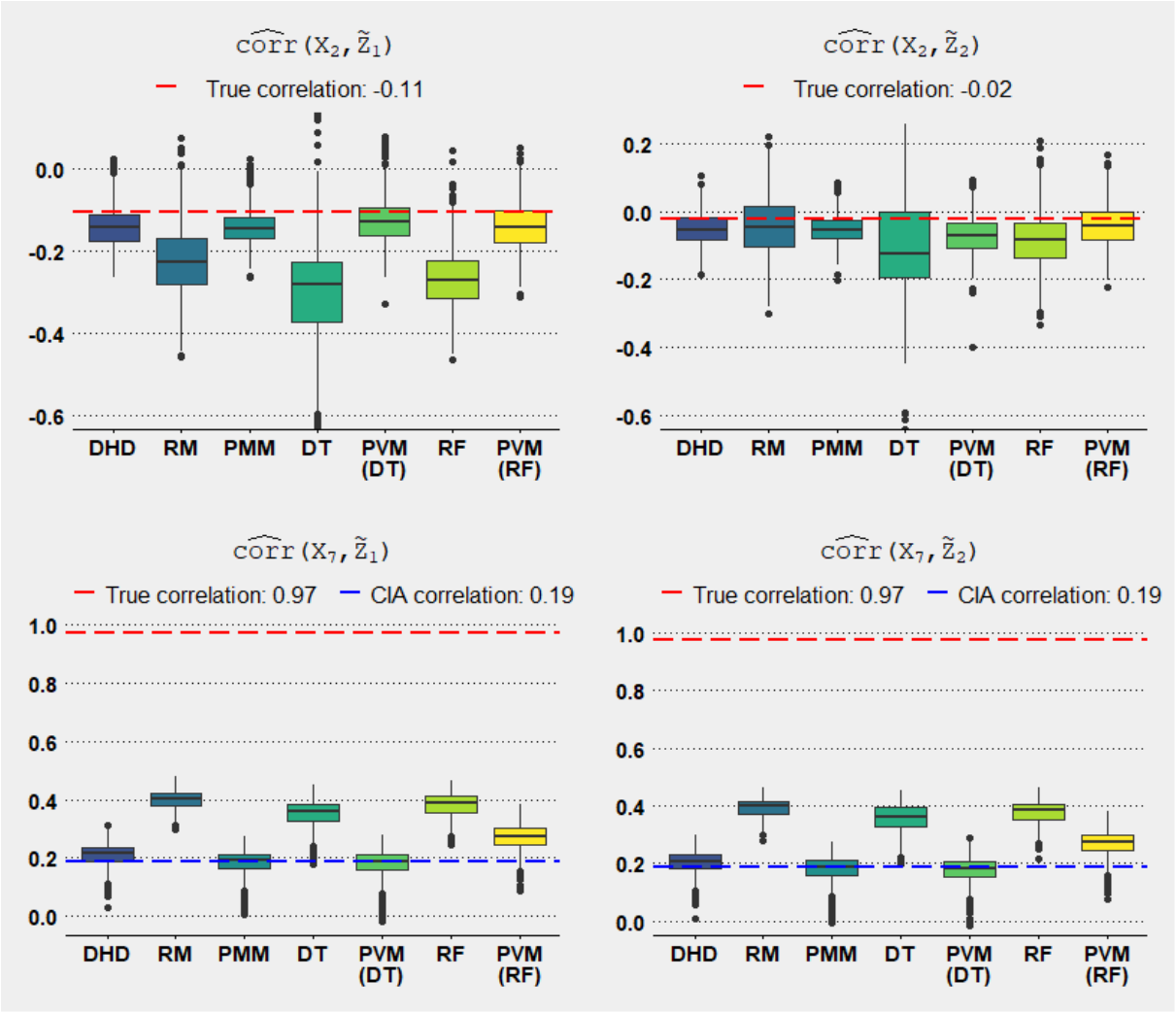
Figure 5.12 shows the resulting correlations under a high donor ratio (n_1), Figure 5.13 in turn illustrates the results for the scenario of a low donor ratio (n_2). It is particularly striking for $\widehat{\text{corr}}(X_2, \widetilde{Z}_1)$ that the nearest neighbour methods DHD, PMM, PVM-DT and PVM-RF yield relatively unbiased results for n_1 and n_2, while the pure prediction methods RM, DT and RF overestimate the original correlations. Consequently, the mean values for $\widehat{\text{corr}}(X_2, \widetilde{Z}_1)$ resulting from DHD, PMM, PVM-DT and PVM-RF are close to the true correlation of -0.11 (see Tab. A.7). The RMSE values of the nearest neighbour procedures for $\widehat{\text{corr}}(X_2, \widetilde{Z}_1)$ are again lower than the RMSE values of RM, DT and RF (see Tab. A.8). For $\widehat{\text{corr}}(X_2, \widetilde{Z}_2)$, the results are very similar to those from the CIA compliance scenario, whereby DT in particular tends to be overestimated under n_1. Under n_2, on the other hand, it can be seen for $\widehat{\text{corr}}(X_2, \widetilde{Z}_2)$ that the DT correlations have a relatively high variance.



Source: EU-SILC SUF DE (2015); EU-SILC SUF FR (2015).

Figure 5.12: MC distributions for $\widehat{\rho}_{\mathbf{X}\widetilde{\mathbf{Z}}}$ with n_1 under CIA Violation

With regard to the correlations between X_7 and $\mathbf{Z}$ (lower part of Fig. 5.12 and 5.13), it should first be noted that X_7 was excluded from the fusion process. Therefore, strictly speaking, X_7 does not represent a common variable here, as already indicated. Instead, it can be seen as a specific Y or Z variable. Since X_7 was not imputed but is already present in the recipient data file, it would be a third specific variable from the recipient data source and could therefore be referred to as Y_3 according to our notation. For the sake of a uniform and clear notation, the variable will nevertheless continue to be referred to as X_7. However, it is important to note that the evaluation of the correlations between X_7 and $\mathbf{Z}$ now correspond to an evaluation on the third validity level, since it was presumed in the simulation that X_7 and $\mathbf{Z}$ were never jointly observed. In this respect, the CIA must now be assumed again here. The theoretical correlation between X_7 and Z_1 as well as between X_7 and Z_2, which would result if the CIA were true, is 0.19 in each case (marked by the blue horizontal line). The effects observable for $\widehat{\text{corr}}(X_7, \widetilde{Z}_1)$



Source: EU-SILC SUF DE (2015); EU-SILC SUF FR (2015).

Figure 5.13: MC distributions for $\widehat{\rho}_{\mathbf{X}\widetilde{\mathbf{Z}}}$ with n_2 under CIA Violation

and $\widehat{\text{corr}}(X_7, \widetilde{Z}_2)$ under n_1 and n_2 are very similar to the results of $\rho_{\mathbf{YZ}}$ previously discussed and illustrated in the upper part of Figures 5.8 and 5.10. While the nearest neighbour methods largely replicate the CIA correlations around 0.19, RM, DT and RF come closer to the original correlations, but still remain located far below the benchmark correlations of 0.97.

5.3.3 Discussion

The results presented under the explicit scenarios of compliance and violation of the CIA as well as under high and low donor ratios include multivariate imputation solutions for DHD, PMM, PVM-DT and PVM-RF. For PMM, PVM-DT and PVM-RF, on the other hand, univariate, that is, variable-by-variable imputations are also possible. Appendix B.1 includes the respective results for the univariate PMM and PVM solutions with regard to the correlations between $\mathbf{Y}$ and

Z, illustrated in Figures B.1 to B.4. Here, no substantial change compared to the multivariate results are observable. This is encouraging for practical applications, as multivariate imputation solutions are likely to be preferred by most researchers and our results suggest that this preference is at least not countered by performance problems compared to univariate imputation. Additional results for the correlations between **X** and **Z** also indicate, as expected, no significant difference between the multivariate and univariate imputation solution for PMM, PVM-DT and PVM-RF (EU-SILC SUF DE 2015; EU-SILC SUF FR 2015). Apart from this, however, the results presented give rise to discussion.

First, even in the favourable case of CIA compliance, it was striking that DHD, a traditional and frequently used data fusion method in practice, does not provide adequate results and underestimates the correlations between the variables **Y** and **Z** originally not jointly observed. One problem with DHD could be that all **X** variables were used in the fusion process, regardless of whether they are even a relevant influencing variable for the **Z** variables to be imputed. The use of irrelevant **X** variables in turn leads to relatively inefficient matches. A prominent possibility to mitigate this problem, which is often discussed in the context of covariate-based nearest neighbour methods, is the selection of relevant **X** variables via stepwise regression (see e.g. Lamarche et al. 2020), whereby only the **X** variables selected in this way are subsequently considered for DHD.

In the multivariate case with $p_{don} > 1$, however, it must then be determined which of the **Z** variables is the dependent variable of the stepwise regression. One possibility would be to carry out a stepwise selection for each Z variable to be imputed and then apply DHD for each Z variable separately with the respective selected **X** variables. If different **X** variables are selected for different **Z** variables, this would be equivalent to a univariate imputation solution. However, this potential univariate imputation solution cannot be influenced in advance, but depends solely on which **X** variables are selected for the respective **Z** variables by the stepwise regression. Alternatively, for $p_{don} > 1$ the stepwise regression could also be calculated on all **Z** variables separately, whereby now the intersection of the selected **X** variables could then serve as the final set of common variables for DHD. However, this is at best recommendable for a few **Z** variables (such as for $p_{don} \leq 3$), which moreover have a very similar correlation structure with the common **X** variables. Otherwise, too careless or even complete exclusion of relevant **X** variables would be the consequence.

In order to investigate the effects of stepwise regression for DHD, additional simulations were carried out using backward deletion based on the `regsubsets()` function from the `leaps` pack-

age (Lumley and Miller 2022). For the backward deletion, both Z_1 and, in another simulation, Z_2 were used as dependent variables. However, the results show only extremely marginal and no substantial changes in all relevant scenarios compared to DHD without backward deletion (EU-SILC SUF DE 2015; EU-SILC SUF FR 2015).

Thus, stepwise regression cannot overcome the problem associated with DHD that no or only insufficient information about the relationship between **X** and **Z** is included in the fusion process. All **X** variables or, in the case of prior selection by stepwise regression, all selected common variables have the same weight in the fusion process of DHD and other covariate-based nearest neighbour approaches. Covariate-based nearest neighbour methods thus implicitly assume that all selected common variables have the same explanatory power with regard to the **Z** variables to be matched. However, this is unrealistic. Some **X** variables are more important for identifying nearest neighbours suitable for the specific **Z** variables, while others are less relevant. All other methods incorporate the differential influence of the common **X** variables in the fusion process through the underlying regression or the respective statistical learning method.

In general, the crucial task for nearest neighbour methods such as DHD, PMM and PVM is to provide distance calculations that are as adequate as possible. This is by no means trivial. Exact matches, that is, zero distances between the recipient and donor observations according to the common **X** variables, are unlikely in practical applications and impossible for continuous variables. Moreover, in principle, the higher the number of common **X** variables and the smaller the donor pool, the less likely are exact matches (Meinfelder and Schaller 2022). In particular, a low number of donors leads on average to greater distances between recipients and their assigned donors (Andridge and Little 2010). The crucial question, then, is how potential nearest neighbour procedures deal with the absence of exact matches. Since PMM and PVM, unlike DHD, incorporate information about the relevance of certain **X** variables with respect to the **Z** variables to be fused, PMM and PVM can produce more efficient matches overall than DHD, leading to better performance of PMM and PVM under CIA compliance. However, performance disadvantages can also be observed for PVM-DT, especially with a low donor ratio. The problem for PVM-DT is probably too undifferentiated matching classes. Although the matching classes do not result from zero distances due to the multivariate case, there are nevertheless many 'most similar' donors for PVM-DT according to Equation (4.14) for each recipient unit, which in turn implies a random draw within these most similar donors. Thus, a too high random component hinders more efficient matches between recipients and donors. For PVM-RF, this problem disappears because RF produces almost no identical predictions and thus PVM-RF is not prone to yield more than one most similar donor unit for the recipients.

However, the nearest neighbour methods predominantly produce the theoretical CIA correlation, which leads to extremely poor results in the case of CIA violation. Since the CIA implicitly underlies the potential data fusion methods, it is assumed that the CIA correlations correspond to the true correlations. In this respect, for nearest neighbour methods, the CIA correlation reflects the maximum achievable correlation after imputation. In the CIA violation scenario for high original correlations, the CIA correlation suggested medium associations between **Y** and **Z** (of 0.25 and 0.21), although the true correlation was significantly higher (0.87 and 0.85). Here, the exaggeration effects described in Section 4.5 in the context of statistical learning methods now prove helpful. These effects apparently also apply to RM in this case, whose imputations are also based on pure predictions. Although the medium correlations of 0.25 and 0.21 reflect the true correlations between **Y** and **Z** according to the CIA, RM and the SL prediction methods overestimate these correlations considerably. Thus, if the actual correlations were similar to the medium CIA correlations, RM, DT and RF would significantly overestimate the correlations, leading to biased results. In conventional imputation applications without an identification problem (and thus without the need for identifying assumptions for the joint distribution of **Y** and **Z**), RM, DT and RF would produce significantly worse results compared to the nearest neighbour methods. Since in the CIA violation scenario the true correlation is significantly higher than the CIA correlation actually presumed, the exaggeration effect of RM, DT and RF is beneficial here, since it results in a smaller underestimation of the correlations. However, this exaggeration effect weakens somewhat with a low donor ratio. In addition, such exaggeration effects were only rudimentary observable for medium original correlations under n_1. The fact that the exaggeration effects seem to increase with a larger donor pool suggests that an even larger donor pool ($n_{don} > 4000$) could further improve the results for RM, DT and RF under CIA violation in this case.

Overall, with regard to the explicit scenarios, it became clear that the CIA has a considerably greater influence on the quality of the results than varying donor-recipient ratios. DHD proved to be particularly sensitive to a small donor pool. DHD's poor distance processing is further exacerbated by the low donor ratio. Similar, but less strong effects are observed for PVM-DT. All other methods, with the exception of RM when the CIA is fulfilled, also suffer from a low donor pool, but to a much lesser extent than DHD. The low sensitivity for RM under CIA compliance is likely due to the very high linear correlation.

5.4 Concluding Remarks

The aim of this thesis, to comprehensively evaluate a concrete plethora of possible data fusion procedures in different data contexts, could now be realised in this chapter by means of a classic use case of data fusion. The data fusion is based on samples with a conventional and easy-to-handle sample size. In addition, the larger dataset serves as donor sample in the specific application case. In the current scientific discourse on official statistics in the EU, the adequate data fusion of EU-SILC and HBS is a prominent topic, as it is intended to create an important database for the more precise measurement of social and economic living standards in the EU. The associated objective is the common measurement of household income and consumption expenditures. Obtaining the joint distribution of income and consumption is thus crucial for the subsequent analyses. The present simulation results show that non-parametric, covariate-based nearest neighbour methods should not automatically be considered as an appropriate data fusion method. This could already be shown by Meinfelder and Schaller (2022) for Random Hot Deck. These findings are particularly important because Eurostat and other NSIs often use covariate-based nearest neighbour methods for data fusion (see e.g. Donatiello et al. 2014; Lamarche et al. 2020). While such methods may often adequately reproduce the marginal distribution of $\mathbf{Z}$, this is not necessarily the case for joint distributions, which is the actual aim of data fusion. Rather, future data fusions of EU-SILC and HBS should be applied using alternative imputation or SL procedures. If real values are to be imputed in a multivariate imputation scenario, PMM and PVM-RF are particularly promising methods.

The present results also show that in the case of approximate CIA compliance, a low donor ratio implies only small performance disadvantages compared to a high donor ratio. In this respect, the HBS could theoretically also serve as the recipient dataset and EU-SILC as the donor data source, provided that content-related or strategic considerations suggest this and small performance disadvantages are considered tolerable. However, the fulfilment of the CIA plays a central role in the performance of the fusion methods. This underlines the relevance of the suggestion by Donatiello et al. (2016) to include common variables as similar as possible to the $\mathbf{Y}$ or $\mathbf{Z}$ variables in the fusion process, if available.

In principle, it should be noted that simulation studies can never claim general validity. General conclusions from the present simulations are subject to the assumption that the results are also transferable to other data situations. Furthermore, the results of this chapter are only limited to continuous $\mathbf{Z}$ variables. Conclusions for categorical $\mathbf{Z}$ variables to be fused cannot be derived

from this. It therefore seems all the more relevant to consider a further application of data fusion, which in some aspects clearly differs from the data situation of this chapter, and furthermore targets the imputation of categorical instead of metric $\mathbf{Z}$ variables. The next chapter is thus devoted to a data fusion constellation that is contrary in various aspects.

Chapter 6

Data Fusion of Tax Statistics and Microcensus

While the data fusion use case of EU-SILC and HBS represents a classical data fusion scenario, we now introduce the data fusion use case of matching the Tax Statistics (TS) with the Microcensus (MC). We will see that this data fusion use case comprises some atypical properties compared to conventional data fusion applications. This is important in order to cover a variety of possible data fusion constellations and thus meet the need of investigating different scenarios. First, we outline the motivation of this particular data fusion use case. The simulation design, including the underlying database and the manipulations to cover the explicit scenarios, is then described. In addition, an empirical evaluation seems useful here, since this data fusion constellation affords target-oriented possibilities for empirical evaluation. Hence, the design of the empirical evaluation is also outlined. Subsequently, the results from the simulations and the empirical evaluation are presented and discussed.

6.1 Motivation and Data Fusion Scenario

The motivation of this data fusion use case closely follows Emmenegger et al. (2023) and is attributed to the challenging data situation on income. Income is a crucial indicator for assessing individual well-being and the welfare of a society and has therefore been studied in depth since the seminal works of Mincer (1958). Consequently, many studies related to economic inequality (Cowell 2000), poverty risks (Ravallion and Chen 1997) or to the concentration and distribution of income (Atkinson 2007; Piketty 2015) examine income as principle measure of

interest. However, consistent and high-quality research on the above topics requires an integrated database that includes reliable information on the full distribution of individual incomes in addition to reliable socio-demographic characteristics (Emmenegger et al. 2023).

In official statistics in Germany, the motivation in high-quality income modelling is currently in particular based on the generation of income variables in the context of microsimulations. The research group 'Multisectoral Regional Microsimulation Model (MikroSim)' (FOR 2559), funded by the German Research Foundation (DFG), is dedicated to creating a synthetic microdatabase to enable researchers to conduct microsimulations on diverse topics (Münnich et al. 2021). Here, income is one important indicator within the synthetic database. The associated objective is to be able to analyse incomes taking social disaggregation variables such as education and working time into account and to make them usable for microsimulations (see e.g. Li and O'Donoghue 2013).

In this respect, the data situation on income comprises many types of datasets, which, by contrast, only depict individual parts of the entire income distribution. Most official statistics and scientific analyses regarding income inequality rely on household surveys (BMAS 2017), while other studies as well as numerous policy evaluations are based on tax income records (Piketty 2015). However, both data sources provide inconsistent estimates with regard to the development of income distribution (see e.g. Bach et al. 2009; Bartels and Schröder 2016; Burkhauser et al. 2018). The survey data indicate a decline in inequality after 2005, while the concentration of top incomes, however, has steadily increased during the same period (Deutscher Bundestag 2017). This reveals data-specific artefacts in the inequality trends in Germany and leads to uncertainties in the interpretation of the scientific results and the associated policy implications. Accordingly, the Scientific Service of the German Bundestag has found that income-related analyses are to be interpreted with respect to the strengths and weaknesses of the several datasets (Deutscher Bundestag 2017). This already indicates that there is no perfect income data in Germany (Emmenegger et al. 2023).

In this context, two studies in particular are of interest for income measurement in official statistics in Germany: the Tax Statistics (TS) and the German Microcensus (MC). Table 6.1 contains a comparison of both datasets including their advantages and disadvantages. As can be seen in Table 6.1, the Tax Statistics (TS) is a complete survey of all tax units in Germany. The original tax data comprises over 40,000,000 tax units. This data is collected via the tax authorities and reflects the information provided in the tax returns. Accordingly, the TS contain complete and reliable information on the taxable income of all tax units in Germany and thus reflect the

entire income distribution from the bottom to the top. This dataset thereby provides the most reliable information on income in Germany. Following the tax system in Germany, the tax units include on the one hand individual assessments, that is, the tax data of a single person, and on the other hand joint assessments, that is, the tax data of married couples filing a joint tax return. However, with regard to high-quality income analyses, important covariates for social disaggregation such as education, family structures, occupation and working hours are missing in the administrative tax data. Hence, the distributional strength of the Tax Statistics is typically not exploited because there are few meaningful explanatory variables for income analyses. In contrast, reliable information on such socio-demographic variables are in turn contained in the German Microcensus, which is the largest official survey sample in Germany with a sample size of 1 % of the population. The survey units reflect individuals in private households. Information on income, on the other hand, is only available in the Microcensus in the form of 24 income classes (Statistische Ämter der Länder 2014: 65). These rudimentary income data also exhibit typical survey-related problems, such as self-response bias, top censoring or reports of classified or heaped data (Angel et al. 2019). Due to the reliable income information in the Tax Statistics, which is also available in regional depth, it seems reasonable to use the Tax Statistics for income analyses. However, such analyses are insufficient if income-relevant covariates of social disaggregation are missing (Emmenegger et al. 2023).

Consequently, the central aim of this data fusion use case is therefore, in a first step, to enhance the Tax Statistics by the socio-demographic variables education and working time from the Microcensus in order to enable high-quality models on income. Individual's level of education and the amount of working time reflect two of the most important factors for the level of income. The effects of years of schooling on income are obvious and have been studied extensively since the early works of Mincer (1958), who analysed returns to education. Mincer (1958) also considers the importance of working hours by including hourly wages. In addition, many studies examine effects of educational attainment and working hours on income distribution (Haughton and Khandker 2009; Atkinson and Bourguignon 2014), including at the regional level (Lee et al. 2016; Panori and Psycharis 2019). Hence, education and working time are chosen as socio-demographic variables in a first step because the inclusion of the interdependencies between income and these two variables is essential for valid analyses of income distribution (Emmenegger et al. 2023).

Furthermore, since the Tax Statistics represent a complete survey of all tax units, but not of all individuals, and do not contain any information on non-taxpayers, possible analyses are limited to the subpopulation of taxpayers. Accordingly, the Microcensus has to be limited to

Table 6.1: Comparison of the Tax Statistics (TS) and the Microcensus (MC)

Overview

Dataset	**Tax Statistics**	**Microcensus**
Source	tax authorities	official survey
Coverage	all taxpayers	1 % of the population
Observed units	tax units (individuals or married couples)	individuals in households
Sample size[a,b]	12,757,629 single tax units 10,355,771 married tax units	162,575 single tax units 105,936 married tax units

Advantages (+) and Shortcomings (−)

Dataset	**Tax Statistics**	**Microcensus**
Frame	+ full register of taxpayers	− 1 % sample
Observed units	− tax units	+ individuals in households
Unobserved units	− non taxpayers	− non-sampling elements
Income variables	− determined by tax law + continuous information	+ defined in survey design − classified in 24 classes
Income distribution	+ tails included	− shortcomings at the tails
Quality	+ very high quality	− self-response bias
Socio-demographic information	− limited to basic variables (like sex or age)	+ Variety of socio-demographic variables (like education and working time)
Spatial scale	+ fully exploitable at municipality level	− limited by sample size
Timeliness	− late availability (after 3 years)	+ yearly
Costs	+ automatic data transmission	− response burden

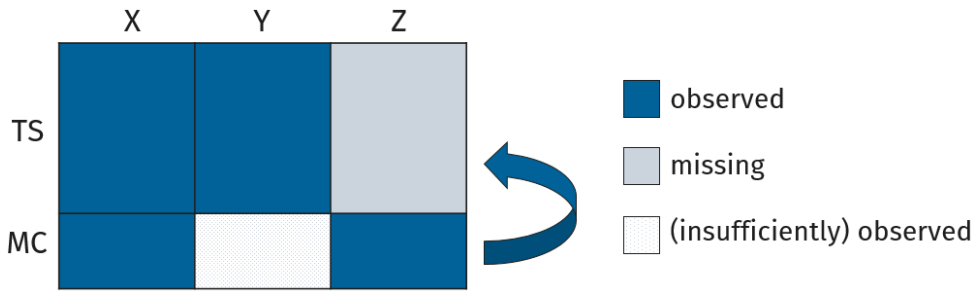
[a] Restricted to tax units of working age between 16 and 65 years with a (taxable) income > 0;
[b] Based on TS and MC from 2014.

Table closely based on Emmenegger et al. (2023).

the group of taxpayers, which is already apparent when looking at the specified sample sizes of both studies in Table 6.1. Analogously to Emmenegger et al. (2023), we define taxpayers in the Microcensus as individuals whose main source of income includes one of the following: (1) income from work, (2) assets, savings, dividends, renting or leasing, (3) pensions or

(4) wage-replacement benefits. Non-taxpayers, instead, are basically individuals whose main source of income comprises income from relatives (like parents or spouses) or from unemployment benefits. Furthermore, since our analysis objective is mainly based on methodological aspects regarding the comparison of different data fusion methods, we restrict both datasets to individual tax units (singles) of working age between 16 and 65 years for the sake of simplicity. In addition, the focus on individual tax units reduces the complexity regarding the composition of tax units and the aggregation of income due to the rules on joint taxation in Germany (Emmenegger et al. 2023).

By enhancing the Tax Statistics with the specific education and working time variables from the Microcensus the strengths of both datasets, the reliable income data from the Tax Statistics and the socio-demographic information from the Microcensus, are to be exploited more effectively. This data fusion use case is illustrated in Figure 6.1. While the high-quality income variable from the Tax Statistics represents the specific Y variable, education (Z_1) and working time (Z_2) reflect the specific $\mathbf{Z}$ variables from the Microcensus to be imputed within the Tax Statistics. As already indicated, the Microcensus also contains insufficient and rudimentary income information that was collected within 24 income classes and is likely to be subject to the survey-related problems mentioned above. This income information from the Microcensus will be used in the following by means of an appropriate strategy, which is useful for simulation and evaluation purposes to overcome and analyse possible problems of the CIA.



Source: Emmenegger et al. (2023).

Figure 6.1: Data Fusion Scenario of TS and MC

In contrast to the data fusion of EU-SILC and HBS from Chapter 5, it is apparent in Figure 6.1 that the significantly larger dataset, the Tax Statistics, serves as the recipient dataset, while the Microcensus represents the donor sample. Specifically, the sampling ratio for single working-age tax units with a taxable income greater than zero is $\frac{n_{MC}}{n_{TS}} = \frac{162{,}575}{12{,}757{,}629} \approx 0.013$. This seems realistic, since the Tax Statistics is a full survey, while the Microcensus is a 1 % sample of the population. Hence, with regard to the explicit scenarios of the donor-recipient ratio, a low donor

ratio can be observed according to Figure 6.1. Looking at the implicit data-related scenarios, it is clear that the summed sample size of the studies to be matched is significantly larger than $170,000$, which limits the available data fusion methods. Thus, as discussed in Section 3.4, DHD cannot be applied due to computational limitations and is therefore eliminated as a potential fusion method. Both aspects, the low donor ratio and the size of the datasets to be fused result in an atypical data fusion scenario, while the low donor ratio is not preferred from a methodological point of view, and the size of the datasets yields disadvantages in terms of computational effort. However, this is contrasted by the substantively justified motivation of ensuring high-quality income models. This emphasises the need to investigate potential data fusion procedures in different data situations in order to obtain an appropriate fusion result despite the sometimes atypical data constellation.

It should also be noted that we use different categories for the specific working time variable (full-time, part-time, non-working). With regard to the implicit scenarios, both $\mathbf{Z}$ variables to be fused, education and working time, therefore now have an ordered categorical scale level. This would in principle further limit the available data fusion methods, as PMM is an imputation method for metric variables only. However, since education and working time are ordered categorical variables and at least rank gradations are possible, PMM should nevertheless be implemented and evaluated in the upcoming simulations. In addition, unlike DHD, there are no computational reasons against the use of PMM, which is why PMM can in principle be considered as a method to be evaluated here. This will provide additional scientific insights into how PMM, an imputation method originally designed for metric variables, performs with ordered categorical variables.

To sum up, the Tax Statistics and the Microcensus reflect extremely large data sources, whereby in addition the smaller dataset, the Microcensus, is intended to serve as a donor file. The size of the data sources limits the available data fusion methods. Moreover, the variables to be fused, education and working time, have ordered categorical scale levels, whereas metric variables were imputed in the data fusion of EU-SILC and HBS. In the following section, the research design for the evaluation of the data fusion methods in the context of the data fusion of Tax Statistics and Microcensus is presented, again considering different scenarios.

6.2 Research Design

The research design consists of a Monte Carlo simulation as well as an empirical evaluation of the fused data file of the Tax Statistics and the Microcensus. Throughout this section we closely follow Emmenegger et al. (2023) and additionally incorporate further scenarios into the simulations. First, the interpolation of the income information from the Microcensus is described, as this is relevant for both the simulation study and the empirical evaluation. We continue with a description of the simulation database and the simulation design, incorporating the explicit scenarios in particular, before moving on to the design of the empirical evaluation.

6.2.1 Income Interpolation

This section draws from Emmenegger et al. (2023). As has already become clear, income in the Microcensus is merely available in classified form, which only provides inadequate income information. A metric variable, on the other hand, offers considerable information advantages due to better quantification. Therefore, it is useful for our purposes to transfer the classified income variable from the Microcensus into a continuous income variable.

This is done through the generalised Pareto interpolation according to Blanchet et al. (2022) by means of the `R` package `gpinter` (Blanchet 2018). Thereby, the Pareto parameters are based on frequencies from the 24 income classes observed in the Microcensus. In this respect, we rely on the generalised Pareto interpolated income from previous work by Emmenegger et al. (2023). An alternative could be linear interpolation. However, the generalised Pareto interpolation provides a more realistic income distribution that takes on a smooth and uninterrupted shape and allows for the most appropriate representation of high incomes obtainable from the Microcensus (Emmenegger et al. 2023).

Despite the presence of a metric variable, significant survey-related drawbacks of the MC income information remain. For example, a common criticism regarding income in the Microcensus is that respondents insufficiently include irregular income components and transfers in their reported income (Hochgürtel 2019). Accordingly, the income information suffers from under-coverage, which is particularly severe at the tails of the distribution, that is, at particularly low or high incomes (see Tab. 6.4). To counter this problem, Emmenegger and Münnich (2023) imputed top incomes of the Microcensus by means of the observed incomes from the Tax Statistics. Due to the quality concerns of the MC incomes, it seems particularly attractive to

use the Tax Statistics for reliable income analyses and to enhance them with socio-demographic variables from the Microcensus, which corresponds to the aim of the underlying data fusion application (Emmenegger et al. 2023).

In this respect, for the comprehensive evaluation of potential data fusion procedures, the presence of a metric MC income variable is useful for several reasons. On the one hand, Donatiello et al. (2016) already suggests including characteristics closely related to the specific $\mathbf{Y}$ and $\mathbf{Z}$ variables in the fusion process as a common variable in order to be able to mitigate potential problems of CIA. Thus, a metric MC income variable can be directly included in the fusion process as a common variable. Including or excluding the metric income variable as a common variable can also be used in this context to analyse effects with respect to the explicit scenarios of compliance or violation of the CIA. In addition, the metric income variable is also useful for the simulations, as the Microcensus can then serve as a surrogate population for a meaningful simulation study. Furthermore, the metric MC income variable is an important basis for the possibility of conducting at least a rough empirical evaluation. The next section continues with details on the database for simulation purposes.

6.2.2 Simulation Database

The Monte Carlo study is based only on the German Microcensus from the year 2014 and is restricted to individual working-aged tax units. Hence, the Microcensus as simulation database comprises $N_{MC} = 162,575$ observations (see Tab. 6.1). Here, all relevant variables $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ are observed, albeit with insufficient and imprecise income information, which we interpolated as described in the previous section. For simulation purposes, this database serves as surrogate population for the Tax Statistics, and from this surrogate TS study we draw $k = 1,000$ subsamples according to the Microcensus sampling procedure that in turn serve as proxy for the Microcensus. Therefore, analogously to the simulations from Chapter 5, all simulations are exclusively based on one dataset, here the Microcensus, in order to obtain meaningful benchmarks for the 'true' joint distribution of $\mathbf{Y}$ and $\mathbf{Z}$ as basis for the evaluations.

Analogously to Emmenegger et al. (2023), six variables observed in both the Tax Statistics and the Microcensus were identified as common $\mathbf{X}$ variables. These variables have already been harmonised by previous work of Emmenegger et al. (2023) in order to make them comparable between the Tax Statistics and the Microcensus for the implementation of the data fusion (for example, with regard to the same coding in both datasets). For simulation purposes, we also use these $p = 6$ common variables $X_1, \ldots, X_6$. Table 6.2 provides a respective overview of these

characteristics including information on the value range and the measurement level. In addition, we need to specify a substitute variable from the Microcensus to approximate the TS income variable (Y). Here, the interpolated income variable serves as a corresponding substitute. Note that the income variable within the Microcensus reports net income. For details on the income concepts of TS and MC, see Emmenegger et al. (2023). The specific **Z** variables from the Microcensus, in contrast, can be used directly for the simulations and do not have to be substituted. An overview of the specific variables (Y, Z_1, Z_2) including value range and measurement level is also included in Table 6.2.

The variable *employment status* (X_3) distinguishes between (1) employee, (2) self-employed, (3) civil servant and (4) other (Statistische Ämter der Länder 2014: 17-18). The *family type* (X_5) comprises combined information on marital status and the presence of own children (Statistische Ämter der Länder 2014: 8-9, 102). *Federal State* (X_6) includes all 16 federal states in Germany.

Table 6.2: Overview of Relevant Variables for Simulations, TS/ MC

	Variables	**Range / Scale Level**
X: **Common Variables**	X_1: Sex	1 to 2 / categorical
	X_2: Age	acc. X_2 / metric
	X_3: Employment Status	1 to 4 / categorical
	X_4: Number of Kids	acc. X_4 / metric
	X_5: Family Type	1 to 5 / categorical
	X_6: Federal State	1 to 16 / categorical
Y: **Sub. TS Variables**	Y: Generalised Pareto Interpolated Income	acc. Y / metric
Z: **MC Variables**	Z_1: Education Level (low, middle, high)	1 to 3 / categorical
	Z_2: Working Time (full-time, part-time, none)	1 to 3 / categorical

Source: Microcensus (2014).

The three-scaled *education level* (Z_1) comprises the levels (1) low, (2) middle and (3) high and is based on the OECD definition of levels of education with the following modifications: 'low' reflects the education below upper secondary (ISCED levels 0, 1, 2), 'middle' corresponds to upper secondary or post-secondary non-tertiary education (ISCED levels 3A, 3B, 3C) and 'high' represents tertiary education (ISCED levels 4, 5A, 5B, 6) (OECD 2014). We chose a three-scaled education variable because most of the income variation can be explained by these

three categories, which is therefore a typical choice in the literature (see e.g. Flood et al. 2008). All three education levels reflect the highest educational attainment as suggested by Mincer (1958). *Working time* (Z_2) is also a three-scaled variable representing (1) full-time, (2) part-time and (3) non-working (Emmenegger et al. 2023).

With regard to the underlying database, the correlation structure between the common **X** variables and the specific **Z** variables is also of interest, as this could provide initial indications as to whether the CIA applies. In this respect, Table 6.3 shows the corresponding matrix of associations. Depending on the scale level, we use Cramer's V for associations between unordered categorical **X** variables and **Z**, while Kendall's τ is applied for associations between metric **X** variables and **Z**. The associations in Table 6.3 predominantly show low associations between **X** and **Z** of less than 0.1, except for associations between sex (X_1) and working hours (Z_2), between age (X_2) and working hours (Z_2), between family type (X_5) and working time (Z_2) and between employment status (X_3) and both specific **Z** variables. However, in contrast to the correlation structure between **X** and **Z** for the data fusion use case of EU-SILC and HBS (see Sec. 5.2.1 and Tab. 5.2), there is no **X** variable here that is extremely highly correlated with Z_1 or Z_2, which could indicate an increased violation of the CIA. This will also be evident later in the corresponding benchmarks of the simulations. In this respect, the inclusion of the interpolated income information from the Microcensus in the data fusion process seems to make sense in order to overcome the CIA, which also allows the CIA-related explicit scenarios to be examined.

Table 6.3: Associations Between **X** and **Z**, TS/ MC

	Z_1	Z_2
X_1: Sex [a]	0.0627	0.2435
X_2: Age [b]	0.0203	0.1104
X_3: Employment Status [a]	0.1738	0.6954
X_4: Number of Kids [b]	−0.0918	0.0010
X_5: Family Type [a]	0.0327	0.1390
X_6: Federal State [a]	0.0910	0.0532

[a] Cramer's V;
[b] Kendall's τ.

Source: Microcensus (2014).

6.2.3 Monte Carlo Study

The Monte Carlo study based on the Microcensus data just described closely follow Emmenegger et al. (2023). While the simulations of Emmenegger et al. (2023) already cover the explicit CIA-related scenarios, we additionally extend the simulation design to include different donor-recipient ratios. First, we provide details on the sampling procedure for each of the $k = 1,000$ simulation runs to replicate the original data fusion use case of matching TS with MC. Subsequently, further scenarios are incorporated into the simulation procedure and the evaluation criteria are specified.

To simulate the original data fusion use case of TS and MC, as already indicated in the previous section, the Microcensus represents the surrogate population and thus could serve as proxy for the Tax Statistics. However, instead of using the whole Microcensus as substitute for the TS, as in Emmenegger et al. (2023), we draw subsamples from the database that represent the TS substitutes, which is useful for the later incorporation of the high donor ratio scenario. From this TS proxies, small Microcensuses are drawn with the original Microcensus sampling procedure.

Hence, the sampling procedure for each of the $k = 1,000$ simulation runs is as follows: First, from the MC data we draw a subsample (without replacement) considered as the TS substitute that represents two thirds (≈ 67 %) of the database, which leads to a sample size of $n = 108,925$ for the TS substitute. This dataset now represents the Tax Statistics. From this TS proxy comprising $n = 108,925$ observations, we further draw a small Microcensus according to the concrete sampling design of the German Microcensus. This sampling procedure equals a one-stage cluster sampling (Statistisches Bundesamt 2015) and involves the following: First, different clusters are specified, which consist of combined information of a building size category and regional information, which in our case is based on federal states. Subsequently, we randomly draw 1 % of the selection districts (called *Auswahlbezirke* in German) within each of the specified clusters. Finally, all observation units within the drawn selection districts form the final subsample representing the small Microcensus, and this subsample comprises about 1 % of the TS substitute. This small Microcensus now reflects the MC proxy. Therefore, within each simulation run, we obtain a TS proxy from which we delete the education and working time information, and a small Microcensus reflecting the MC proxy that contains these socio-demographic information. Hence, the concrete data fusion scenario of TS and MC (see Fig. 6.1) is artificially generated. For each of the $k = 1,000$ simulation runs, we then impute the missing education and working time information within the TS substitute by means of the data fusion algorithms from Chapters 3 and 4 with the exception of the DHD method. Analogously to the simulations

in Chapter 5, single imputation ($M = 1$) is applied (Emmenegger et al. 2023).

Thus, it became clear that we obtain a sampling ratio of $\frac{n_{MC}}{n_{TS}} \approx 0.01$ for the TS and MC substitutes. This corresponds to a scenario with a quite low donor ratio and simulates the original data fusion of TS and MC. To stay consistent with the simulations from Chapter 5, we refer to this low donor-recipient ratio as n_2. However, in order to investigate the explicit scenarios of different donor-recipient ratios, we additionally aim to consider a high donor ratio. To simulate a high donor ratio between the recipient and the donor study, we consider the two-thirds sample as the donor study and the small Microcensus as the recipient study, which serves comparability with the scenario of a low donor ratio. This yields a donor pool that is 100 times larger than the recipient dataset, which we refer to as n_1. Therefore, the explicit scenarios of a high or low donor-recipient ratio can be investigated by running the simulation study both with $n_1 = \frac{n_{MC}}{n_{TS}} \approx 100$ and $n_2 = \frac{n_{MC}}{n_{TS}} \approx 0.01$. Note that the prior drawing of a two-thirds sample within each simulation run for n_1 ensures that a random component is included, otherwise each model from the donor data would look the same and contain no variation (which is not desirable in simulation studies). To ensure comparability with the scenario of a low donor ratio, a two-thirds sample is also drawn from the database for the TS proxy for n_2, as mentioned above.

The explicit scenarios of compliance or violation of the CIA are also to be incorporated into the simulations in order to investigate CIA-related effects, which we do closely following Emmenegger et al. (2023). As already indicated, the CIA is violated if solely the common $\mathbf{X}$ variables listed in Table 6.2 are included in the fusion process. This can be shown using the simulation database by calculating the regression parameters that would result if the CIA holds (see Sec. 6.3.2 and Tab. 6.7). To circumvent the CIA and thus create a scenario in which the CIA is fulfilled, the rudimentary income information form the Microcensus is to be exploited by including the generalised Pareto interpolated income as a common variable in the data fusion process. This approach closely follows Donatiello et al. (2016) as they propose to include variables that are closely related to the specific $\mathbf{Y}$ and $\mathbf{Z}$ variables in the fusion process. Since the Microcensus comprises rudimentary income information, the CIA can be circumvented by adequately including this income information in the data fusion process, which is ensured through the interpolated income. Accordingly, for both n_1 and n_2 we conduct the simulation study twice: Once using income as a common variable, thereby satisfying the CIA, and once excluding income as a common fusion variable, which induces a violation of the CIA (Emmenegger et al. 2023).

In order to evaluate the performance of the respective data fusion methods using applications

that are as close to reality as possible, the evaluations are based on regression parameters β of education and working hours obtained from exemplary regression models on income. Besides education and working time, it is essential to ensure that the regression models include a subset of the common $\mathbf{X}$ variables in order to avoid *uncongeniality* issues (see Meng 1994; Xie and Meng 2017). Therefore, in addition to education and working time, the exemplary regression models on income include the variables sex, age, number of kids, family type and employment status, while for employment status only dummies on self-employed and civil servants are considered. After each simulation run, the income models are calculated on the recipient study including the education and working time variables imputed by the respective data fusion algorithms. From this, we obtain education and working time parameters $\widehat{\beta}_{educ}$ and $\widehat{\beta}_{w-time}$ for each data fusion method. These parameters are then compared to the 'true' quantities of interest, that is, the education and working time parameters obtained from income models estimated on the Microcensus database specified in the previous section. This allows us to evaluate to what extent the respective data fusion algorithms can reproduce the benchmark regression parameters from possible income models (Emmenegger et al. 2023).

In addition, we also briefly consider results of the univariate distribution of the imputed education and working time variables for further discussion of the results. For this, we show the Monte Carlo means of the relative frequencies of $\widetilde{Z}_{educ}$ and $\widetilde{Z}_{w-time}$ for each method.

The MC simulation is based on the same programme basis and essentially on the same `R` packages as the simulations from the previous chapter (see Sec. 5.2.2). Instead of the `lm()` function, however, the package `nnet` (Ripley and Venables 2022) is now used for RM due to the categorical scale level of $\mathbf{Z}$. In addition, `mice` (van Buuren 2022) is now used for PMM. However, `mice` does not comprise a multivariate PMM imputation, while the advantage of `BaBooN` (Meinfelder and Schnapp 2015) is the availability of such a multivariate PMM solution. Due to the size of the datasets and the associated computational aspects related to the runtime, `mice` is more useful here, which is why we use `mice` for PMM in the upcoming simulations. Hence, no differences between the univariate and multivariate imputation solutions can be examined with regard to the imputation scenarios for PMM. Although multivariate imputation has some advantages in practice, no substantial differences between univariate and multivariate PMM should be expected in terms of reproducing joint distributions between $\mathbf{Y}$ and $\mathbf{Z}$. This is also underlined by the results from Chapter 5, which is why the exclusively univariate consideration of PMM can be regarded as acceptable. However, the developed PVM function allows both univariate and multivariate imputation, which is why both imputation scenarios can be considered for PVM. Again, we basically present the results for the multivariate imputation solution, while Appendix B.2 includes

the results for univariate PVM.

To sum up, the specified MC simulation is based on the Microcensus only, while a simulation design was introduced that replicate the original data fusion use case of matching TS and MC as realistically as possible. The simulation design also comprises further manipulations to cover the explicit scenarios with regard to the CIA and the donor-recipient ratio. We base our evaluations on exemplary regression models on income in order to stay close to the analysis objective of income modelling.

6.2.4 Empirical Evaluation

In addition to the simulation study, the underlying data situation of TS and MC suggests at least a rough empirical evaluation. In this respect, the concept of the empirical evaluation and the related descriptions in this section closely follow Emmenegger et al. (2023). The empirical evaluation first involves implementing the data fusion on the real TS and MC datasets from 2014. In other words, the Tax Statistics is enhanced by the socio-demographic variables educational attainment and working time from the Microcensus using the underlying data fusion procedures RM, PMM, DT, PVM-DT, RF and PVM-RF. The programme basis and the packages used are identical to those specified in the previous section.

The empirical evaluation is two-fold: On the one hand, the tax data income medians conditional on the six education and working time variables are compared to the conditional income medians observed in the Microcensus. This is done on a regional level of the federal states and corresponds to an evaluation based on the second validity level. On the other hand, with regard to the data fusion to improve income modelling, it would have to be critically questioned whether adding the education and working time variables actually improves the income models based on the enhanced Tax Statistics. This is because certain model variables included in both datasets, such as age and employment status, probably already capture parts of the explanatory power of the education and working time characteristics. Thus, the question arises about the actual need of including education and working time within the TS income models, given the fact that additional uncertainty is caused by the data fusion process. In this respect, an evaluation of the adjusted R^2 seems an appropriate measure to illustrate potential model improvements. Therefore, the adjusted R^2 of income models with and without the six resulting education and working time variables within the Tax Statistics is compared to the exemplary income models with and without education and working time from the Microcensus (Emmenegger et al. 2023).

Additionally and analogously to the simulation study, a brief look at the marginal distribution of the imputed variables $\widetilde{Z}_{educ}$ and $\widetilde{Z}_{w-time}$ seems useful. Hence, we compare the marginal distribution of $\widetilde{Z}_{educ}$ and $\widetilde{Z}_{w-time}$ from the enhanced Tax Statistics to those obtained from the Microcensus, that is, the donor data file.

This empirical evaluation is conducted with and without using the interpolated MC income characteristic as a common fusion variable. Thus, the empirical application covers at least the explicit CIA-related scenarios. An investigation of the explicit scenarios of a high or low donor-recipient ratio is not useful here, since the size of the datasets and the respective sampling ratios are given by the data situation and can only be appropriately manipulated within simulation studies.

Note that the income concepts differ between the Tax Statistics and the Microcensus (for details see Emmenegger et al. 2023). While the Microcensus comprises monthly net income, the Tax Statistics reports yearly taxable income defined by the German tax law, which is not directly comparable to the MC net income. Therefore, Emmenegger et al. (2023) already transformed the relevant taxable income information from the TS to a monthly net income variable in order to ensure comparability between TS and MC income information. Concerning the TS, we rely on this new TS income variable for the purpose of the empirical application and evaluation.

However, as has already become clear, the interpolated income distribution from the Microcensus still deviates from the observed TS incomes caused by under-coverage and under-reportage in the survey data, which is particularly problematic for a meaningful evaluation of the conditional income medians. We therefore aim for an appropriate level of comparability between both datasets. Hence, for the initially biased income information in the Microcensus, we apply reweighting methods (as described in the next section) in order to adjust the income distribution of the Microcensus to that of the Tax Statistics (Emmenegger et al. 2023).

Note that the empirical evaluation of the joint distribution of $\mathbf{Y}$ and $\mathbf{Z}$ is typically not possible in common data fusion applications, since no information regarding their joint distributions is available. However, as income is at least roughly observed, the income interpolation as well as the reweighting procedure described in the next section ensure a comparability between the two datasets that can be considered acceptable. Thus, at least rough insights on the performance of the data fusion procedures in real applications can be gained (Emmenegger et al. 2023).

6.2.5 Reweighting

This section closely follows Emmenegger et al. (2023). As already indicated, we apply reweighting methods in order to adjust the income distribution of the Microcensus to that of the Tax Statistics. In this respect, Table 6.4 displays the income distribution within TS and MC according to the 24 MC income classes. The third column (no. of obs. TS) represents the number of observations from the Tax Statistics that belong to the respective income classes and thus provide the benchmark for the reweighting. The fourth column (extr. pop. MC) contains the extrapolated population from the 1 % Microcensus sample for each of the income classes. Note that the extrapolated MC population may exceed the number of observations from the TS in some income classes due to extrapolation. The coverage, which is the fifth column, is given by $\frac{Extr.pop.MC}{No.obs.TS}$. According to Table 6.4, especially the middle of the income distribution between 300 and 2,000 Euro are over-coveraged by the MC, while the bottom and the top of the distributions, that is, the lower and higher incomes, are underrepresented (Emmenegger et al. 2023).

To adjust the incomes from the Microcensus to those from the Tax Statistics, weight calibration techniques of Deville and Särndal (1992) are applied, where the income weights within the Microcensus are corrected to represent the German taxpayer population. For this purpose, a minimum distance criterion is used that minimises the sum of the differences between the original and the corrected weights. Linear distance functions that are not restricted to a certain range are used for this. Ultimately, with this well-studied calibration technique, we obtain weights that are positive for all observations within the Microcensus. These calibrated weights range from 0.0014 to 1.1672 and therefore do not need to be adjusted further. Note that Brzezinski et al. (2019) employ a similar approach with regard to the Polish taxpayer population (Emmenegger et al. 2023).

For the evaluation purposes, we rely on the interpolated and reweighted MC income variable already created by Emmenegger et al. (2023). The reweighted, interpolated income distribution within the Microcensus now represents exactly the same (unconditional) quantiles as the observed incomes from the Tax Statistics. This underlines the usefulness of this approach to achieve a meaningful empirical evaluation (Emmenegger et al. 2023).

Table 6.4: Comparison of Income Frequencies in TS and MC 2014

Income class	Income range[a]			No. of obs. TS[b]	Extr. pop. MC[c]	Coverage
0		equal to	0	199,777	47,847	0.24
1	0	to under	150	416,230	96,361	0.23
2	150	to under	300	306,416	232,263	0.76
3	300	to under	500	418,994	884,733	2.11
4	500	to under	700	468,755	1,403,315	2.99
5	700	to under	900	593,790	1,975,034	3.33
6	900	to under	1,100	866,782	2,179,991	2.52
7	1,100	to under	1,300	900,618	2,507,952	2.78
8	1,300	to under	1,500	902,285	2,367,916	2.62
9	1,500	to under	1,700	938,785	2,157,226	2.30
10	1,700	to under	2,000	1,439,555	2,429,688	1.69
11	2,000	to under	2,300	1,449,274	1,694,422	1.17
12	2,300	to under	2,600	1,370,980	1,070,314	0.78
13	2,600	to under	2,900	1,102,155	590,685	0.54
14	2,900	to under	3,200	858,737	514,595	0.60
15	3,200	to under	3,600	857,596	360,813	0.42
16	3,600	to under	4,000	578,167	201,573	0.35
17	4,000	to under	4,500	456,458	163,786	0.36
18	4,500	to under	5,000	274,605	100,364	0.37
19	5,000	to under	5,500	171,457	65,843	0.38
20	5,500	to under	6,000	112,767	40,847	0.36
21	6,000	to under	7,500	171,031	54,608	0.32
22	7,500	to under	10,000	107,061	43,072	0.40
23	10,000	to under	18,000	82,302	27,714	0.34
24		over	18,000	31,788	13,190	0.41

[a] In Euro;
[b] Number of observations from the Tax Statistics;
[c] Extrapolated population from the Microcensus.
Source: Emmenegger et al. (2023), based on Microcensus (2014) and Tax Statistics (2014).

6.3 Simulation Results

We now present the results of the simulation study, first for the case where income is included as a common variable in the data fusion process, thus fulfilling the CIA. The results are presented both for the high and the low donor ratio scenarios. We then discuss the simulation results when income is excluded as a common variable, thus violating the CIA, also for varying sample sizes. All relevant tables containing the exact values of the graphical diagnostics from this section can be found in Appendix A.2. As already pointed out, all results from this section and from Appendix A.2 comprise multivariate imputation solutions, while Appendix B.2 includes the results for univariate PVM. Note that the simulations carried out are partly built on Emmenegger

et al. (2023). However, the simulations presented here are extended to PVM-DT and PVM-RF and to the explicit scenarios of varying sample sizes of the recipient and the donor data, while also providing benchmarks for theoretical CIA parameters and additional diagnostics concerning the marginal distribution.

6.3.1 CIA Compliance

With regard to the results under CIA compliance where income is included in the data fusion process, Table 6.5 includes the respective education and working time parameters from the exemplary income models based on the simulation database. These 'true' parameters serve as benchmark for evaluation purposes. Here, for educational attainment we obtain parameters of 0.33 and 0.61 for the middle and high education categories, while low education serves as reference category. For working time, the benchmark parameters amount to -0.44 and -0.63 for part-time and non-working, with full-time representing the reference category. Since the generalised Pareto interpolated income serve as common variable and simultaneously represents the specific Y variable from the recipient data file, the CIA is fulfilled, which is why no information on theoretical CIA parameters is required.

Table 6.5: Benchmark Parameters for β_{educ} and β_{w-time} under CIA Compliance

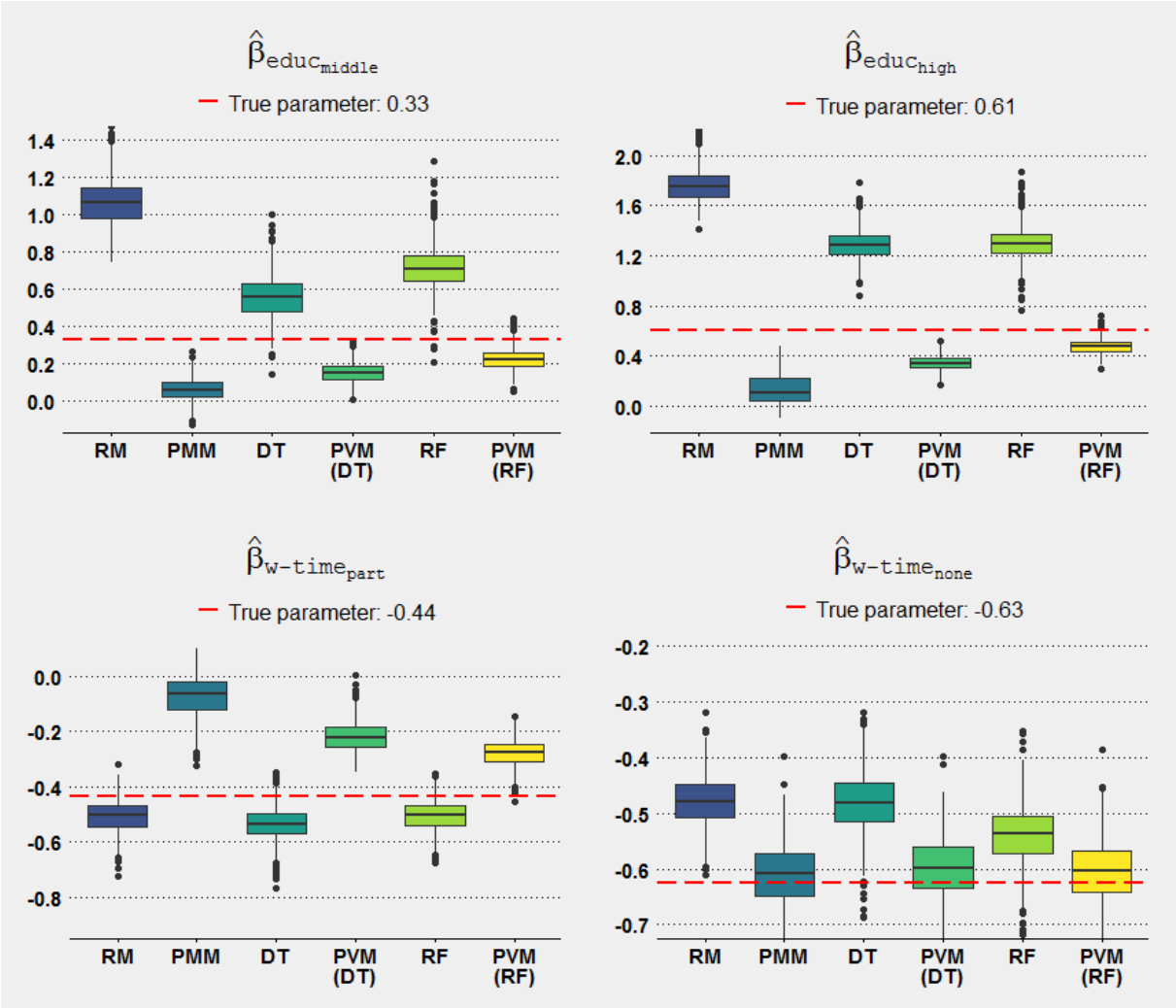
β_{educ_m}	β_{educ_h}	β_{w-time_p}	β_{w-time_n}
0.3277	0.6059	-0.4357	-0.6255

Source: Microcensus (2014).

The simulation results can now be used to assess the extent to which the potential data fusion methods are able to reproduce the corresponding education and working time parameters. In this respect, Figure 6.2 illustrates the Monte Carlo distributions of the $k = 1,000$ simulation runs for each method under n_1, that is, under a donor pool that is about 100 times higher than the number of recipient observations. In the upper part of Figure 6.2, we see the MC distributions for the education parameters, while the lower part depicts those of the working time coefficients. The horizontal red line represents the corresponding benchmark parameters from the simulation database.

As apparent in the upper part of Figure 6.2, RM, DT and RF tend to overestimate the education parameters, which is especially striking for RM. RM produces mean coefficients of 1.07 for $\widehat{\beta}_{educ_m}$ and 1.76 for $\widehat{\beta}_{educ_h}$, which are thus biased on average with $1.07 - 0.33 = 0.74$ and

$1.76 - 0.61 = 1.15$, respectively (see Tab. A.9). Since the Root Mean Squared Error (RMSE)[1] brings together the diagnostics on bias and variance, we again consider the RMSE as the most appropriate indicator for overall performance evaluation. In this respect, by far the highest RMSE values with regard to $\widehat{\beta}_{educ}$ are observed for RM, as illustrated in the left side of Figure 6.3. The exact RMSE values are shown in Table A.10. Lower overestimations are found for DT and RF. PMM, PVM-DT and PVM-RF in turn seem to underestimate the education parameters, while PVM-RF yields the best performance and results in the lowest RMSE values for $\widehat{\beta}_{educ_m}$ and $\widehat{\beta}_{educ_h}$ (0.12 and 0.14) under n_1. The education parameters resulting from PMM, on the other hand, are largely biased towards 0.



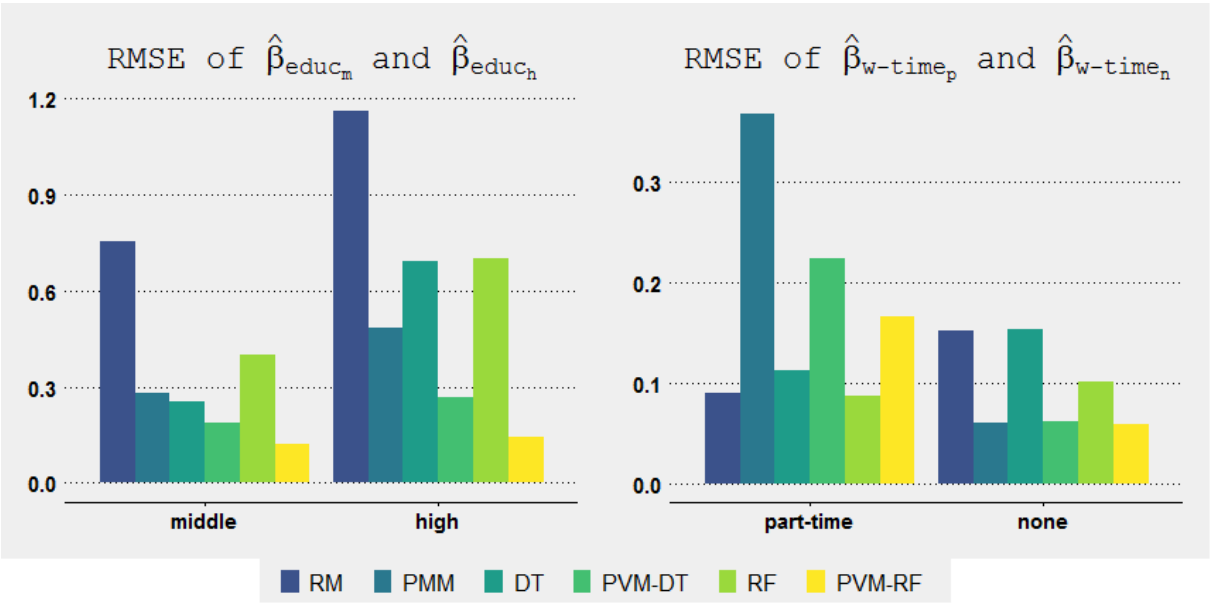
Source: Microcensus (2014).

Figure 6.2: MC distributions for $\widehat{\beta}_{educ}$ and $\widehat{\beta}_{w-time}$ with n_1 under CIA Compliance

For the working time parameters under n_1 (lower part of Fig. 6.2 and right side of Fig 6.3)

[1] $\text{RMSE}(\widehat{\beta}) = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(\widehat{\beta}_i - \beta)^2}$

we see the same effects for the part-time parameter that we already obtained for the education parameters, but now more in favour of RM, DT and RF. While PMM, PVM-DT and PVM-RF again underestimate the part-time parameter, RM, DT and RF seem to overestimate it. However, lower exaggerations for RM, DT and RF are found compared to the underestimations of PMM, PVM-DT and PVM-RF. The best performance under n_1 with regard to $\widehat{\beta}_{w\text{-}time_p}$ is obtained by RF according to the RMSE values, with RM and DT performing similarly well. RF produces mean estimates of -0.51 that come close to the benchmark value of -0.44, and the RF parameters for $\widehat{\beta}_{w\text{-}time_p}$ also provide the lowest RMSE value, which amount to 0.09. The RMSE values for PVM-DT and PVM-RF are about two times higher compared to those of DT and RF, respectively, as shown in Figure 6.3 and Table A.10. In contrast, for the non-working time parameter we obtain rather opposite effects. RM, DT and RF tend to slightly underestimate the true parameter of -0.63, while the estimates for $\widehat{\beta}_{w\text{-}time_n}$ resulting from PMM, PVM-DT and PVM-RF are on average quite close to the original non-working time parameter. The better performance of RM, DT and RF for $\widehat{\beta}_{w\text{-}time_p}$ thus seems to be somewhat at the expense of their performance for $\widehat{\beta}_{w\text{-}time_n}$. Conversely, this can also be stated for PMM, PVM-DT and PVM-RF, as the better performance for $\widehat{\beta}_{w\text{-}time_n}$ seems to be slightly at the cost of their performance for $\widehat{\beta}_{w\text{-}time_p}$.

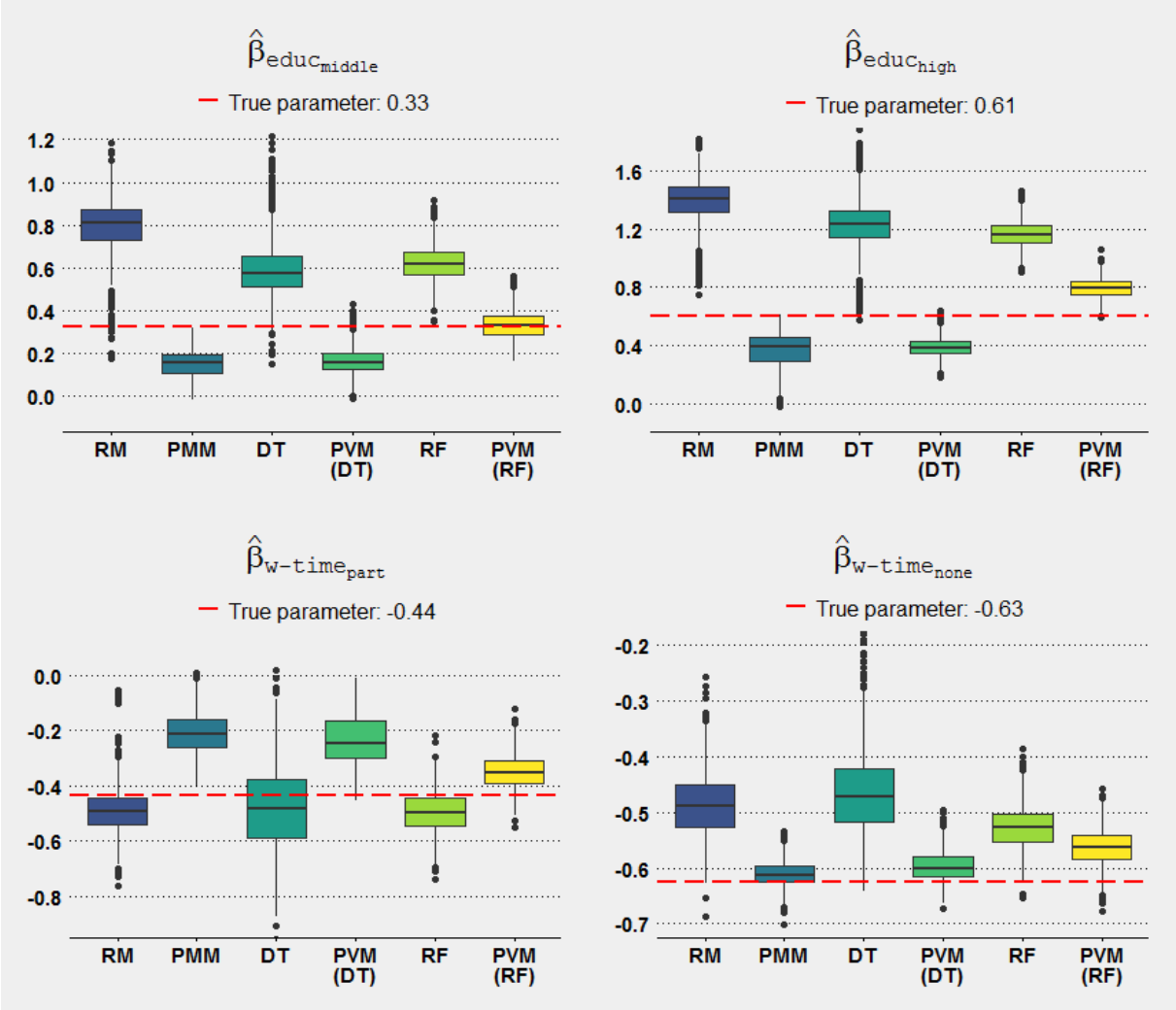


Source: Microcensus (2014).

Figure 6.3: RMSE of $\widehat{\beta}_{educ}$ and $\widehat{\beta}_{w-time}$ with n_1 under CIA Compliance

The results for the low donor ratio scenario under n_2, where the donor pool includes only 1 % compared to the number of recipient observations, are illustrated in Figures 6.4 and 6.5. For the education parameters, compared to n_1, we see somewhat lower exaggerations for RM and

RF, and at the same time lower underestimations of PMM and PVM-RF, while PVM-RF now reproduce the middle education parameter quite well on average, but tends to overestimate the high education parameter. The changes for DT and PVM-DT compared to n_1 are lower. DT tend to produce slightly higher parameters for $\widehat{\beta}_{educ_m}$ under n_2, while PVM-DT performs marginally better under n_2 in terms of the high education coefficient. The overall best performance under n_2 according to the RMSE values is, analogously to n_1, obtained by PVM-RF for both education parameters.

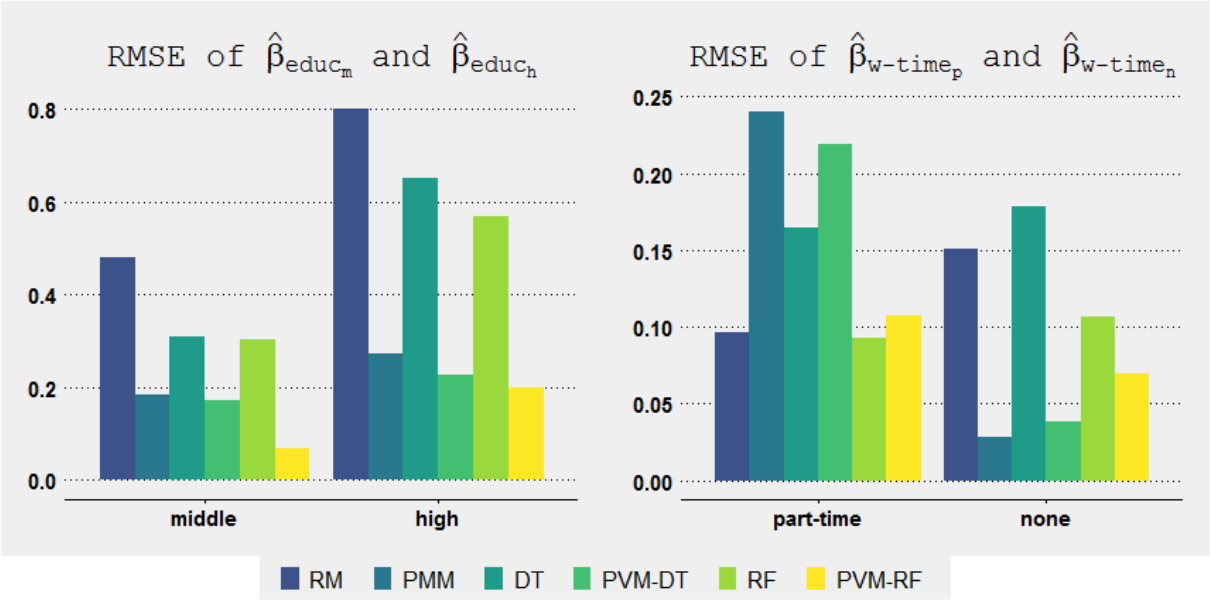


Source: Microcensus (2014).

Figure 6.4: MC distributions for $\widehat{\beta}_{educ}$ and $\widehat{\beta}_{w-time}$ with n_2 under CIA Compliance

With regard to the working time parameters, PMM, PVM-DT and PVM-RF again yield at least for $\widehat{\beta}_{w\text{-}time_p}$ lower underestimations, while the resulting parameters from RM, DT and RF for $\widehat{\beta}_{w\text{-}time_p}$ are on average slightly less overestimated compared to n_1. However, for RM, DT and RF only marginal changes are to be found with respect to $\widehat{\beta}_{w\text{-}time_p}$. PMM now shows a substantial improvement over n_1, which is particularly striking for $\widehat{\beta}_{w\text{-}time_p}$. Nevertheless, the RMSE

value for PMM is still highest with respect to $\widehat{\beta}_{w\text{-}time_p}$. For the non-working time parameter, no substantial changes are to be found under n_2 compared to n_1, but it can be stated that the boxes for PMM, PVM-DT and PVM-RF are narrower under n_2 for $\widehat{\beta}_{w\text{-}time_n}$, thus indicating a lower variance. For RM and DT, in contrast, the variance seems to increase under n_2, which is especially striking for DT with regard to the part-time coefficients. The overall best performance for the working time coefficients under n_2 can generally be observed for PVM-RF, as PVM-RF yields low RMSEs for both part-time and non-working coefficients, while all other methods only produce low RMSEs for one of the working time parameters that are higher at the other coefficient.



Source: Microcensus (2014).

Figure 6.5: RMSE of $\widehat{\beta}_{educ}$ and $\widehat{\beta}_{w-time}$ with n_2 under CIA Compliance

Overall, both learning-based nearest neighbour approaches, PVM-DT and PVM-RF, tend to be superior to the other fusion methods under the CIA compliance scenario. This is particularly observable for the education coefficients. For the working time parameters, the results are less clear in favour of PVM-DT and PVM-RF. Especially for the part-time coefficient, the performance of the prediction methods RM and RF is slightly better.

These findings can be further classified and discussed by looking at the univariate distribution of $\widetilde{Z}_{educ}$ and $\widetilde{Z}_{w-time}$ resulting from the respective simulation runs for each method. Table 6.6 shows the Monte Carlo means of the relative frequencies of the education and working time categories for each fusion method under n_1 and n_2 compared to the benchmark distribution of $\mathbf{Z}$ of the underlying database. In this respect, the effect already discussed in Section 4.5 is striking

that the corresponding SL methods DT and RF predominantly impute the mode category. This also applies to RM, which is also a predictive method, but is subject to distributional assumptions. For education, the predominant imputation of the mode category (medium education) is at the expense of the other categories (low and high education) for RM, DT and RF. For working time, on the other hand, part-time and non-working have similarly low relative frequencies in the MC database. Here, the predominant imputation of the mode category (full-time) of RM, DT and RF seems to be at the expense of the part-time category. This is probably due to the fact that non-working can be explained very well by the 'other' category of employment status (X_3), which predominantly includes non-employed individuals. PMM, PVM-DT and PVM-RF, on the other hand, match the marginal distribution of $\mathbf{Z}$ relatively well. At least for the education variable, under CIA fulfilment, this is consistent with the fact that PVM-DT and PVM-RF, in contrast to RM, DT and RF, better reproduce the joint distribution between income and education, while RM, DT and RF significantly overestimate the parameters. PMM yields good results in terms of marginal distributions, but relatively poor results for the joint distributions. With regard to the working time variable, the concentration on the full-time category and the fact that a common variable (X_3) can almost perfectly explain one of the smaller categories, non-working time, seems to be a challenging data situation.

Furthermore, the overall sensitivity to a high or low donor ratio is moderate for all methods in the CIA compliance scenario. However, it seems that in this case some data fusion approaches rather tend to profit from a lower donor ratio. This was especially apparent for RM, RF, PVM-DT and PVM-RF with respect to the education parameters, and partly also for the working time coefficients. For PMM, the high donor ratio could further compromise the probably misspecified model, since PMM actually requires a metric scale level for $\mathbf{Z}$. This source of error is then compounded by a higher number of donor observations underlying PMM. Generally, the higher the number of available donors, the more method-specific artefacts such as exaggerations, underestimations or misspecifications seem to occur in the CIA compliance scenario. Nevertheless, methods with good performance in the CIA compliance scenario, such as PVM-RF (at least for three of the four considered parameters) provide acceptable results for both high and low donor ratios. The next section is devoted to the scenario of CIA violation.

Table 6.6: Relative Frequencies of $\widetilde{Z}_{educ}$ and $\widetilde{Z}_{w-time}$ under CIA Compliance

		Education			**Working Time**		
		low	middle	high	none	part-time	full-time
$\mathbf{n_1}$	**RM**	0.04	0.70	0.26	0.13	0.07	0.80
	PMM	0.14	0.50	0.36	0.13	0.16	0.71
	DT	0.04	0.72	0.24	0.13	0.06	0.81
	PVM-DT	0.14	0.50	0.36	0.13	0.16	0.71
	RF	0.05	0.65	0.30	0.13	0.10	0.77
	PVM-RF	0.14	0.49	0.37	0.13	0.16	0.71
$\mathbf{n_2}$	**RM**	0.06	0.66	0.28	0.13	0.08	0.79
	PMM	0.15	0.49	0.37	0.12	0.17	0.71
	DT	0.05	0.67	0.28	0.13	0.07	0.80
	PVM-DT	0.14	0.49	0.37	0.13	0.16	0.71
	RF	0.06	0.63	0.31	0.13	0.09	0.78
	PVM-RF	0.10	0.57	0.34	0.13	0.13	0.74
	MC Database	0.14	0.49	0.37	0.13	0.16	0.71

Note that all values under n_1 and n_2 for each method reflect the Monte Carlo means of the relative frequencies from all $k = 1,000$ simulation runs. The distribution of the MC database serves as benchmark.

Source: Microcensus (2014).

6.3.2 CIA Violation

Excluding income as a common variable from the data fusion process, Table 6.3 already suggests that the CIA is a bold assumption here. This is also apparent in Table 6.7, which includes the benchmark parameters from exemplary income models (that remain equal to those of Tab. 6.5) as well as the theoretical regression coefficients that would result if **Y** and **Z** were truly independent given **X**.

However, the computation of the theoretical regression parameters under CIA is not as trivial as in the simulation of Chapter 5, where Equation (5.1) according to Rässler (2002: 36) could be used directly to obtain proxies for the CIA correlations assuming partial uncorrelation. In order to calculate the respective CIA parameters, we make use of the fact that regression parameters can be computed via covariances. This first involves to apply Equation (5.1) to obtain the

covariance matrix that would result if the CIA holds. Thus, we obtain a covariance matrix

$$\Sigma_{\text{CIA}} = \begin{pmatrix} \Sigma_{\mathbf{XX}} & \Sigma_{\mathbf{XY}} & \Sigma_{\mathbf{XZ}} \\ \Sigma_{\mathbf{YX}} & \Sigma_{\mathbf{YY}} & \widetilde{\Sigma}_{\mathbf{YZ}} \\ \Sigma_{\mathbf{ZX}} & \widetilde{\Sigma}_{\mathbf{ZY}} & \Sigma_{\mathbf{ZZ}} \end{pmatrix}, \tag{6.1}$$

where $\widetilde{\Sigma}_{\mathbf{YZ}}$ and $\widetilde{\Sigma}_{\mathbf{ZY}}$ represent the covariance matrices under CIA calculated by Equation (5.1). Note that all elements in Σ_{CIA} from (6.1) are themselves covariance matrices again. For example, in this case $\Sigma_{\mathbf{XX}}$ is a 25×25-dimensional matrix because it reflects a model matrix (to account for the categorical variables) of all common $\mathbf{X}$ variables. This covariance matrix under CIA from (6.1) is now reduced to the relevant variables for the exemplary regression models, denoted as $\mathbf{X}_m$, whereby the $\mathbf{Z}$ variables are now also counted among these independent variables $\mathbf{X}_m$ for the income models. $\mathbf{Y}_m = Y$ reflects the generalised Pareto interpolated income as dependent variable. Thus, for the exemplary regression models we obtain the CIA covariance matrix

$$\Sigma_{\text{CIA},m} = \begin{pmatrix} \Sigma_{\mathbf{X}_m\mathbf{X}_m} & \Sigma_{\mathbf{X}_m\mathbf{Y}_m} \\ \Sigma_{\mathbf{Y}_m\mathbf{X}_m} & \Sigma_{\mathbf{Y}_m\mathbf{Y}_m} \end{pmatrix}. \tag{6.2}$$

Finally, the regression parameters under CIA are computed via

$$\beta_{\text{CIA}} = \Sigma_{\mathbf{Y}_m\mathbf{X}_m} \cdot \Sigma^{-1}_{\mathbf{X}_m\mathbf{X}_m}, \tag{6.3}$$

and the lower part of Table 6.7 shows the resulting CIA parameters. Note that the respective CIA parameters are again subject to the assumption of partial uncorrelation, as there is no normal distribution for the variables of interest.

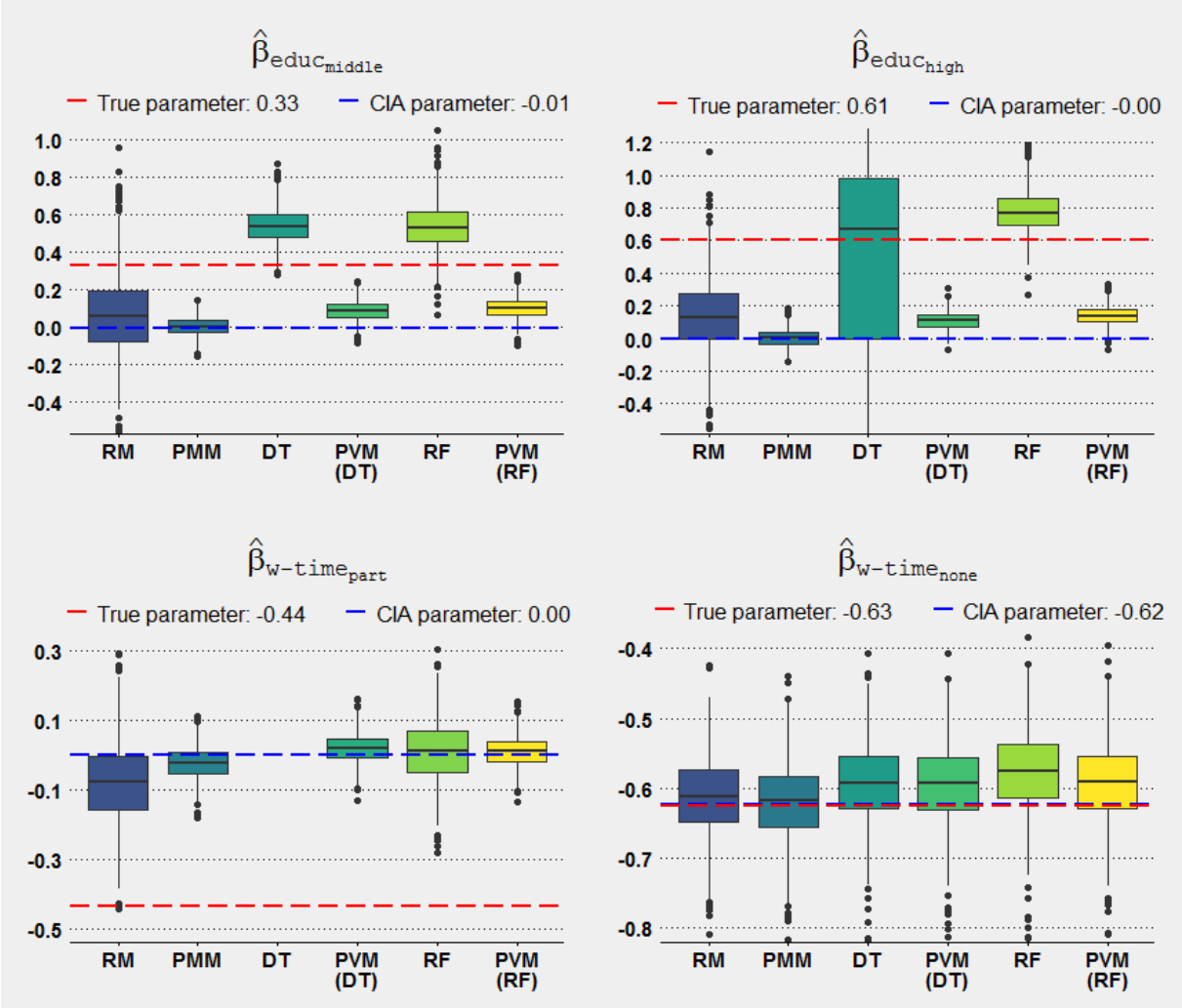
Table 6.7: Benchmark Parameters for β_{educ} and β_{w-time} under CIA Violation

	β_{educ_m}	β_{educ_h}	β_{w-time_p}	β_{w-time_n}
True parameter	0.3277	0.6059	−0.4357	−0.6255
CIA parameter	−0.0108	−0.0021	0.0009	−0.6227

Source: Microcensus (2014).

Here, we see that the CIA is indeed violated, since the CIA parameters are close to 0 for at least three of the four coefficients. In contrast, the CIA parameter for the non-working category

is quite similar to the true benchmark parameter. Since this is not the case for the part-time category, the CIA is still at least partly violated but also partly fulfilled. According to the associations between $\mathbf{X}$ and $\mathbf{Z}$ from Table 6.3, we observe rather high associations between the employment status (X_3) and working hours (Z_2) and at least moderate associations between sex (X_1) and working time (Z_2) as well as between family type (X_5) and working time (Z_2). From this, the partial violation and fulfilment of the CIA for working hours seems realistic. In addition, as has already become clear, the employment status (X_3) comprises a category 'other', which is highly dominated by non-working respondents. Thus, the category 'other' almost perfectly explains the non-working category of the working time (Z_2) characteristic, which is why the CIA fulfilment for the non-working time parameter is plausible.

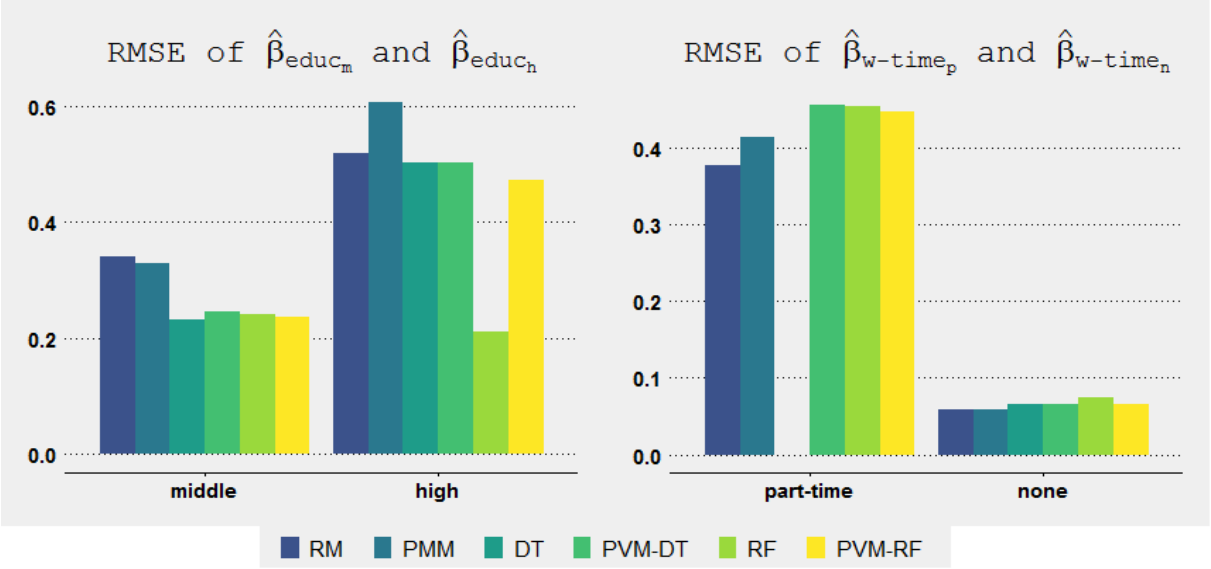


Source: Microcensus (2014).

Figure 6.6: MC distributions for $\widehat{\beta}_{educ}$ and $\widehat{\beta}_{w-time}$ with n_1 under CIA Violation

Figures 6.6 and 6.7 illustrate the results of all $k = 1,000$ Monte Carlo simulation runs under n_1, that is, under a high donor ratio scenario. Again, we observe the interesting phenomenon that

potential exaggeration effects may be partially beneficial for DT and RF in this case, at least with regard to the educational parameters. With respect to $\widehat{\beta}_{educ_m}$, DT and RF produce mean correlations of 0.54 (see Tab. A.11). Thus, DT and RF still induce a bias of $0.54 - 0.33 = 0.21$ with regard to reproduce the middle education parameter, while all other methods underestimate the true parameter and rather produce the CIA coefficient. However, the RMSE values for $\widehat{\beta}_{educ_m}$ illustrated in Figure 6.7 are quite similar for DT, PVM-DT, RF and PVM-RF. PVM-DT and PVM-RF have a slightly higher bias compared to DT and RF, but have a lower variance, resulting in relatively similar RMSEs. With regard to the high education parameter, DT and RF provide coefficients of 0.59 and 0.77 on average, respectively, while all other methods again predominantly yield parameters around the CIA coefficient. In terms of content, RM, PMM, PVM-DT and PVM-RF would thus imply that higher education has almost no or at most a minor influence on income, which is, though, unrealistic. RF provides the best overall performance with regard to the education parameters along the RMSE values.



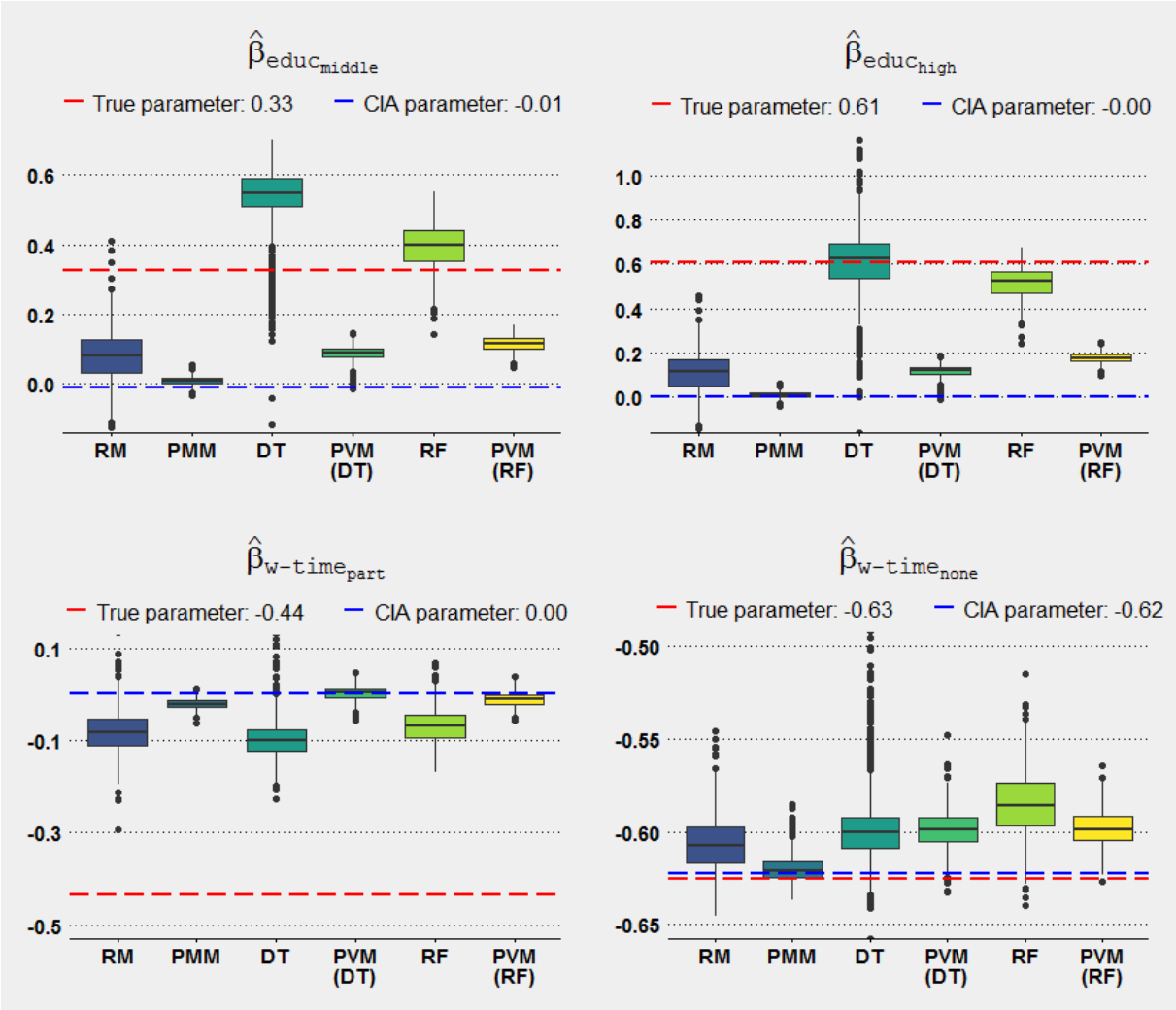
Source: Microcensus (2014).

Figure 6.7: RMSE of $\widehat{\beta}_{educ}$ and $\widehat{\beta}_{w-time}$ with n_1 under CIA Violation

What is striking here, however, is the high variance for DT with respect to $\widehat{\beta}_{educ_h}$. Consequently, the RMSE values for the high education coefficients resulting from DT are significantly higher than those of RF, as indicated in Figure 6.7 and Table A.12. This reflects the fundamental disadvantage of Decision Trees, which tend to have a high variance because different data can imply very different models. It is also noticeable that the fully parametric RM approach, in contrast to DT and RF, performs similarly poorly to the nearest neighbour methods PMM, PVM-DT and PVM-RF with regard to both education parameters. This could be due to an underlying

violation of distributional assumptions of RM, whereby in the case of logistic regression for categorical **Z** variables, at least linearity is assumed between the common **X** variables and the log odds of **Z**. At the very least, it seems obvious that a linear relationship between the $p = 6$ common **X** variables and **Z** is much less obvious than between income and education in the case of the CIA compliance scenario (where income additionally reflects a common variable).

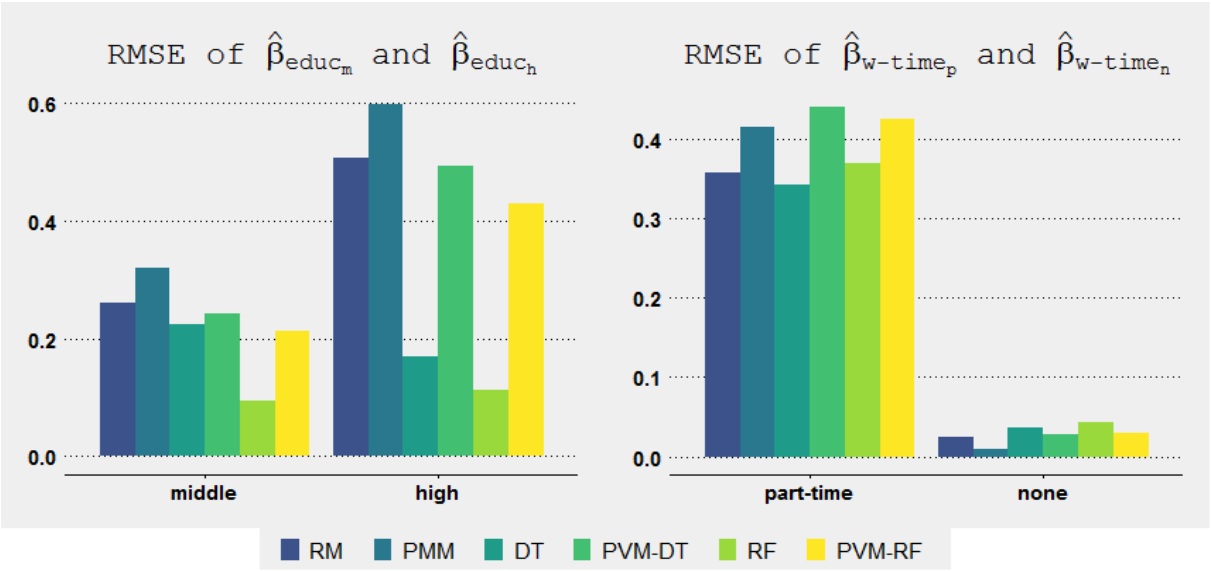
Concerning the working time parameters under n_1, we see that no method is able to reproduce the part-time coefficient. All potential data fusion approaches result in coefficients that are close to the CIA parameter. It is striking that the part-time category was not imputed by DT in any simulation run under n_1, which is why no part-time coefficients are observed for DT in this case. In contrast, concerning the non-working time parameter, all methods are predominantly able to approximately reproduce the benchmark coefficient of -0.63. However, the non-working time coefficient is subject to the favourable conditions of approximate CIA fulfilment.



Source: Microcensus (2014).

Figure 6.8: MC distributions for $\widehat{\beta}_{educ}$ and $\widehat{\beta}_{w-time}$ with n_2 under CIA Violation

The results under n_2 and thus under a low donor ratio are illustrated in Figures 6.8 and 6.9. With regard to the education parameters and the tree-based prediction methods DT and RF, both seem to benefit from the smaller donor pool in this case. RF now performs better for both education parameters. The RMSE values for RF under n_2 are significantly lower than under n_1 (see Fig. 6.9 and Tab. A.12). Similarly, the parameter values for education resulting from RF are now 0.39 and 0.52 on average, which are relatively close to the true coefficients of 0.33 and 0.61 (see Tab. A.11). Hence, the lower donor pool indicates mitigated exaggerations for RF and thus an improved performance compared to the high donor ratio. This mitigation of the respective exaggeration effects under n_2 was already observed in the CIA compliance scenario. DT, on the other hand, shows improvements for $\widehat{\beta}_{educ_h}$ under n_2. This becomes clear when comparing the respective RMSE values for the high education parameter resulting from DT under n_1 and n_2. The improvements for DT concerning the high education parameter are in particular due to the substantially lower variance compared to n_1. Nevertheless, RF performs better overall than DT in terms of reproducing the education parameters. In addition, along the RMSEs, RM also seems to benefit slightly from the lower donor pool in terms of the middle education parameter. This is again due to the lower variance, as the resulting middle education parameters for RM under n_2 are on average still quite similar to those under n_1, indicating a similar bias (see Tab. A.11). Concerning all other methods, no substantial change for the education parameters are observed under n_2 compared to n_1.



Source: Microcensus (2014).

Figure 6.9: RMSE of $\widehat{\beta}_{educ}$ and $\widehat{\beta}_{w-time}$ with n_2 under CIA Violation

When looking at the working time parameters under n_2, it is apparent for the part-time coef-

ficient that RF yields slightly better results than under n_1, but is still strongly biased. For DT, no comparison is possible with the performance for $\widehat{\beta}_{w\text{-}time_p}$ under n_2 with that under n_1. All methods again perform well in preserving the non-working time parameter, while the variances for all methods decrease compared to n_1. Consequently, all data fusion approaches yield even lower RMSE values under n_2 compared to n_1 with respect to the non-working time coefficient.[2]

Table 6.8: Relative Frequencies of $\widetilde{Z}_{educ}$ and $\widetilde{Z}_{w-time}$ under CIA Violation

		Education			**Working Time**		
		low	middle	high	none	part-time	full-time
$\mathbf{n_1}$	**RM**	0.02	0.78	0.21	0.13	0.02	0.85
	PMM	0.15	0.50	0.35	0.12	0.16	0.71
	DT	0.06	0.82	0.12	0.13	0	0.87
	PVM-DT	0.14	0.49	0.37	0.13	0.16	0.71
	RF	0.03	0.71	0.25	0.13	0.04	0.83
	PVM-RF	0.14	0.49	0.37	0.13	0.16	0.71
$\mathbf{n_2}$	**RM**	0.03	0.71	0.26	0.13	0.04	0.84
	PMM	0.14	0.49	0.36	0.12	0.17	0.71
	DT	0.04	0.71	0.25	0.13	0.03	0.84
	PVM-DT	0.14	0.49	0.36	0.13	0.16	0.71
	RF	0.05	0.65	0.30	0.13	0.05	0.82
	PVM-RF	0.12	0.52	0.36	0.13	0.15	0.72
	MC Database	0.14	0.49	0.37	0.13	0.16	0.71

Note that all values under n_1 and n_2 for each method reflect the Monte Carlo means of the relative frequencies from all $k = 1,000$ simulation runs. The distribution of the MC database serves as benchmark.

Source: Microcensus (2014).

The underlying results can again be further classified and discussed with a look at the marginal distribution of $\widetilde{Z}_{educ}$ and $\widetilde{Z}_{w-time}$. In this respect, Table 6.8 contains the relative frequencies of the education and working time categories averaged over all $k = 1,000$ simulation runs for each method under n_1 and n_2. The distribution of the MC database is analogous to Table 6.6 and again reflects the benchmark. In Table 6.8 we first see that PMM, PVM-DT and PVM-RF show almost no differences in the univariate distribution compared to the CIA compliance scenario. This illustrates how the nearest neighbour methods PMM, PVM-DT and PVM-RF

[2]Note that some boxes for $\widehat{\beta}_{w\text{-}time_n}$ in Figure 6.6 (bottom right) appear as narrow as in Figure 6.8 (bottom right), but the scales of the vertical axes reflecting the parameter values differ in the two plots.

seem to achieve acceptable results at first sight in the case of CIA violation according to the marginal distribution, but this is by no means true for the joint distribution between income and education and between income and working hours. Furthermore, it can be observed that for RM, DT and RF the CIA violation in this case seems to condition the predominant imputation of the mode category even more strongly than in the simulation under CIA fulfilment. The even more frequent imputation of the mode category of RM, DT and RF this time comes more at the expense of the second most frequent category for education (which is high education) and more at the expense of the part-time category for working time. Even though DT and RF reproduce the univariate distribution of $\mathbf{Z}$ unsatisfactorily, their performance with regard to the joint distribution between $\mathbf{Y}$ and $\mathbf{Z}$ is superior to that of the other methods. RM, on the other hand, produces a suboptimal result with regard to both the univariate distribution of $\mathbf{Z}$ and the joint distribution of $\mathbf{Y}$ and $\mathbf{Z}$, which could be due to a possible violation of distributional assumptions.

6.3.3 Discussion

The results presented in the previous two sections refer to the multivariate imputation solution for PVM-DT and PVM-RF. The results for univariate PVM under the CIA compliance and CIA violation scenarios, both with a high and a low donor ratio under n_1 and n_2, are included in Appendix B.2 and illustrated in Figures B.5 to B.8. Here, slight changes regarding the working time parameters are shown for PVM-DT and PVM-RF in the scenario of CIA compliance under n_1. The respective parameter estimates of PVM-DT and PVM-RF are somewhat more biased towards 0 for $\widehat{\beta}_{w\text{-}time_p}$ in the univariate case, while PVM-DT and PVM-RF show slight improvements with regard to the non-working time parameter compared to the multivariate PVM solution (see Fig. B.5). For n_2, an increased variance and a stronger bias for PVM-DT compared to multivariate PVM-DT with respect to $\widehat{\beta}_{w\text{-}time_p}$ under CIA compliance can be observed (see Fig. B.6). Marginal improvements for PVM-DT and PVM-RF for $\widehat{\beta}_{w\text{-}time_n}$ under n_1 and n_2 are observable in the univariate case when the CIA is violated (see Fig. B.7 and B.8). However, these changes are marginal at most. Overall, no substantial differences can be observed between the multivariate and the univariate PVM imputation. The possible preference for multivariate imputation in practice is therefore at least not countered by expected performance losses compared to univariate imputation.

Contrary to the results from Chapter 5, it is evident in the context of the underlying simulations that, in the case of CIA fulfilment, all prediction methods, both RM as a classical imputation

approach and the SL methods DT and RF, tend to overestimate the associations between $\mathbf{Y}$ and $\mathbf{Z}$. While in Chapter 5 there were sometimes very high original correlations between $\mathbf{Y}$ and $\mathbf{Z}$ (which provided less potential for overestimation), the benchmark correlations between $\mathbf{Y}$ and $\mathbf{Z}$ are lower in this case. Kendall's τ, which reflects a suitable measure for correlations between metric and ordered categorical characteristics, is 0.29 for the correlation between income and education and -0.24 for the correlation between income and working hours (Microcensus 2014). Thus, there are medium associations between $\mathbf{Y}$ and $\mathbf{Z}$ (note that $-1 \leq$ Kendall's $\tau \leq 1$). This is likely to increase the vulnerability of RM, DT and RF to overestimate the correlations in the CIA compliance scenario. The advantage of PVM in the case of CIA fulfilment is to mitigate such exaggeration effects.

However, if the CIA is violated and biased towards 0, the disadvantage of potential overestimation effects of DT and RF could turn into an advantage, which was at least evident with regard to the reproduction of the education parameters under CIA violation. With the fully parametric RM approach, on the other hand, the potential to benefit from possible overestimation effects in the case of CIA violation seems to depend on the validity of the distributional assumptions. The nearest neighbour methods PMM, PVM-DT and PVM-RF, on the other hand, again produce estimates biased towards the CIA parameters. However, especially in the CIA violation scenario, all methods had difficulties reproducing the part-time coefficient adequately. The corresponding distribution of the working time variable seems challenging as it is relatively biased in favour of the full-time category, while the 'none' category can be almost perfectly explained by the 'other' category of employment status (X_3). From the combination of both aspects, a particular difficulty may be derived for all methods to adequately impute the part-time category. In such a case, potential exaggeration effects of DT and RF, which could be useful in the case of the CIA violation, seem to largely evaporate.

With respect to the explicit scenarios of a high and low donor ratio, it should be noted that a low donor ratio is not necessarily a disadvantage. In contrast to the simulations in Chapter 5, some fusion methods seem to benefit from a low donor pool. In this case, however, the underlying donor sample is much larger than the donor study in the data fusion use case of EU-SILC and HBS. A larger donor pool as the basis for model calculation could lead to even greater exaggerations in the prediction methods, suggesting that using the smaller dataset as recipient data might even be beneficial when fusing quite large datasets.

With regard to the PMM approach, which was originally focused on metric variables, it must be stated that the fundamental reservations for PMM could not be dispelled in the case of un-

derlying ordered categorical **Z** variables. PMM was almost always shown to be inferior to the PVM approaches in the case of CIA fulfilment and to the SL methods DT and RF in the case of CIA violation. The crucial source of error is likely to be the erroneous pretence that education and working time are metric variables, which causes PMM's performance to suffer.

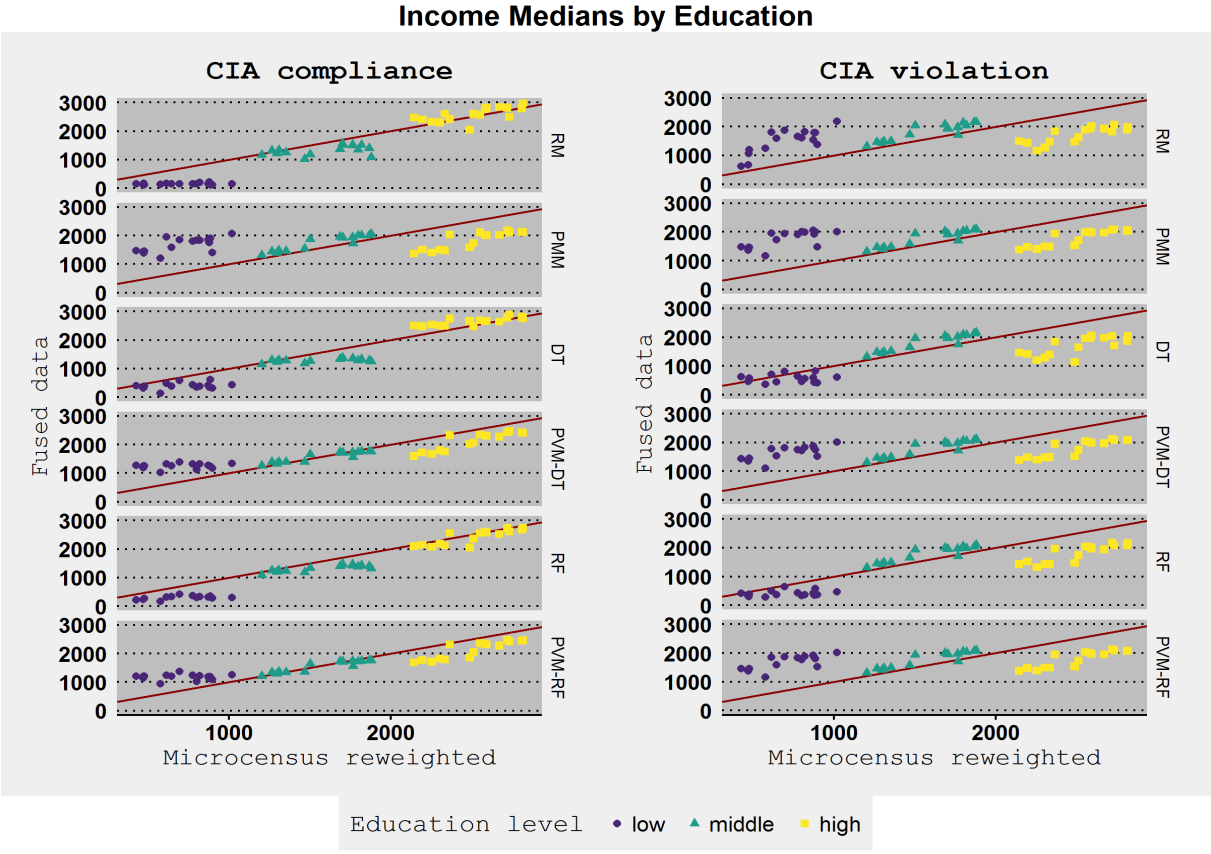
Overall, the simulation results show that the explicit scenarios of fulfilling or violating the CIA have a significantly stronger impact on the performance of different fusion methods than a high or low donor ratio, with the low donor ratio tending to be preferred in this context. Depending on the compliance or violation of the CIA, either PVM-RF or RF has proven to be a particularly promising data fusion method in the underlying simulations. The next section complements these simulation results with an empirical application.

6.4 Empirical Results

An empirical application and evaluation of the respective data fusion methods provides at least rough additional insights on the performance of the respective data fusion methods. This is especially due to the nature of the underlying data sources, an income register (TS) and a comprehensive survey with a large sample and the obligation to provide information (MC). Modifications in terms of interpolating and reweighting the income information within the MC ensures an acceptable degree of comparison between the Microcensus and the fused data of Tax Statistics and Microcensus. For the evaluations in this section, we initially conduct the real data fusion of Tax Statistics and Microcensus, that is, we impute the missing education and working time variables within the Tax Statistics by means of the respective data fusion approaches. With regard to PVM-DT and PVM-RF, we apply the multivariate imputation solution. The evaluations in this section build on Emmenegger et al. (2023) and are extended to both PVM methods and to diagnostics concerning the marginal distribution.

First, we evaluate the conditional income medians at the regional level of the federal states, since one advantage of the Tax Statistics as a register-based dataset also lies in the possibility of high-quality regional analyses. For this, we compare income medians conditioned on the education and working time levels between the fused data and those observed in the reweighted Microcensus. In this respect, Figure 6.10 illustrates the income medians for each federal state conditioned on educational level, which is indicated by the three different colours. Here, the y-axis reflect the conditional income medians obtained from the fused data, that is, the Tax Statistics enhanced by the socio-demographic variables, while the x-axis represent the respec-

tive income medians from the reweighted Microcensus. This comparison is illustrated for each of the respective data fusion procedures to match TS and MC, represented by the six panels. The red line marks the bisector. The left-hand side reflects the results when income is included in the data fusion process, thus fulfilling the CIA, while the right-hand side refers to the results when income is excluded as a common variable, thus violating the CIA.

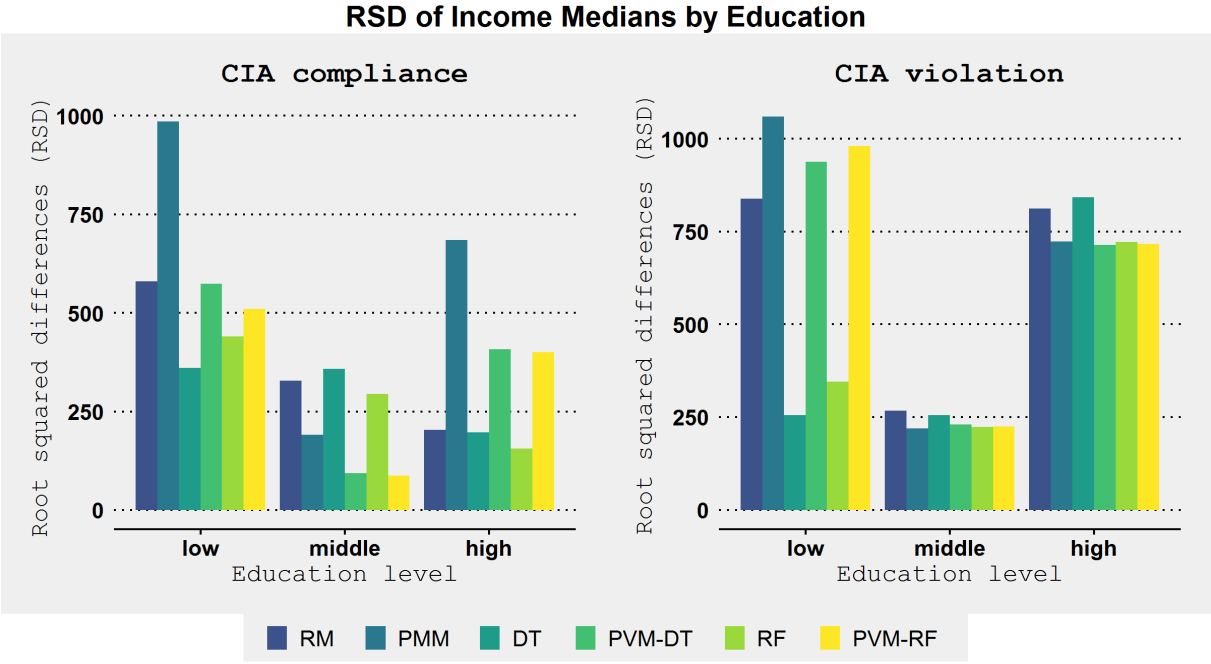


Source: Microcensus (2014); Tax Statistics (2014).

Figure 6.10: Income Medians of Federal States by Education from MC and Fused Data

With regard to the income medians conditioned on the education levels resulting from each data fusion method, it is apparent in Figure 6.10 for the CIA compliance scenario that the education variable imputed by RM seems to underestimate the income medians for the low educated individuals within all federal states. This can also be observed for DT and RF, but to a lower extent, since the income medians conditioned on low education are closer to the red line. PVM-DT and PVM-RF, in contrast, show opposite effects and seem to overestimate the income medians for low educated individuals. The majority of the regional income medians conditioned on middle and high education seem to be well represented by RM, PVM-DT, RF and PVM-RF within the fused data file. In line with the simulation results, we observe problems for PMM to adequately reproduce the joint distribution between income and education. Concerning the CIA

violation scenario on the right-hand side of Figure 6.10, we see that RM, PMM, PVM-DT and PVM-RF struggle to reproduce the association between income and educational attainment, since along the y-axis all points are in the same range regardless of the different education levels. DT and RF are better able to map the correlation between income and education under CIA violation, but are prone to underestimate the median incomes of the highly educated. In addition, it is noticeable that DT and RF can reproduce the differences in median incomes between groups with low education levels on the one hand and middle and high education levels on the other. In contrast, income differences between groups with medium and high education levels are not well represented by DT and RF within the fused dataset under CIA violation.



Source: Microcensus (2014); Tax Statistics (2014).

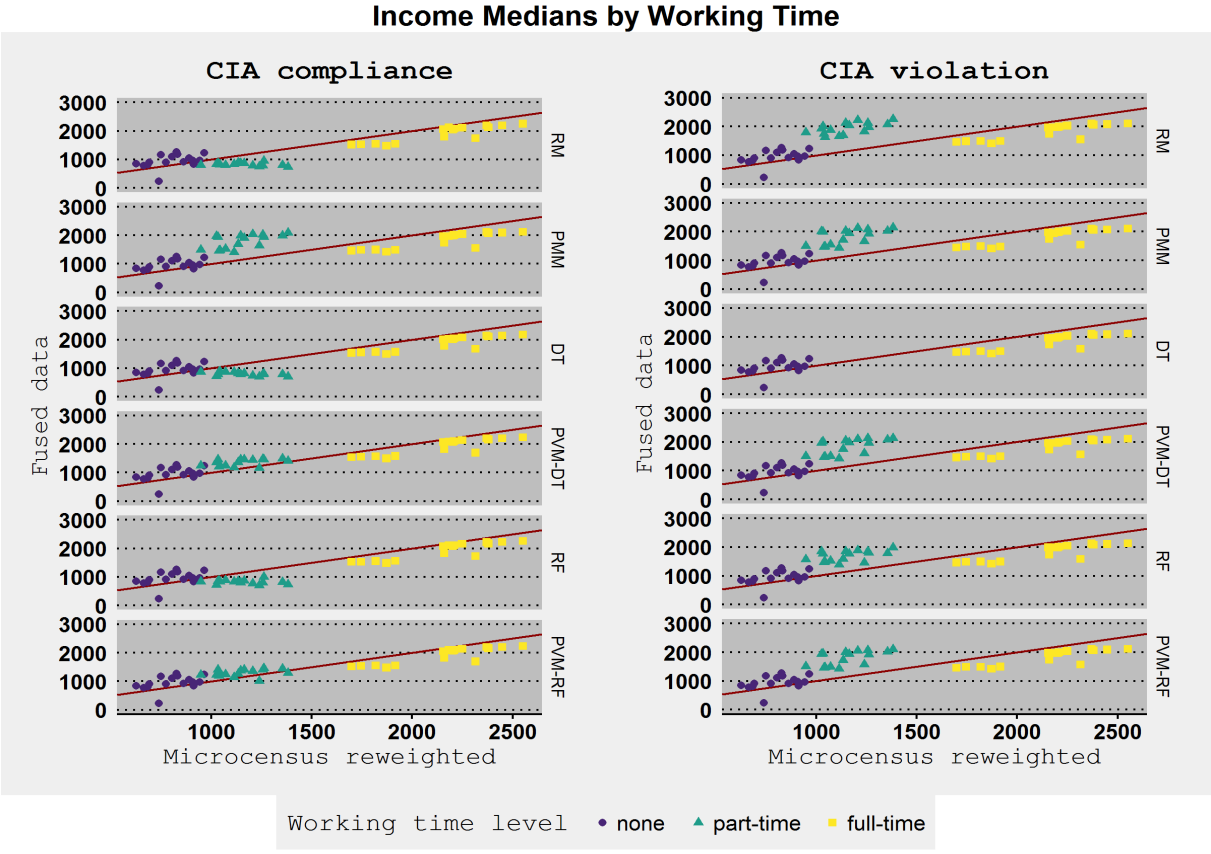
Figure 6.11: RSD of Federal States' Income Medians by Education

The discussed results are summed up in Figure 6.11 by illustrating the Root Squared Differences (RSD) between the conditional income medians from the Microcensus and the enhanced Tax Statistics averaged over all 16 federal states analogously to Emmenegger et al. (2023):

$$\text{RSD} = \frac{1}{16}\sum_{i=1}^{16}\sqrt{(m_{TS}(R_i\,|\,Z_r) - m_{MC}(R_i\,|\,Z_r))^2}, \tag{6.4}$$

with $m(R_i)$ representing the income median from the ith federal state conditional on Z_r, which is in this case educational attainment. In addition, m_{MC} reflects the medians from the reweighted Microcensus, while m_{TS} corresponds to the fused data file, that is, the enhanced Tax Statistics (Emmenegger et al. 2023). The results in Figure 6.11 indicate for the CIA compliance scenario

that DT and RF are on average able to appropriately represent the income medians of low and high educated units over all 16 regions, while PVM-DT and PVM-RF are superior to RF and DT in terms of the middle educated. RM shows low RSD values for income medians of the highly educated, but rather higher RSD values for low and middle educated observation units. PMM, on the other hand, yields high RSD values for the groups with low and high education levels. With regard to the CIA violation scenario, it is apparent that DT and RF indicate the best overall performance, since both approaches result in substantially lower RSD values for the low educated compared to all other methods, while the RSD values for middle and high educated are similar for all data fusion approaches. In this respect, the empirical results support the findings from the simulation study that DT and RF are better able to cope with the challenging scenario of CIA violation. All exact values of Figure 6.11 are shown in Table A.13.



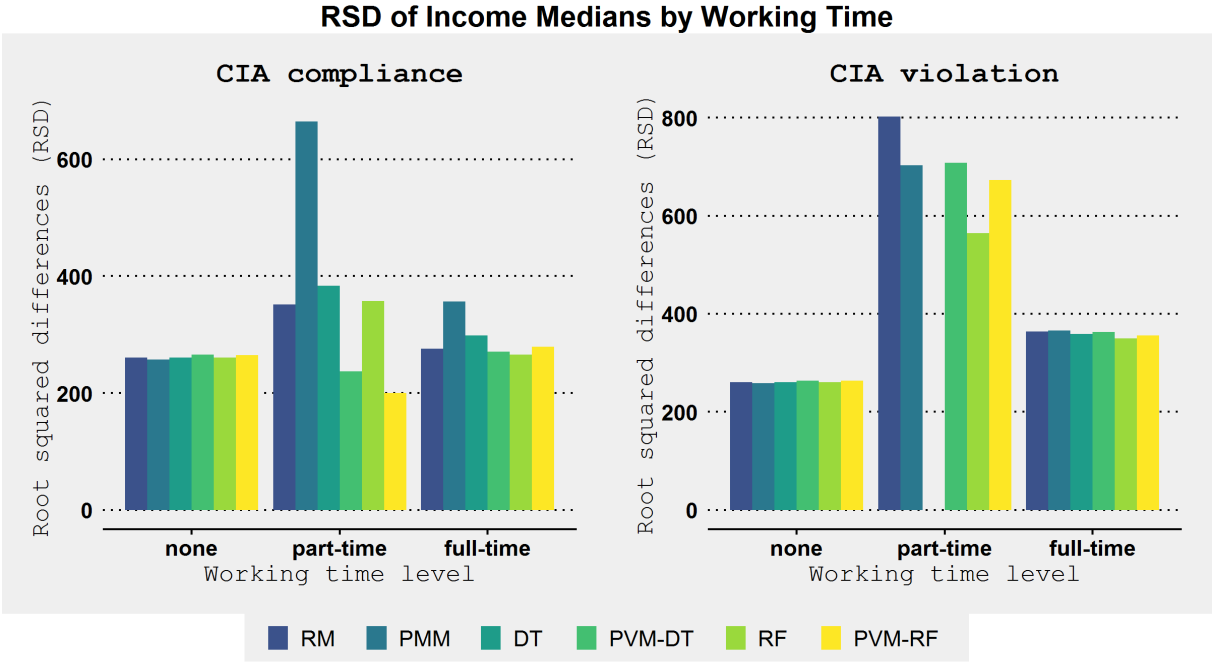
Source: Microcensus (2014); Tax Statistics (2014).

Figure 6.12: Income Medians of Federal States by Working Time from MC and Fused Data

The empirical results of income medians conditional on working time are illustrated in Figure 6.12. Here, we see for the CIA compliance scenario that all methods perform quite similar with regard to the regional income medians of non-working units and full-time workers. RM, DT and RF seem to underestimate the income medians of part-time workers, while PMM seems to

overestimate them. PVM-RF and PVM-DT, on the other hand, indicate superior results over all other methods in terms of part-time workers. For the CIA violation scenario, again all methods predominantly produce similar results for the income medians of non-working individuals and full-time workers. Additionally, all methods seem to be prone to overestimate the income median of part-time workers over all 16 regions. DT, on the other hand, never imputed the part-time category, which corresponds to the general disadvantage of single Decision Trees to predominantly impute the mode category at the expense of other categories.

These results for the specific working time variable to be fused are illustrated compactly across all 16 federal states by means of the RSD values in Figure 6.13. The exact RSD values are again listed in Table A.13. The RSD values highlight that all methods mostly yield similar results for the income medians conditioned on the non-working and the full-time category both under CIA compliance and under CIA violation. PMM yields by far the highest RSD for the group of part-time workers in the CIA compliance scenario, while PVM-DT and PVM-RF yield the lowest RSD values under the fulfilment of the CIA. RF in turn yields slightly lower RSD values for the income medians of part-time workers in the scenario of CIA violation.



Source: Microcensus (2014); Tax Statistics (2014).

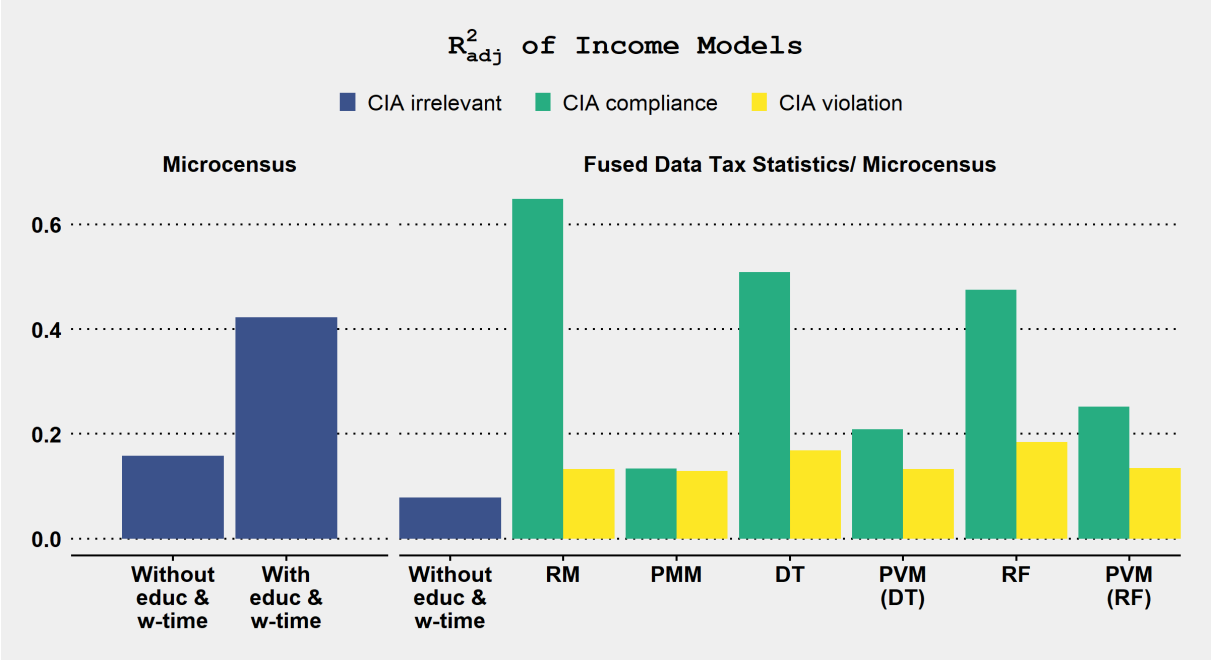
Figure 6.13: RSD of Federal States' Income Medians by Working Time

Apart from the rough evaluation of deviations between income medians conditional on education and working time observed from the Microcensus and the enhanced Tax Statistics, it could further be challenged whether the matched education and working time variables yield sub-

stantial improvements for potential income models. To evaluate possible model improvements, we consider the adjusted R^2 resulting from the exemplary income models specified in Section 6.2.3 as most appropriate measure in this context. Recall that the exemplary income models include the variables sex, age, number of kids, family type and employment status, while for employment status only dummies on self-employed and civil servants are considered. Figure 6.14 shows the adjusted R^2 obtained from the respective income models. The two bars on the left illustrate the adjusted R^2 resulting from the exemplary income models using the Microcensus, in one case excluding the socio-demographic variables education and working hours (first bar) and in the other case including these characteristics in the model (second bar). All other bars show the adjusted R^2 obtained from income models using the enhanced Tax Statistics. The third blue bar in the plot reflects the adjusted R^2 from an exemplary income model based on the Tax Statistics without including the two imputed socio-demographic variables education and working time. The green and yellow bars show the adjusted R^2 of the income models when including the imputed socio-demographic variables. These variables were imputed within the Tax Statistics by the underlying data fusion methods considered. Again, the explicit CIA-related scenarios are represented within the plot. The green bars refer to the adjusted R^2 values that result when income has been included in the data fusion process (thus fulfilling the CIA), while the income models behind the yellow bars include the imputed variables when income has been excluded as a common variable (thus violating the CIA). The exact values are shown in Table A.14.

With regard to the exemplary income models obtained from the Microcensus only, Figure 6.14 and Table A.14 indicate an adjusted R^2 of about 0.16 when excluding education and working hours from the models, and an adjusted R^2 of about 0.42 when including the socio-demographic variables. Hence, the adjusted R^2 when including education and working hours is almost three times higher than the adjusted R^2 of the restricted model. This highlights the importance of including information on the social disaggregation in income modelling. The adjusted R^2 of an income model from the restricted Tax Statistics without information on education and working time amounts to 0.08. All data fusion approaches contribute to improve the explanatory power of the income models within the Tax Statistics when including the imputed education and working time variables into the model, but to different extents. In the scenario of CIA compliance, RM, DT and RF seem to produce quite high and possibly exaggerated model improvements, given that the models within the Microcensus yield approximately a three times higher adjusted R^2 when education and working time are included. The adjusted R^2 of RF under CIA compliance (0.48), however, is about six times higher than those of the restricted model obtained from

TS (0.08). In this respect, the adjusted R^2 resulting from PVM-DT (0.21) and PVM-RF (0.25) are lower than those of the unrestricted MC model (0.42), but seem to be more realistic, as the adjusted R^2 from the restricted TS model (0.08) is already lower than that from the restricted MC model (0.16). On the other hand, if the CIA is violated, all methods seem to provide insufficient model improvements, while DT and RF at least yield an adjusted R^2 of 0.17 and 0.18, respectively. The model improvements of PMM, on the other hand, appear to be insufficient for both the CIA compliance and the CIA violation scenario.



Source: Microcensus (2014); Tax Statistics (2014).

Figure 6.14: Adjusted R^2 of Income Models with MC and Fused Data

In addition and analogously to the simulation study, a brief look at the univariate distribution of education and working time within the fused dataset is useful here. For this, Table 6.9 provides the respective relative frequencies of the fused data under CIA compliance and CIA violation in comparison to the distribution of $\mathbf{Z}$ within the Microcensus. The results underline the tendency of RM, DT and RF to predominantly impute the mode category, with this effect being even more striking when the CIA is violated (which is in line with the simulation results). PMM, PVM-DT and PVM-RF produce marginal distributions that are more close to those in the Microcensus. In contrast to the simulation study, all methods seem to underestimate the non-working category equally, while the category of the full-time employed is overestimated by all methods, albeit to different extents. Once again, it becomes clear that good or bad performance with respect to the univariate distribution does not provide any indication of performance with respect to joint distributions. For example, although DT and RF produce biased results in preserving

the univariate distribution of **Z** in the CIA violation scenario, their performance with regard to preserve joint distributions appear to be superior to those of the other methods under CIA violation.

Table 6.9: Relative Frequencies of $\widetilde{Z}_{educ}$ and $\widetilde{Z}_{w-time}$ of MC and Fused Data

		Education			**Working Time**		
		low	middle	high	none	part-time	full-time
Fused Data: CIA Compliance	**RM**	0.0284	0.5214	0.4502	0.0556	0.0705	0.8739
	PMM	0.1041	0.4798	0.4161	0.0542	0.1834	0.7624
	DT	0.0161	0.5296	0.4544	0.0556	0.0476	0.8967
	PVM-DT	0.0976	0.4489	0.4536	0.0547	0.1505	0.7948
	RF	0.0197	0.4612	0.5192	0.0556	0.0981	0.8463
	PVM-RF	0.0911	0.4244	0.4845	0.0552	0.1619	0.7829
Fused Data: CIA Violation	**RM**	0.0003	0.7237	0.2759	0.0556	0.0251	0.9193
	PMM	0.1192	0.4838	0.3970	0.0543	0.1803	0.7654
	DT	0.0279	0.8151	0.1570	0.0556	0	0.9444
	PVM-DT	0.1142	0.4860	0.3998	0.0548	0.1773	0.7678
	RF	0.0132	0.6322	0.3546	0.0556	0.0668	0.8776
	PVM-RF	0.1122	0.4874	0.4004	0.0551	0.1819	0.7630
Microcensus		0.1394	0.4949	0.3657	0.1255	0.1641	0.7104

Source: Microcensus (2014); Tax Statistics (2014).

It should be noted that, in contrast to the simulations, it was only possible to rudimentarily estimate the extent to which the possible joint distribution between income and the socio-demographic variables education and working hours can be represented in the fused data file. However, due to the quality of both datasets and the strategies of interpolation and reweighting, a database could be created that offers an acceptable degree of comparability. The empirical results largely support the simulation results. However, DT and RF show a similarly good performance as PVM-DT and PVM-RF with regard to the median income conditional on education in the case of CIA fulfilment. However, this is contradicted by the findings on the goodness of fit using the adjusted R^2. These could not invalidate the reservations derived from exaggeration effects for all three prediction methods RM, DT and RF and possibly produced too high model improvements under CIA fulfilment.

6.5 Concluding Remarks

The intended evaluation of various data fusion approaches has now been complemented by an underlying use case of data fusion that is fundamentally different from the data fusion scenario of EU-SILC and HBS discussed in the previous chapter. While EU-SILC and HBS represent conventional samples and the larger sample is used as donor study, TS and MC reflect extremely large data sources, which excludes DHD from the analyses. In addition, the smaller dataset, MC, reflects the donor file and categorical **Z** variables are to be imputed in the recipient dataset instead of metric variables. The motivation of the data fusion of TS and MC is to ensure high-quality income analyses, which is not possible if the respective datasets are used separately. Since a rough empirical application seemed promising in this case, an empirical evaluation was carried out in addition to the simulations.

Bringing together the results from the simulations and the empirical evaluation, it can be stated that PVM-RF represents the most promising approach under the explicit scenario of CIA compliance. Within the simulations, PVM-RF was able to reproduce the education and working time coefficients to an acceptable extent, apart from slight performance disadvantages under a high donor ratio for the part-time coefficient. The empirical results indicate tendencies that PVM-DT and PVM-RF overestimate the income medians of low educated individuals while underestimating those of the highly educated. If the CIA is violated, however, all methods except of DT and RF yield poor results. Under a low donor ratio, RF was able to largely reproduce the associations between income and the socio-demographic variables education and working time under CIA violation according to the simulations and the empirical results. This can also be partially observed for DT, whereby a disadvantage of DT is the overestimation of the mode category, which in extreme cases leads to the fact that not all categories are imputed (as was the case with the part-time category of working time). In the case that a specific variable to be matched has high relative frequencies for the mode category, while one of the other categories can be well explained by one of the common variables (as was the case for working time), all methods showed partial problems in imputing a category that has low relative frequencies and cannot be almost perfectly explained by another common variable.

Furthermore, the results indicate that a low donor ratio can even be advantageous if the underlying donor dataset implies a sufficiently large donor pool due to its sample size. This is an important finding, given that the concrete data fusion scenario of TS and MC implies a low donor ratio. In this respect, the results indicate that an adequate data fusion between both datasets

is at least not countered by the explicit scenario of a low donor ratio. Of particular relevance, on the other hand, is the handling of the CIA. The results of the simulations and the empirical application show that the strategy already applied in Emmenegger et al. (2023) of adequately including the rough income information from MC in the data fusion process is promising, especially for PVM. In this case, however, caution is advised when using SL prediction methods. Yet, also the part-time parameter could be imputed to an acceptable extent under CIA compliance. Therefore, with regard to the concrete data fusion use case of matching TS and MC, the results suggest to include the income information in the fusion process to circumvent the CIA and to perform the data fusion by means of PVM-RF in order to adequately represent the joint distribution of income and the socio-demographic characteristics education and working hours in the fused data file. This is possible with the underlying donor-recipient ratio that reflects a low amount of donor observations compared to the number of recipient units. Hence, improved income models based on an income tax register can be implemented with the additional consideration of characteristics on the social disaggregation. Thus, the strength of both datasets, TS and MC, are exploited more effectively.

In line with Section 4.4, it should be noted that PVM typically implies zero distances between the predictions of the recipient and donor data files for categorical $\mathbf{Z}$ variables to be imputed. Thus, the imputation process is subject to a stochastic component. This basic idea also underlies the programme implementations of DT and RF in the `R` package `mice` (van Buuren 2022). Therefore, the application of `mice` and the method arguments 'cart' and 'rf' should yield relatively similar results, at least in the univariate case and with categorical $\mathbf{Z}$ variables to be imputed. However, `mice` only includes a univariate imputation solution (van Buuren 2022). The more general PVM approach, on the other hand, provides both a univariate and a multivariate imputation solution. Without the PVM approach, there would be no adequate multivariate imputation solution in the concrete use case of data fusion of TS and MC. Since the analyses subsequent to the data fusion relate in particular to regression models on income, the multivariate imputation solution would be preferable here in practice. This ensures that the cell combinations of education and working time remain consistent by identifying an overall donor across both $\mathbf{Z}$ variables.

In addition, `mice` also comprises an imputation model for RM via the method argument 'polyreg' (see van Buuren 2022), which is based on a data augmentation method according to White et al. (2010) in order to mitigate potential performance problems of fully parametric approaches. This mice-polyreg implementation was used by Emmenegger et al. (2023) for their analyses, and they referred to this approach as 'multinom'. The results, which are restricted to a low donor ratio

scenario, indicate a good performance for mice-polyreg in the case of CIA compliance and a poor performance in the CIA violation scenario (Emmenegger et al. 2023). Hence, if the CIA is violated, neither the original RM approach, nor mice-polyreg seems beneficial in this case. However, if the CIA holds, then mice-polyreg is able to mitigate potential exaggerating effects. In addition, it is of course also possible to implement PVM based on logistic regressions in the case of categorical $\mathbf{Z}$ variables, which is expected to yield similar mitigation effects.

Furthermore, it is to be noted that the results presented with regard to PMM, DT and RF are in the low donor ratio scenario largely similar to those in Emmenegger et al. (2023), but differ to some extent. This is particularly because slight modifications have been made here with respect to the concrete common $\mathbf{X}$ variables and the independent variables of the exemplary income models in order to avoid *uncongeniality* issues (Meng 1994; Xie and Meng 2017) and to calculate the CIA parameters consistently.

Simulation studies do not provide general validity, and it has to be presumed that the results are transferable to comparable data situations. In this respect, it is encouraging that the empirical results, which are based on rudimentary but as comparable as possible benchmarks, largely support the simulation results. The underlying simulations and evaluations thus completed the intended investigations in this work. The next chapter summarises the results and provides implications for the respective data constellations tailored to data fusion problems in official statistics.

Chapter 7

Conclusion

The investigations in this thesis were motivated by the steadily growing importance of data fusions in official statistics in order to exploit the manifold data treasures of official data. At the same time, confidentiality aspects and methodological considerations of relieving the response burden lead to official statistics data sources that are mostly devoted to a particular objective. Legal restrictions and compliance with necessary data protection are essential for trust in official statistics, which is why a direct linkage of the data sources via unique identifiers is typically not possible. In addition, when dealing with survey samples, a horizontal overlap of observation units is extremely unlikely. While data fusions are thus an important topic in official statistics, the literature lacks a comprehensive evaluation of a concrete plethora of different data fusion methods, encompassing both classical imputation approaches and statistical learning procedures. Furthermore, SL approaches require a general extension that mitigates potential exaggeration effects of SL procedures, provides a multivariate imputation solution, and imputes real values instead of artificial values in the case of metric $\mathbf{Z}$ variables.

The underlying thesis served the purpose of addressing these research gaps. For this, we first introduced and defined different types of scenarios that represent concrete data and imputation situations tailored to official statistics. We classified them into explicit, implicit and imputation scenarios and selected two concrete scenarios for each of these scenario types, which claim to cover a broad range of possible data fusion constellations in official statistics. Subsequently, the concrete plethora of potential data fusion methods were specified by three classical imputation approaches, DHD, RM and PMM, and two prominent and widespread SL procedures, DT and RF. With PVM, a new method for the toolbox of data fusion and missing data in general was introduced. Since PVM serves as a general approach that can be based on various prediction methods, we chose to investigate PVM based on single Decision Trees (PVM-DT) and Ran-

dom Forests (PVM-RF). While all data fusion methods considered were directly classified with regard to the implicit and imputation scenarios, comprehensive evaluations were required to assess their performance with respect to the explicit scenarios.

Two current but fundamentally different data fusion use cases in official statistics served as basis for the simulations and evaluations. One was the data fusion of EU-SILC and HBS, representing a classical data fusion scenario with conventional sample sizes and a high donor-recipient ratio, while metric $\mathbf{Z}$ variables are to be imputed. The other was the data fusion of TS and MC, which is based on rather large data sources and a low donor-recipient ratio, with ordered categorical $\mathbf{Z}$ variables to be fused. In particular, we incorporated into the simulation designs the explicit scenarios of compliance or violation of the CIA and of high or low donor-recipient ratios. Concerning the evaluations, special attention was paid to the joint distribution of $\mathbf{Y}$ and $\mathbf{Z}$ by looking at the correlations and regression parameters, respectively, which corresponds to the third validity level of data fusion according to Rässler (2002: 29-32). Furthermore, additional evaluation criteria were taken into account that appeared to be meaningful and useful depending on the specific data fusion use case. Therefore, in the context of matching EU-SILC and HBS, the reproduction of the joint distribution of the variables $\mathbf{X}$ and $\mathbf{Z}$ already present in the donor data file was also considered, which is a kind of minimum requirement for a data fusion. In the context of the data fusion of TS and MC, on the other hand, a look at the marginal distributions of the categorical $\mathbf{Z}$ variables and an empirical evaluation were particularly useful. In contrast to the data fusion of EU-SILC and HBS, TS is a full survey and MC is a large sample where comparability seems realistic when MC incomes are adjusted to incomes obtained from TS using reweighting methods. Therefore, an empirical evaluation seemed promising for the data fusion of TS and MC.

7.1 Implications for Official Statistics

Several implications for official statistics can now be derived from the results of the simulations and evaluations. In this respect, Table 7.1 complements Tables 3.1 and 4.2 with the explicit scenarios and gives an overview of which data fusion method should be considered under which data and imputation constellation. The first important implication is that DHD, which reflects a traditional and popular data fusion approach that appears to be the default method for practical applications, should not be considered as a suited data fusion method, irrespective of the underlying explicit scenario. Even the minimum requirement of reproducing the correlations

already observed in the donor data file could not be guaranteed by DHD. One restriction of this implication is that DHD could only be considered in the simulations concerning the data fusion of EU-SILC and HBS. However, an additional disadvantage of DHD arises in terms of computing capacity, as common programme implementations, such as the `R` package `StatMatch` (D'Orazio 2022), are not able to process such large datasets with conventional computing capacities. The implicit scenario of large sample sizes thus excludes the use of DHD due to the nature of the data in terms of the data fusion of TS and MC.

Table 7.1: Evaluating Classical versus Statistical Learning Approaches

			DHD	RM	PMM	DT	PVM (DT)	RF	PVM (RF)
Explicit Scenarios	**CIA**	(Appr.) fulfilled	⛔	⚠	✅	⚠	⚠	⚠	✅
		Violated	⛔	⚠	⛔	⚠	⛔	⚠	⛔
	Donor-recipient ratio	High	⛔	✅	✅	✅	⚠	✅	✅
		Low	⛔	✅	✅	✅	⚠	✅	✅
Implicit Scenarios	**Summed sample size**	$\leq 170,000$	✅	✅	✅	✅	✅	✅	✅
		$> 170,000$	⛔	✅	✅	✅	✅	✅	✅
	Scale level of Z	Metric	✅	✅	✅	✅	✅	✅	✅
		Categorical	✅	✅	⚠	✅	✅	✅	✅
Imputation Scenarios	**Imputation solution**	Univariate	⛔	✅	✅	✅	✅	✅	✅
		Multivariate	✅	⛔	✅	⛔	✅	⛔	✅
	Imputed metric values	Observed values	✅	⛔	✅	⛔	✅	⛔	✅
		Artificial values	⛔	✅	⛔	✅	⛔	✅	⛔

Interpretation: ✅ Method should be considered;
⚠ Method should be considered with caution;
⛔ Method should not be considered.

With regard to the PMM approach, another implication is that the reservations against PMM regarding the imputation of ordered categorical **Z** variables could not be eliminated. While we did not examine unordered categorical **Z** variables to be fused, it is reasonable to assume that if there are already caveats for ordered categorical variables, performance is unlikely to be

better for unordered categorical **Z** characteristics. Yet, in some cases PMM produced acceptable results in terms of ordered categorical **Z** variables. However, these results cannot convincingly dispel the reservations, which is why PMM should be considered with caution when imputing categorical variables, as pronounced by the exclamation mark in Table 7.1. Hence, the implicit scenario of a categorical **Z** variable indirectly restrict the use of PMM by the nature of the data, but does not invalidate PMM as a viable data fusion approach in general.

Further implications can be derived from the explicit scenarios, which we defined as scenarios that are expected to have a direct impact on the performance of data fusion procedures. Throughout the simulations, only PMM and PVM-RF proved useful if the CIA is at least approximately fulfilled. For PMM, however, this implication is restricted to metric **Z** variables. An important conclusion for future data fusions is therefore that PMM in the metric case and PVM-RF for categorical or metric **Z** variables should be preferred over all other approaches investigated in this work if there is evidence that the CIA at least approximately holds. Caution is advised when applying RM, DT, PVM-DT and RF under CIA compliance. In our evaluations, RM, DT and RF provided acceptable results concerning the data fusion of EU-SILC and HBS, but revealed overestimation effects in the matching case of TS and MC, where moderate true associations between **Y** and **Z** were present. Hence, RM, DT and RF are prone to overestimate correlations between different variables. PVM-DT, in turn, appeared to be prone to slightly underestimate the associations between the variables of interest in both considered data fusion use cases under the scenario of CIA compliance. Therefore, caution is also advised for PVM-DT if the CIA approximately holds, listed by a respective exclamation mark in Table 7.1.

In case of CIA violation, however, none of the nearest neighbour approaches, not even PMM and PVM, should be considered as potential data fusion techniques for the underlying data fusion problem. With regard to the prediction methods RM, DT and RF, their disadvantage under CIA compliance becomes advantageous under CIA violation due to potential exaggerations. The results from this thesis indicate that at least the SL prediction methods DT and RF are able to reproduce the joint associations between the specific variables **Y** and **Z** to some extent under CIA violation for moderate true associations of 0.29 and -0.24 (as was the case for TS and MC), although they still induce slight overestimates. However, too high underestimations in case of high (0.87 and 0.85) and medium original correlations (0.44 and 0.48) remained within the data fusion use case of EU-SILC and HBS under CIA violation. Potential exaggeration effects of RM are expected to depend on whether the distributional assumptions of linearity applies. Therefore, when the CIA is violated, all three prediction methods, the classical RM approach and the SL methods DT and RF, seem to be more promising compared to the nearest neighbour

approaches, but they still may fail to fully compensate underestimations of the correlations. For this reason, caution is also advised for RM, DT and RF under CIA violation. However, if the CIA is violated, then RF proved to be the most promising approach according to the results of this thesis. In general, it should be noted that the underlying results with respect to the CIA violation scenario are restricted to a violation towards 0, that is, the CIA correlation is lower than the original correlation. This is the most common problem regarding a violation of the CIA, but it is still theoretically possible that the CIA correlation is higher than the true correlation.

Regarding the scenarios of high and low donor-recipient ratios, it should first be noted that these have less impact on the performance of the data fusion procedures than the CIA-related scenarios. DHD in particular proved vulnerable to a low donor ratio, with DHD also failing to produce convincing results under a high donor ratio. Furthermore, a high donor ratio tends to yield stronger exaggerations for the prediction methods RM, DT and RF. This implication may be useful for future data fusions if there is evidence that the CIA strongly biases the correlations towards zero. Here, stronger exaggeration effects of RM, DT and RF could then be targeted by a large donor pool. In this case, the larger dataset should then serve as the donor data file. With regard to the nearest neighbour approaches PMM and PVM, a low donor ratio is not necessarily a disadvantage. In the simulations regarding the data fusion of TS and MC, PMM, PVM-DT and PVM-RF benefited from the low donor ratio. A possible explanation could be that DT and RF produced lower overestimates and thus better results under a low donor ratio, which in turn should provide a more favourable basis for the distance calculation of PVM-DT and PVM-RF. However, the overall deviations between the high and low donor ratio scenarios were moderate within each CIA-related scenario across all simulations. Since DHD failed in both donor-recipient scenarios and PVM-DT had partial performance problems across all evaluations, it is advisable not to consider DHD in either a high or low donor ratio scenario and to be cautious when using PVM-DT. All other methods, PMM, DT, RF and PVM-RF, are generally able to provide acceptable results for both high and low donor ratio scenarios. However, the extent to which a satisfactory result can be achieved is essentially determined by the CIA-related scenarios.

It also became clear throughout Chapter 6 that the evaluation of marginal distributions of the $\mathbf{Z}$ variables to be fused is insufficient for data fusion purposes and gives at best a rough indication on whether the joint distribution of $\mathbf{Y}$ and $\mathbf{Z}$ is adequately reproduced. This is also an important additional implication, given that some data fusion evaluations, especially in official statistics, seem to be based primarily on marginal distributions (Webber and Tonkin 2013; Serafino and Tonkin 2017).

In summary, PVM-RF proved to be the most promising and flexible data fusion approach across all evaluations conducted throughout this thesis, if there is evidence that the CIA at least approximately holds. This can also be stated for PMM, but with the restriction to metric **Z** variables. The additional advantage of PVM-RF as promising data fusion approach is its flexibility with respect to the implicit scenarios. This is due to the fact that PVM-RF is neither dependent on the underlying scale level of **Z** nor on the size of the datasets. Moreover, PVM-RF meets the desirable imputation scenarios of imputing observed and thus plausible metric values and of providing a multivariate imputation solution, irrespective of the scale level of the underlying variables to be imputed. Consequently, Table 7.1 shows that PVM-RF can be considered as a suited data fusion approach for most of the selected data fusion scenarios, as indicated by the highest number of green checkmarks for PVM-RF. Furthermore, the prediction methods, especially the non-parametric SL prediction approaches DT and RF, are generally a chance and a risk for data fusions at the same time, depending on whether the CIA is fulfilled. If the CIA holds, then RM, DT and RF tend to overestimate associations, which in turn is useful to some extent if the CIA is violated. Overall, RF yields the most promising results in case of CIA violation. Nevertheless, it is still advised to always seek to circumvent the CIA, since none of the potential data fusion approaches can guarantee an appealing result under CIA violation. In this respect, the proposal of Donatiello et al. (2016) to incorporate common **X** variables that are closely related to the specific **Y** or **Z** variables serves as quite practical and pragmatic approach.

7.2 Outlook

Future research could extend to further data fusion scenarios. For the investigation of further scenarios, the introduced scenario classification into explicit, implicit and imputation scenarios provides orientation. In terms of conventional samples to be fused, different sample designs could be another explicit scenario. Substantially different sample designs could limit the comparability of the studies and have a different impact on the performance of different data fusion methods. Regarding the explicit CIA-related scenarios, the analyses regarding a violation of the CIA could be extended to a scenario in which the correlation under CIA is higher than the true correlation and thus the CIA is violated in the direction of 1 or -1 (instead of in the direction of 0). Although this is a rare case, a bias of the CIA towards 1 or -1 is theoretically possible. As an additional imputation scenario, the use of multiple imputation (MI) (Rubin 1978, 1987) could be considered, provided inferential statistical inferences are to be drawn.

Accordingly, the developed PVM algorithm could be extended to MI in the future if the analytical objective is of an inferential nature. Moreover, the multivariate extension of PVM provides a combination of distances and allows for the identification of nearest neighbours across all **Z** variables. However, contrary to the multivariate PMM implementation proposed by Little (1988), no gradation is conducted along the explanatory power of the common **X** variables with respect to the specific **Z** variables to be fused. Future research could complement the multivariate PVM solution by including a gradation process based on the explanatory power of the common **X** variables. For example, appropriate gradations for PVM based on DT or RF could be conducted by means of the node purity. Here, distances between **Z** variables that imply stronger node purity could be penalised more severely than distances between **Z** variables that imply lower node purity. The estimated, cross-validated error of the predictions for the intermediate values could also serve as a basis for gradations in which the smaller the estimated error and thus the better the prediction result, the more strongly distances are penalised. Further research may also dedicate to investigate the proposed PVM approach to many other missing data problems. The advantage of PVM is its general nature. A variety of different prediction methods can underlie PVM. It is flexible with regard to the scale levels of the variable to be imputed (in contrast to PMM). And it provides a general solution for the univariate and multivariate imputation for any prediction method that underlies PVM.

For future data fusions in official statistics, but also in other fields of application, Table 7.1 provides an important orientation for a variety of different data constellations. It should be noted, however, that each data situation brings its own challenges. Therefore, the indications in Table 7.1 cannot claim general validity. However, they do provide a valuable reference for future data fusions, so that it can be directly assessed for practical application which data fusion method could be considered for a specific fusion scenario and could provide promising results. At the beginning of this thesis, we stated that the literature on concrete and practical data fusion methods appears diffuse and leaves practitioners from official statistics and other application areas with ambiguities rather than a concrete plethora of possible data fusion methods tailored to the underlying data situation. This thesis has been capable of addressing these ambiguities through comprehensive evaluations of classical versus statistical learning approaches, resulting in concrete implications regarding a variety of data fusion procedures in different data situations.

Appendix A

Relevant Tables

A.1 EU-SILC/ HBS

Table A.1: Means of $\widehat{\rho}_{\mathbf{Y}\widetilde{\mathbf{Z}}}$ under CIA Compliance

		$\widehat{\text{corr}}(Y_1, \widetilde{Z}_1)$	$\widehat{\text{corr}}(Y_1, \widetilde{Z}_2)$	$\widehat{\text{corr}}(Y_2, \widetilde{Z}_1)$	$\widehat{\text{corr}}(Y_2, \widetilde{Z}_2)$
	DHD	0.7392	0.6883	0.3051	0.3351
	RM	0.8656	0.8229	0.4160	0.4314
	PMM	0.8455	0.8022	0.4004	0.4218
$\mathbf{n_1}$	**DT**	0.8041	0.8028	0.4158	0.4107
	PVM-DT	0.7778	0.7681	0.4025	0.3936
	RF	0.8621	0.8030	0.3719	0.3994
	PVM-RF	0.8517	0.7982	0.3808	0.4108
	DHD	0.5717	0.5205	0.1737	0.1949
	RM	0.8767	0.8380	0.4526	0.4642
	PMM	0.8209	0.7787	0.3666	0.3774
$\mathbf{n_2}$	**DT**	0.7536	0.7494	0.3518	0.3469
	PVM-DT	0.7067	0.6984	0.3297	0.3235
	RF	0.8213	0.7675	0.2988	0.3210
	PVM-RF	0.8159	0.7680	0.3172	0.3350

Source: EU-SILC SUF DE (2015); EU-SILC SUF FR (2015).

Table A.2: RMSE of $\widehat{\rho}_{\mathbf{Y}\widetilde{\mathbf{Z}}}$ under CIA Compliance

		$\widehat{\text{corr}}(Y_1,\widetilde{Z}_1)$	$\widehat{\text{corr}}(Y_1,\widetilde{Z}_2)$	$\widehat{\text{corr}}(Y_2,\widetilde{Z}_1)$	$\widehat{\text{corr}}(Y_2,\widetilde{Z}_2)$
	DHD	0.1479	0.1820	0.1716	0.1815
	RM	0.0532	0.0680	0.1304	0.1356
	PMM	0.0590	0.0800	0.1226	0.1290
$\mathbf{n_1}$	**DT**	0.0906	0.0796	0.1127	0.1300
	PVM-DT	0.1115	0.1097	0.1205	0.1466
	RF	0.0428	0.0678	0.1050	0.1146
	PVM-RF	0.0463	0.0733	0.1105	0.1151
	DHD	0.3039	0.3397	0.2693	0.2945
	RM	0.0228	0.0296	0.0840	0.0816
	PMM	0.0567	0.0826	0.0925	0.1230
$\mathbf{n_2}$	**DT**	0.1188	0.1095	0.1101	0.1533
	PVM-DT	0.1702	0.1658	0.1274	0.1743
	RF	0.0611	0.0942	0.1502	0.1738
	PVM-RF	0.0644	0.0940	0.1358	0.1625

Source: EU-SILC SUF DE (2015); EU-SILC SUF FR (2015).

Table A.3: Means of $\widehat{\rho}_{\mathbf{X}\widetilde{\mathbf{Z}}}$ under CIA Compliance

		$\widehat{\text{corr}}(X_2,\widetilde{Z}_1)$	$\widehat{\text{corr}}(X_2,\widetilde{Z}_2)$	$\widehat{\text{corr}}(X_7,\widetilde{Z}_1)$	$\widehat{\text{corr}}(X_7,\widetilde{Z}_2)$
	DHD	−0.1454	−0.0370	0.7977	0.8000
	RM	−0.1205	−0.0246	0.9961	0.9929
	PMM	−0.1179	−0.0283	0.9667	0.9704
$\mathbf{n_1}$	**DT**	−0.0849	−0.0837	0.9627	0.9628
	PVM-DT	−0.0812	−0.0810	0.9308	0.9214
	RF	−0.1571	−0.0412	0.9543	0.9514
	PVM-RF	−0.1304	−0.0241	0.9536	0.9524
	DHD	−0.1851	−0.0723	0.5623	0.5657
	RM	−0.1130	−0.0222	0.9953	0.9921
	PMM	−0.1305	−0.0435	0.9281	0.9288
$\mathbf{n_2}$	**DT**	−0.0873	−0.0848	0.8861	0.8844
	PVM-DT	−0.0809	−0.0791	0.8321	0.8246
	RF	−0.1792	−0.0618	0.8813	0.8840
	PVM-RF	−0.1483	−0.0468	0.8867	0.8887

Source: EU-SILC SUF DE (2015); EU-SILC SUF FR (2015).

Table A.4: RMSE of $\widehat{\rho}_{\mathbf{X}\widetilde{\mathbf{Z}}}$ under CIA Compliance

		$\widehat{\text{corr}}(X_2,\widetilde{Z}_1)$	$\widehat{\text{corr}}(X_2,\widetilde{Z}_2)$	$\widehat{\text{corr}}(X_7,\widetilde{Z}_1)$	$\widehat{\text{corr}}(X_7,\widetilde{Z}_2)$
	DHD	0.0584	0.0463	0.1902	0.1916
	RM	0.0497	0.0444	0.0262	0.0193
	PMM	0.0473	0.0450	0.0232	0.0253
$\mathbf{n_1}$	**DT**	0.0507	0.0775	0.0295	0.0318
	PVM-DT	0.0528	0.0754	0.0581	0.0723
	RF	0.0661	0.0488	0.0448	0.0472
	PVM-RF	0.0500	0.0442	0.0410	0.0435
	DHD	0.0844	0.0622	0.4156	0.4161
	RM	0.0206	0.0159	0.0256	0.0191
	PMM	0.0294	0.0292	0.0595	0.0642
$\mathbf{n_2}$	**DT**	0.0263	0.0662	0.0964	0.1014
	PVM-DT	0.0326	0.0610	0.1546	0.1661
	RF	0.0745	0.0470	0.1014	0.1017
	PVM-RF	0.0457	0.0343	0.0959	0.0974

Source: EU-SILC SUF DE (2015); EU-SILC SUF FR (2015).

Table A.5: Means of $\widehat{\rho}_{\mathbf{Y}\widetilde{\mathbf{Z}}}$ under CIA Violation

		$\widehat{\text{corr}}(Y_1,\widetilde{Z}_1)$	$\widehat{\text{corr}}(Y_1,\widetilde{Z}_2)$	$\widehat{\text{corr}}(Y_2,\widetilde{Z}_1)$	$\widehat{\text{corr}}(Y_2,\widetilde{Z}_2)$
	DHD	0.2933	0.2496	0.0398	0.0559
	RM	0.5599	0.4774	0.0731	0.1028
	PMM	0.2902	0.2472	0.0365	0.0520
$\mathbf{n_1}$	**DT**	0.5520	0.4717	0.0489	0.0741
	PVM-DT	0.2562	0.2237	0.0254	0.0354
	RF	0.5769	0.4896	0.0639	0.0990
	PVM-RF	0.3442	0.3000	0.0532	0.0707
	DHD	0.2821	0.2369	0.0278	0.0405
	RM	0.5151	0.4397	0.0595	0.0822
	PMM	0.2523	0.2118	0.0185	0.0304
$\mathbf{n_2}$	**DT**	0.4988	0.4012	0.0315	0.0651
	PVM-DT	0.2467	0.2136	0.0203	0.0289
	RF	0.5236	0.4446	0.0456	0.0736
	PVM-RF	0.3525	0.3080	0.0412	0.0563

Source: EU-SILC SUF DE (2015); EU-SILC SUF FR (2015).

Table A.6: RMSE of $\widehat{\rho}_{\mathbf{Y}\widetilde{\mathbf{Z}}}$ under CIA Violation

		$\widehat{\text{corr}}(Y_1,\widetilde{Z}_1)$	$\widehat{\text{corr}}(Y_1,\widetilde{Z}_2)$	$\widehat{\text{corr}}(Y_2,\widetilde{Z}_1)$	$\widehat{\text{corr}}(Y_2,\widetilde{Z}_2)$
	DHD	0.5791	0.6078	0.4022	0.4322
	RM	0.3154	0.3808	0.3654	0.3821
	PMM	0.5824	0.6104	0.4044	0.4352
$\mathbf{n_1}$	**DT**	0.3250	0.3875	0.3900	0.4111
	PVM-DT	0.6155	0.6332	0.4146	0.4512
	RF	0.2999	0.3694	0.3763	0.3872
	PVM-RF	0.5293	0.5585	0.3900	0.4184
	DHD	0.5880	0.6186	0.4090	0.4432
	RM	0.3555	0.4165	0.3772	0.4016
	PMM	0.6181	0.6439	0.4182	0.4533
$\mathbf{n_2}$	**DT**	0.3769	0.4635	0.4058	0.4191
	PVM-DT	0.6240	0.6422	0.4163	0.4546
	RF	0.3486	0.4129	0.3914	0.4104
	PVM-RF	0.5185	0.5485	0.3958	0.4277

Source: EU-SILC SUF DE (2015); EU-SILC SUF FR (2015).

Table A.7: Means of $\widehat{\rho}_{\mathbf{X}\widetilde{\mathbf{Z}}}$ under CIA Violation

		$\widehat{\text{corr}}(X_2,\widetilde{Z}_1)$	$\widehat{\text{corr}}(X_2,\widetilde{Z}_2)$	$\widehat{\text{corr}}(X_7,\widetilde{Z}_1)$	$\widehat{\text{corr}}(X_7,\widetilde{Z}_2)$
	DHD	−0.1256	−0.0333	0.2330	0.2285
	RM	−0.2373	−0.0473	0.4375	0.4312
	PMM	−0.1310	−0.0401	0.2277	0.2235
$\mathbf{n_1}$	**DT**	−0.3020	−0.1305	0.4086	0.4059
	PVM-DT	−0.1308	−0.0743	0.1926	0.1885
	RF	−0.2642	−0.0728	0.4416	0.4357
	PVM-RF	−0.1243	−0.0303	0.2771	0.2730
	DHD	−0.1438	−0.0519	0.2109	0.2044
	RM	−0.2261	−0.0446	0.3985	0.3931
	PMM	−0.1423	−0.0523	0.1804	0.1766
$\mathbf{n_2}$	**DT**	−0.2902	−0.0910	0.3540	0.3566
	PVM-DT	−0.1269	−0.0700	0.1819	0.1769
	RF	−0.2694	−0.0821	0.3811	0.3795
	PVM-RF	−0.1418	−0.0417	0.2706	0.2681

Source: EU-SILC SUF DE (2015); EU-SILC SUF FR (2015).

Table A.8: RMSE of $\widehat{\rho}_{\mathbf{X}\widetilde{\mathbf{Z}}}$ under CIA Violation

		$\widehat{\text{corr}}(X_2, \widetilde{Z}_1)$	$\widehat{\text{corr}}(X_2, \widetilde{Z}_2)$	$\widehat{\text{corr}}(X_7, \widetilde{Z}_1)$	$\widehat{\text{corr}}(X_7, \widetilde{Z}_2)$
	DHD	0.0499	0.0462	0.7402	0.7485
	RM	0.1388	0.0604	0.5362	0.5461
	PMM	0.0546	0.0521	0.7456	0.7535
$\mathbf{n_1}$	**DT**	0.2034	0.1238	0.5653	0.5715
	PVM-DT	0.0579	0.0736	0.7800	0.7877
	RF	0.1634	0.0730	0.5326	0.5419
	PVM-RF	0.0557	0.0533	0.6968	0.7046
	DHD	0.0610	0.0568	0.7600	0.7703
	RM	0.1439	0.0882	0.5724	0.5816
	PMM	0.0540	0.0521	0.7911	0.7987
$\mathbf{n_2}$	**DT**	0.2216	0.1558	0.6176	0.6188
	PVM-DT	0.0583	0.0722	0.7893	0.7980
	RF	0.1766	0.1005	0.5901	0.5955
	PVM-RF	0.0685	0.0640	0.7007	0.7069

Source: EU-SILC SUF DE (2015); EU-SILC SUF FR (2015).

A.2 TS/ MC

Table A.9: Means of $\widehat{\beta}_{educ}$ and $\widehat{\beta}_{w-time}$ under CIA Compliance

		$\widehat{\beta}_{educ_m}$	$\widehat{\beta}_{educ_h}$	$\widehat{\beta}_{w-time_p}$	$\widehat{\beta}_{w-time_n}$
	RM	1.0700	1.7585	−0.5068	−0.4797
	PMM	0.0551	0.1329	−0.0746	−0.6115
	DT	0.5565	1.2897	−0.5332	−0.4816
$\mathbf{n_1}$	**PVM-DT**	0.1463	0.3438	−0.2172	−0.6007
	RF	0.7121	1.2989	−0.5053	−0.5391
	PVM-RF	0.2204	0.4761	−0.2757	−0.6048
	RM	0.7904	1.3925	−0.4893	−0.4866
	PMM	0.1522	0.3624	−0.2075	−0.6109
	DT	0.6009	1.2178	−0.4803	−0.4641
$\mathbf{n_2}$	**PVM-DT**	0.1680	0.3902	−0.2342	−0.5983
	RF	0.6200	1.1675	−0.4930	−0.5272
	PVM-RF	0.3346	0.7924	−0.3484	−0.5636

Source: Microcensus (2014).

Table A.10: RMSE of $\widehat{\beta}_{educ}$ and $\widehat{\beta}_{w-time}$ under CIA Compliance

		$\widehat{\beta}_{educ_m}$	$\widehat{\beta}_{educ_h}$	$\widehat{\beta}_{w-time_p}$	$\widehat{\beta}_{w-time_n}$
$\mathbf{n_1}$	**RM**	0.7529	1.1599	0.0904	0.1520
	PMM	0.2790	0.4861	0.3680	0.0602
	DT	0.2549	0.6934	0.1129	0.1531
	PVM-DT	0.1885	0.2675	0.2244	0.0617
	RF	0.4012	0.7032	0.0869	0.1008
	PVM-RF	0.1227	0.1436	0.1666	0.0591
$\mathbf{n_2}$	**RM**	0.4806	0.7997	0.0965	0.1510
	PMM	0.1858	0.2732	0.2399	0.0279
	DT	0.3102	0.6503	0.1646	0.1783
	PVM-DT	0.1736	0.2274	0.2189	0.0379
	RF	0.3040	0.5688	0.0927	0.1062
	PVM-RF	0.0677	0.1989	0.1075	0.0697

Source: Microcensus (2014).

Table A.11: Means of $\widehat{\beta}_{educ}$ and $\widehat{\beta}_{w-time}$ under CIA Violation

		$\widehat{\beta}_{educ_m}$	$\widehat{\beta}_{educ_h}$	$\widehat{\beta}_{w-time_p}$	$\widehat{\beta}_{w-time_n}$
$\mathbf{n_1}$	**RM**	0.0619	0.1376	−0.0785	−0.6115
	PMM	0.0019	0.0026	−0.0253	−0.6193
	DT	0.5404	0.5907	–	−0.5928
	PVM-DT	0.0863	0.1066	0.0173	−0.5945
	RF	0.5375	0.7743	0.0097	−0.5774
	PVM-RF	0.0967	0.1366	0.0085	−0.5929
$\mathbf{n_2}$	**RM**	0.0784	0.1074	−0.0804	−0.6063
	PMM	0.0081	0.0087	−0.0206	−0.6197
	DT	0.5307	0.5993	−0.0959	−0.5971
	PVM-DT	0.0867	0.1139	0.0040	−0.5987
	RF	0.3937	0.5167	−0.0674	−0.5849
	PVM-RF	0.1164	0.1759	−0.0099	−0.5981

Source: Microcensus (2014).

Table A.12: RMSE of $\widehat{\beta}_{educ}$ and $\widehat{\beta}_{w-time}$ under CIA Violation

		$\widehat{\beta}_{educ_m}$	$\widehat{\beta}_{educ_h}$	$\widehat{\beta}_{w-time_p}$	$\widehat{\beta}_{w-time_n}$
$\mathbf{n_1}$	**RM**	0.3418	0.5192	0.3763	0.0591
	PMM	0.3295	0.6056	0.4128	0.0580
	DT	0.2312	0.5016	–	0.0651
	PVM-DT	0.2469	0.5022	0.4551	0.0653
	RF	0.2425	0.2103	0.4538	0.0750
	PVM-RF	0.2370	0.4726	0.4464	0.0665
$\mathbf{n_2}$	**RM**	0.2605	0.5064	0.3583	0.0244
	PMM	0.3198	0.5974	0.4152	0.0094
	DT	0.2245	0.1705	0.3432	0.0357
	PVM-DT	0.2422	0.4929	0.4400	0.0285
	RF	0.0938	0.1122	0.3703	0.0440
	PVM-RF	0.2123	0.4307	0.4260	0.0290

Source: Microcensus (2014).

Table A.13: RSD of Income Medians by Education and Working Time

		Education			**Working Time**		
		low	middle	high	none	part-time	full-time
CIA Compliance	**RM**	580	328	203	260	351	276
	PMM	986	191	685	257	665	357
	DT	360	358	196	261	384	298
	PVM-DT	574	94	408	266	237	271
	RF	440	294	156	261	358	265
	PVM-RF	510	88	401	265	200	279
CIA Violation	**RM**	838	267	812	261	802	364
	PMM	1059	220	724	258	703	365
	DT	255	255	843	261	–	358
	PVM-DT	938	230	714	264	708	362
	RF	346	223	721	261	565	350
	PVM-RF	981	225	717	264	673	356

Source: Microcensus (2014); Tax Statistics (2014).

Table A.14: Adjusted R^2 of Income Models with MC and Fused Data

Microcensus		**Fused Data Tax Statistics/ Microcensus**						
without educ & w-time	**with educ & w-time**	**without educ & w-time**	**RM**	**PMM**	**DT**	**PVM (DT)**	**RF**	**PVM (RF)**
0.1581	0.4229	0.0789	0.6482 [a]	0.1336 [a]	0.5091 [a]	0.2090 [a]	0.4750 [a]	0.2523 [a]
			0.1325 [b]	0.1297 [b]	0.1685 [b]	0.1325 [b]	0.1847 [b]	0.1344 [b]

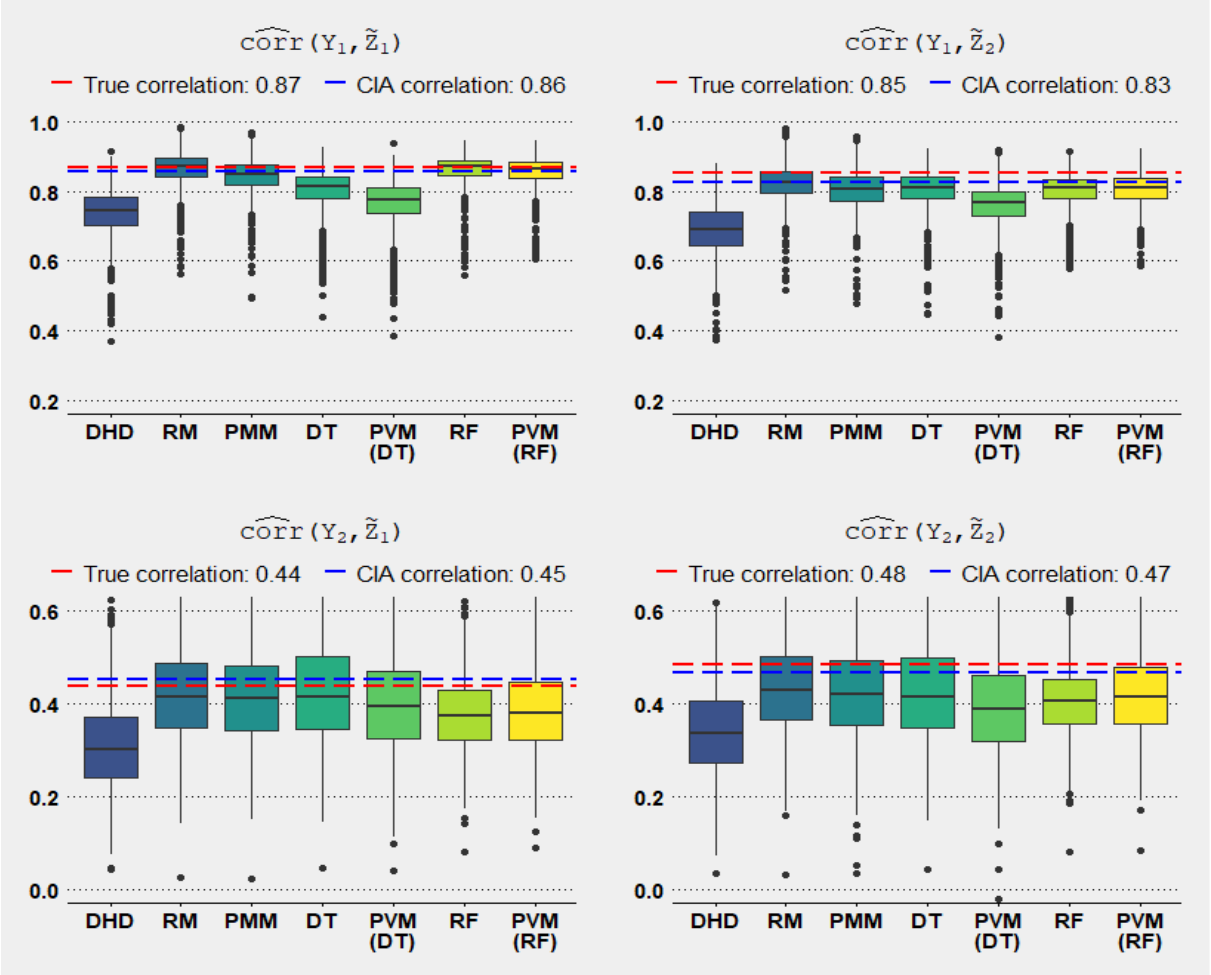
[a] CIA compliance;
[b] CIA violation.

Source: Microcensus (2014); Tax Statistics (2014).

Appendix B

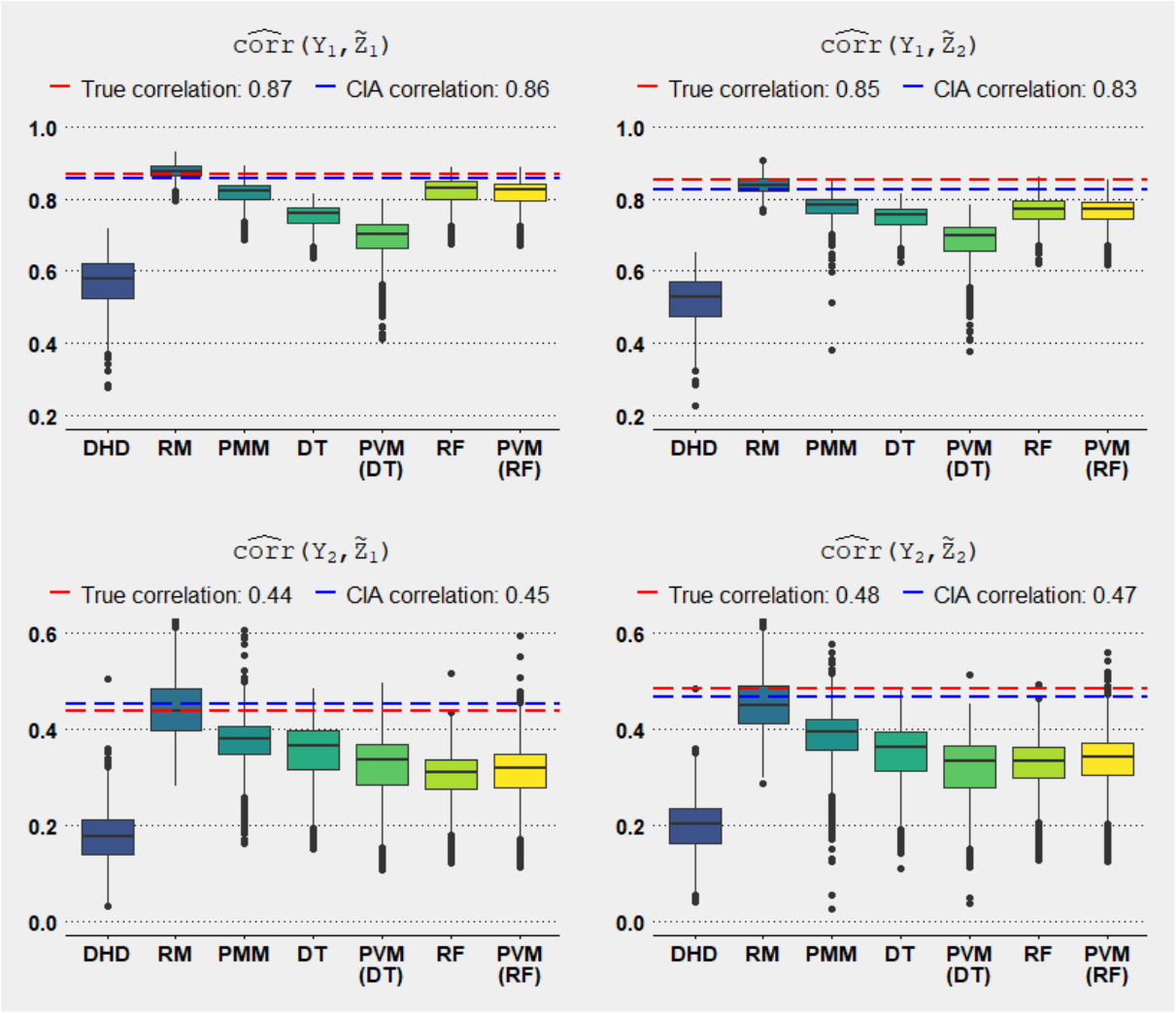
Results for Univariate PMM and PVM

B.1 EU-SILC/ HBS



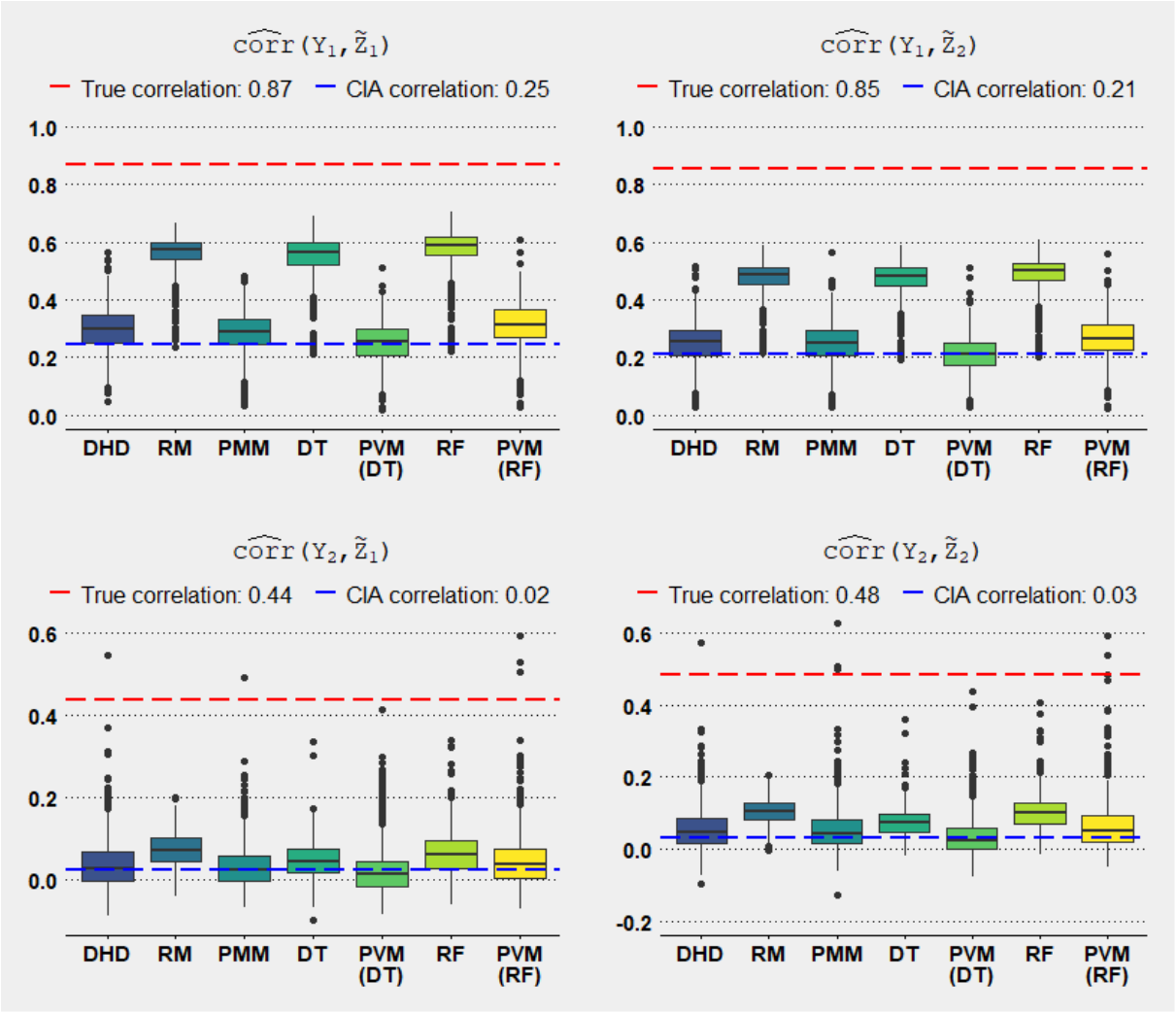
Source: EU-SILC SUF DE (2015); EU-SILC SUF FR (2015).

Figure B.1: MC distributions for $\widehat{\rho}_{\mathbf{Y}\tilde{\mathbf{Z}}}$ with n_1, CIA Compliance, Univariate PMM/ PVM



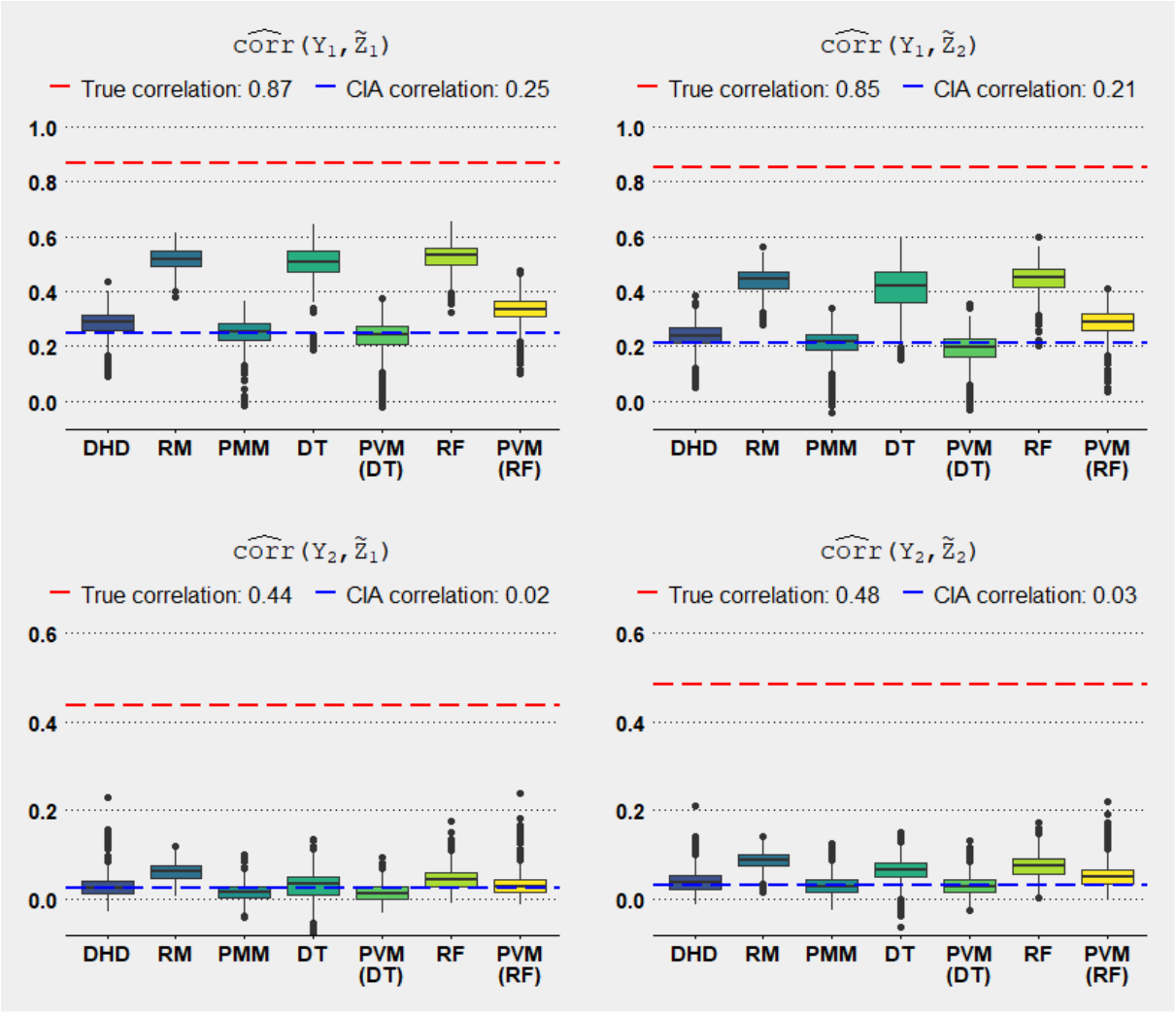
Source: EU-SILC SUF DE (2015); EU-SILC SUF FR (2015).

Figure B.2: MC distributions for $\widehat{\rho}_{\mathbf{Y}\widetilde{\mathbf{Z}}}$ with n_2, CIA Compliance, Univariate PMM/ PVM



Source: EU-SILC SUF DE (2015); EU-SILC SUF FR (2015).

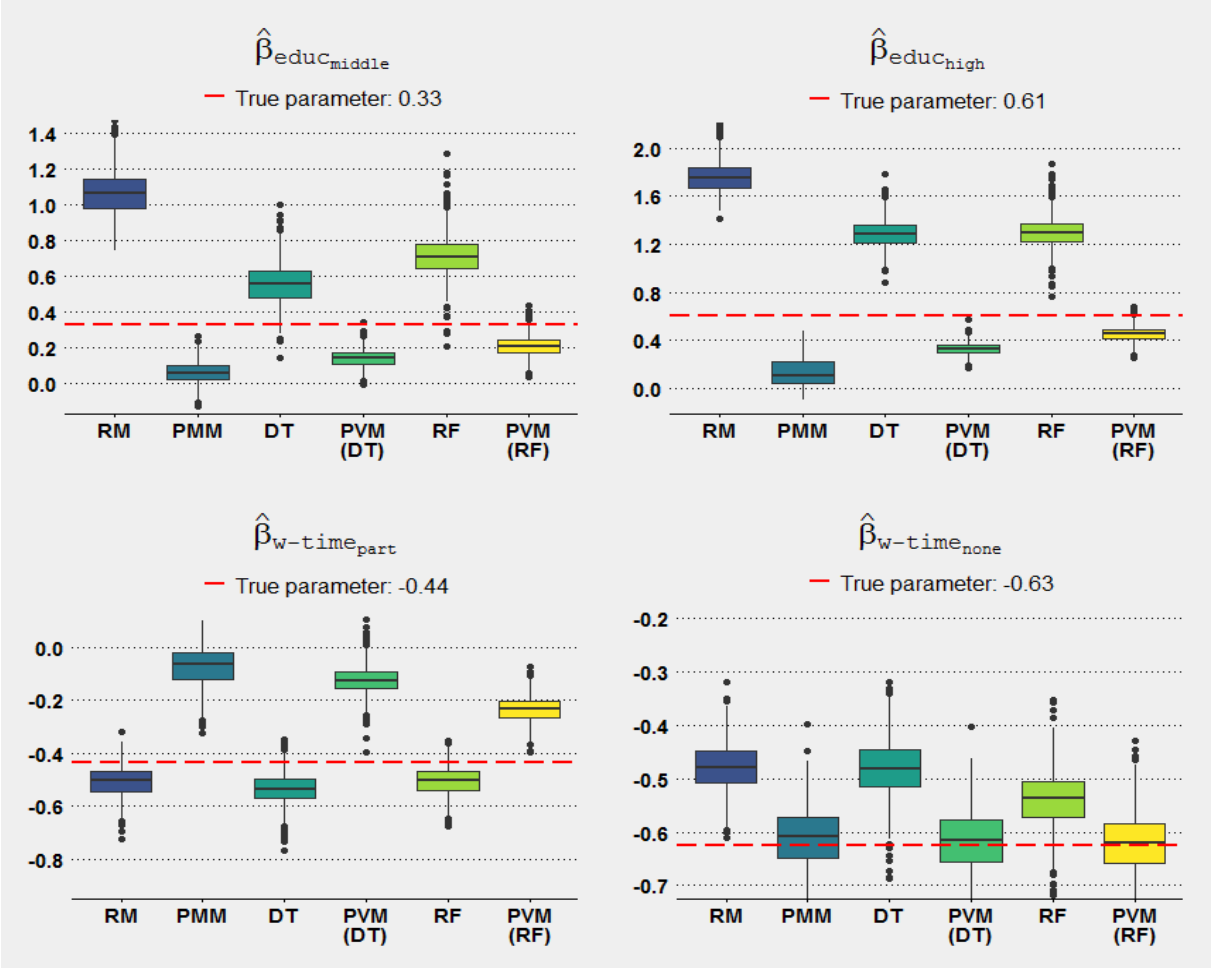
Figure B.3: MC distributions for $\widehat{\rho}_{\mathbf{Y}\widetilde{\mathbf{Z}}}$ with n_1, CIA Violation, Univariate PMM/ PVM



Source: EU-SILC SUF DE (2015); EU-SILC SUF FR (2015).

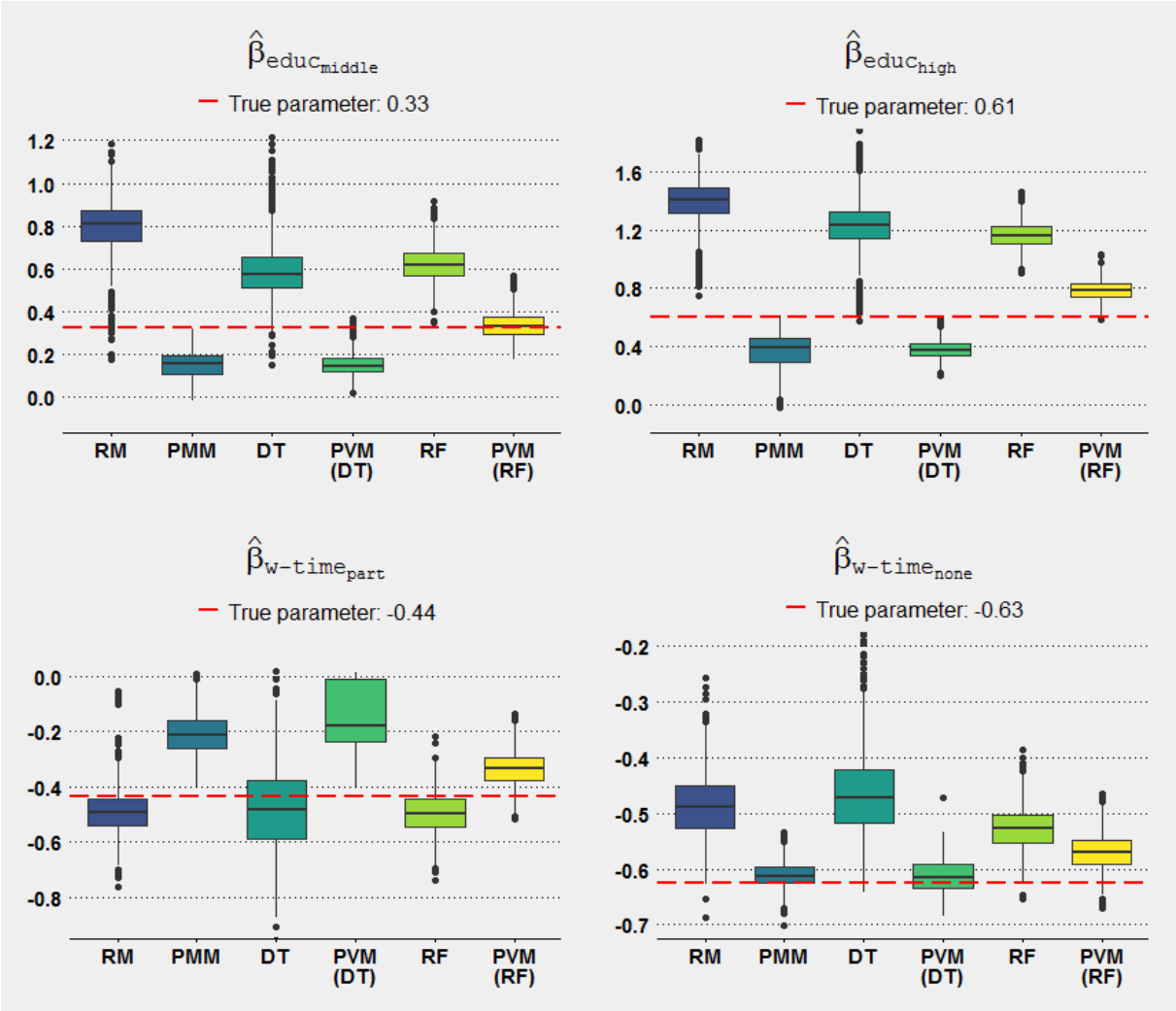
Figure B.4: MC distributions for $\widehat{\rho}_{\mathbf{Y}\widetilde{\mathbf{Z}}}$ with n_2, CIA Violation, Univariate PMM/ PVM

B.2 TS/ MC



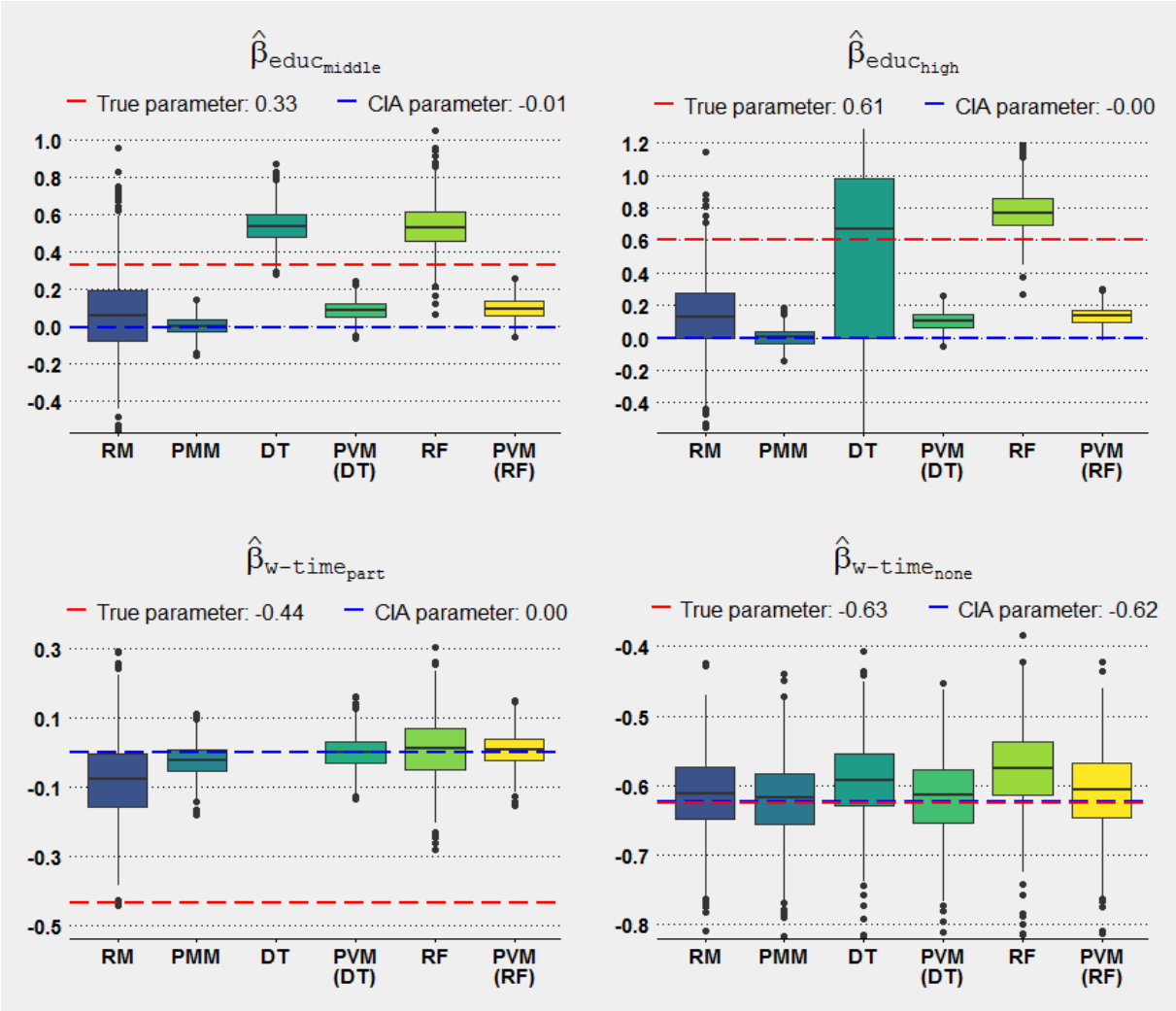
Source: Microcensus (2014).

Figure B.5: MC distributions for $\widehat{\beta}_{educ}$ and $\widehat{\beta}_{w-time}$ with n_1, CIA Compliance, Univariate PVM



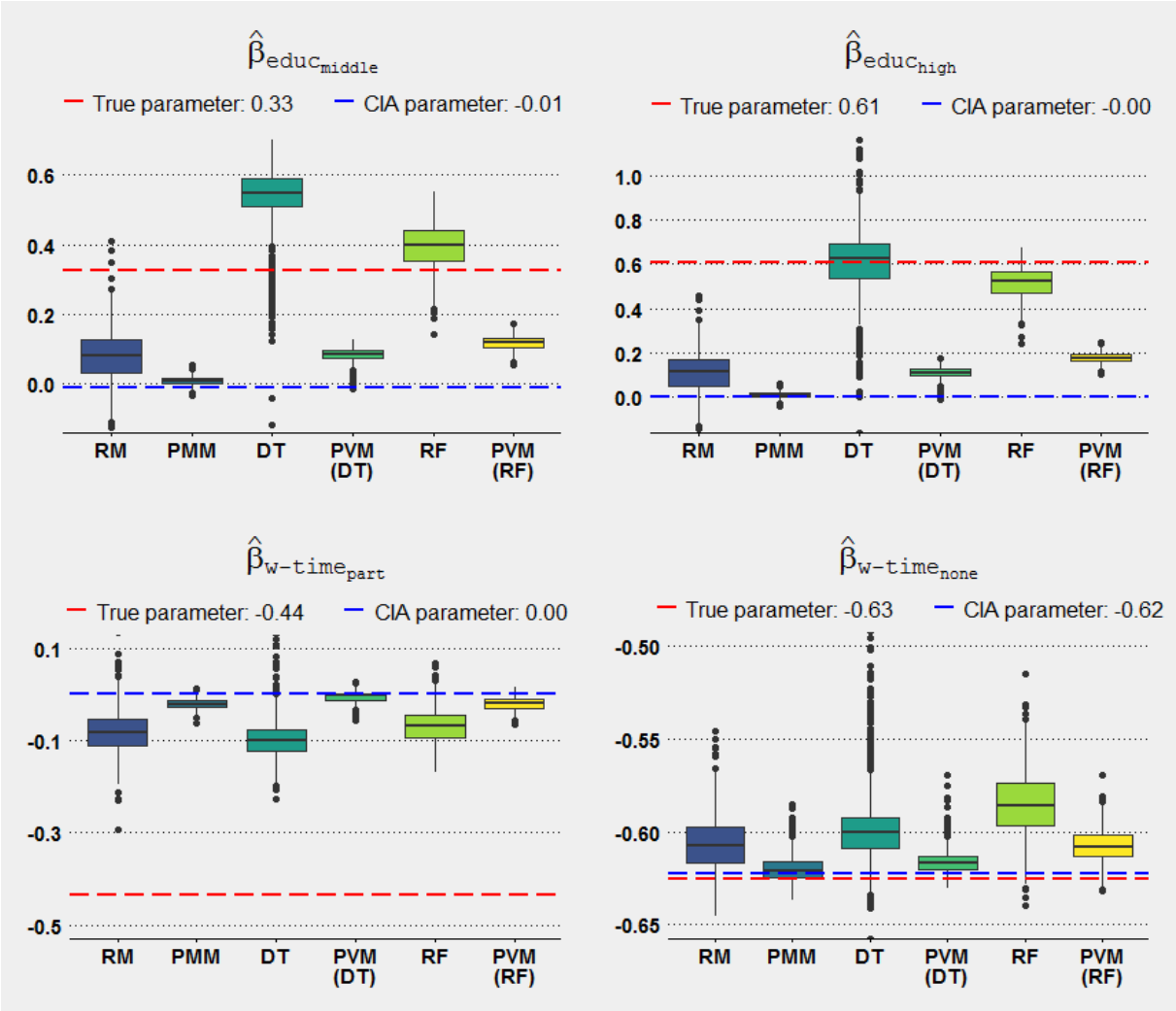
Source: Microcensus (2014).

Figure B.6: MC distributions for $\widehat{\beta}_{educ}$ and $\widehat{\beta}_{w-time}$ with n_2, CIA Compliance, Univariate PVM



Source: Microcensus (2014).

Figure B.7: MC distributions for $\widehat{\beta}_{educ}$ and $\widehat{\beta}_{w-time}$ with n_1, CIA Violation, Univariate PVM



Source: Microcensus (2014).

Figure B.8: MC distributions for $\widehat{\beta}_{educ}$ and $\widehat{\beta}_{w-time}$ with n_2, CIA Violation, Univariate PVM

Bibliography

ALBAYRAK, Ö. and MASTERSON, T. (2017). *Quality of Statistical Match of Household Budget Survey and SILC for Turkey*. Working Paper 885, Levy Economics Institute of Bard College.

ANDRIDGE, R. R. and LITTLE, R. J. A. (2010). A Review of Hot Deck Imputation for Survey Non-response. *International Statistical Review*, **78** (1), pp. 40–64.

ANGEL, S., DISSLBACHER, F., HUMER, S. and SCHNETZER, M. (2019). What did you really earn last year? Explaining measurement error in survey income data. *Journal of the Royal Statistical Society: Series A*, **182** (4), pp. 1411–1437.

ATKINSON, A. B. (2007). Measuring Top Incomes: Methodological Issues. In A. B. Atkinson and T. Piketty (eds.), *Top Incomes over the Twentieth Century: A Contrast Between European and English-Speaking Countries*, 1st edn., Oxford: Oxford University Press, pp. 18–42.

ATKINSON, A. B. and BOURGUIGNON, F. (2014). *Handbook of Income Distribution*. Amsterdam: Elsevier, 2nd edn.

BACH, S., CORNEO, G. and STEINER, V. (2009). From bottom to top: The entire income distribution in Germany, 1992-2003. *Review of Income and Wealth*, **55** (2), pp. 303–330.

BARRY, J. (1988). An investigation of statistical matching. *Journal of Applied Statistics*, **15** (3), pp. 275–283.

BARTELS, C. and SCHRÖDER, C. (2016). Zur Entwicklung von Top-Einkommen in Deutschland seit 2001. *DIW-Wochenbericht*, **83** (1), pp. 3–9.

BLANCHET, T. (2018). `gpinter`: Applying Generalized Pareto Interpolation with gpinter. `R` package. URL: `https://github.com/thomasblanchet/gpinter` (accessed October 2022).

BLANCHET, T., FOURNIER, J. and PIKETTY, T. (2022). Generalized pareto curves: Theory and applications. *Review of Income and Wealth*, **68** (1), pp. 263–288.

BMAS (2017). *Lebenslagen in Deutschland: Der Fünfte Armuts- und Reichtumsbericht der Bundesregierung*. Bundesministerium für Arbeit und Soziales (BMAS), URL: `https://www.armuts-und-reichtumsbericht.de/SharedDocs/Downloads/Berichte/5-arb-langfassung.pdf?__blob=publicationFile&v=6` (accessed October 2022).

BREIMAN, L. (1996). Bagging Predictors. *Machine Learning*, **24** (2), pp. 123–140.

BREIMAN, L. (2001). Random Forests. *Machine Learning*, **45** (1), pp. 5–32.

BREIMAN, L., CUTLER, A., LIAW, A. and WIENER, M. (2022). `randomForest`: Breiman and Cutler's Random Forests for Classification and Regression. `R` package. URL: `https://cran.r-project.org/web/packages/randomForest/randomForest.pdf` (accessed October 2022).

BREIMAN, L., FRIEDMAN, J., STONE, C. J. and OLSHEN, R. A. (1984). *Classification and Regression Trees*. Boca Raton: Chapman & Hall/CRC.

BRZEZINSKI, M., MYCK, M. and NAJSZTUB, M. (2019). *Reevaluating Distributional Consequences of the Transition to Market Economy in Poland: New Results from Combined Household Survey and Tax Return Data*. IZA Discussion Papers 12734, Institute of Labor Economics (IZA).

BURGETTE, L. F. and REITER, J. P. (2010). Multiple Imputation for Missing Data via Sequential Regression Trees. *American Journal of Epidemiology*, **172** (9), pp. 1070–1076.

BURKHAUSER, R. V., HÉRAULT, N., JENKINS, S. P. and WILKINS, R. (2018). Top incomes and inequality in the UK: reconciling estimates from household survey and tax return data. *Oxford Economic Papers*, **70** (2), pp. 301–326.

CONTI, P. L., MARELLA, D. and SCANU, M. (2012). Uncertainty Analysis in Statistical Matching. *Journal of Official Statistics*, **28** (1), pp. 69–88.

COWELL, F. A. (2000). Measurement of Inequality. In A. B. Atkinson and F. Bourguignon (eds.), *Handbook of Income Distribution*, 1st edn., Amsterdam: Elsevier, pp. 87–166.

DALLA CHIARA, E., MENON, M. and PERALI, F. (2019). An Integrated Database to Measure Living Standards. *Journal of Official Statistics*, **35** (3), pp. 531–576.

D'AMBROSIO, A., ARIA, M. and SICILIANO, R. (2012). Accurate Tree-based Missing Data Imputation and Data Fusion within the Statistical Learning Paradigm. *Journal of Classification*, **29** (2), pp. 227–258.

DE MELLO, R. F. and PONTI, M. A. (2018). *Machine Learning. A Practical Approach on the Statistical Learning Theory*. Cham: Springer.

DEUTSCHER BUNDESTAG (2017). Sachstand Einkommensungleichheit und Armutsrisikoquote. Wissenschaftliche Dienste, WD 6 - 3000 - 071/17. URL: `https://www.bundestag.de/resource/blob/538870/8ca1d4131c81ce90b8af45a75381b747/WD-6-071-17-pdf-data.pdf` (accessed October 2022).

DEVILLE, J.-C. and SÄRNDAL, C.-E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, **87** (418), pp. 376–382.

DONATIELLO, G., D'ORAZIO, M., FRATTAROLA, D., RIZZI, A., SCANU, M. and SPAZIANI, M. (2014). Statistical Matching of Income and Consumption Expenditures. *International Journal of Economic Sciences*, **3** (3), pp. 50–65.

DONATIELLO, G., D'ORAZIO, M., FRATTAROLA, D., RIZZI, A., SCANU, M. and SPAZIANI, M. (2016). The role of the conditional independence assumption in statistically matching income and consumption. *Statistical Journal of the IAOS*, **32** (4), pp. 667–675.

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**, pp. 92–104.

D'ORAZIO, M. (2021). Distances with mixed-type variables, some modified gower's coefficients. *arXiv preprint arXiv:2101.02481*.

D'ORAZIO, M. (2022). `StatMatch`: Statistical Matching or Data Fusion. `R` package. URL: `https://cran.r-project.org/web/packages/StatMatch/StatMatch.pdf` (accessed October 2022).

D'ORAZIO, M., DI ZIO, M. and SCANU, M. (2006a). Statistical Matching for Categorical Data: Displaying Uncertainty and Using Logical Constraints. *Journal of Official Statistics*, **22** (1), pp. 137 –157.

D'ORAZIO, M., DI ZIO, M. and SCANU, M. (2006b). *Statistical Matching: Theory and Practice*. Chichester: John Wiley & Sons.

D'ORAZIO, M., FRATTAROLA, D., RIZZI, A., SCANU, M. and SPAZIANI, M. (2018). The statistical matching of EU-SILC and HBS at ISTAT: where do we stand for the production of official statistics. URL: `https://www.istat.it/it/files//2018/11/Scanu_original-paper.pdf` (accessed October 2022).

D'ORAZIO, M. (2019). Statistical learning in official statistics: The case of statistical matching. *Statistical Journal of the IAOS*, **35** (3), pp. 435–441.

EFRON, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, **7** (1), pp. 1–26.

EFRON, B. (2003). Second Thoughts on the Bootstrap. *Statistical Science*, **18** (2), pp. 135–140.

EFRON, B. and TIBSHIRANI, R. J. (1994). *An Introduction to the Bootstrap*. Boca Raton: Chapman & Hall/CRC.

EMMENEGGER, J. and MÜNNICH, R. (2023). Localising the Upper Tail: How Top Income Corrections Affect Measures of Regional Inequality. *Journal of Economics and Statistics*, **243** (3–4), pp. 285–317.

EMMENEGGER, J., MÜNNICH, R. and SCHALLER, J. (2023). Evaluating Data Fusion Methods to Improve Income Modeling. *Journal of Survey Statistics and Methodology*, **11** (3), pp. 643–667.

ENDRES, E., FINK, P. and AUGUSTIN, T. (2019). Imprecise Imputation: A Nonparametric Micro Approach Reflecting the Natural Uncertainty of Statistical Matching with Categorical Data. *Journal of Official Statistics*, **35** (3), pp. 599–624.

EU-SILC SUF DE (2015). European Union Statistics on Income and Living Conditions. Scientific Use File Germany.

EU-SILC SUF FR (2015). European Union Statistics on Income and Living Conditions. Scientific Use File France.

EUROSTAT (2013). *European household income by groups of households*. Methodologies and Working papers KS-RA-13-023, Eurostat.

EUROSTAT (2015). Household Budget Survey, 2015 Wave, EU Quality Report (Version 1). URL: https://ec.europa.eu/eurostat/documents/54431/1966394/HBS_EU_QualityReport_2015.pdf/72d7e310-c415-7806-93cc-e3bc7a49b596 (accessed October 2022).

EUROSTAT (2016). Methodological Guidelines and Description of EU-SILC Target Variables. 2015 operation (Version August 2016). URL: https://circabc.europa.eu/sd/a/afb4601b-4e5c-4f40-86bb-0c3d0d94aa12/DOCSILC065operation2015VERSION08-08-2016.pdf (accessed October 2022).

EUROSTAT (2020). Methodological Guidelines and Description of EU-SILC Target Variables. 2019 operation (Version February 2020). URL: https://circabc.europa.eu/sd/a/b862932f-2209-450f-a76d-9cfe842936b4/DOCSILC065%20operation%202019_V9.pdf (accessed October 2022).

FLOOD, L., KLEVMARKEN, A. and MITRUT, A. (2008). Chapter 8 The Income of the Baby Boomers'. In A. Klevmarken and B. Lindgren (eds.), *Simulating an Ageing Population: A Microsimulation Approach Applied to Sweden (Contributions to Economic Analysis, Vol. 285)*, Bingley: Emerald Group Publishing Limited, pp. 249–292.

FOSDICK, B. K., DEYOREO, M. and REITER, J. P. (2016). Categorical data fusion using auxiliary information. *The Annals of Applied Statistics*, **10** (4), pp. 1907–1929.

FREUND, Y. and SCHAPIRE, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, **55** (1), pp. 119–139.

FUSHIKI, T. (2011). Estimation of prediction error by using K-fold cross-validation. *Statistics and Computing*, **21** (2), pp. 137–146.

GABLER, S. (1997). Datenfusion. *ZUMA-Nachrichten*, **21** (40), pp. 81–92.

GIESEN, D., VELLA, M., BRADY, C. F., BROWN, P., RAVINDRA, D. and VAASEN-OTTEN, A. (2018). Response Burden Management for Establishment Surveys at Four National Statistical Institutes. *Journal of Official Statistics*, **34** (2), pp. 397–418.

GILULA, Z., MCCULLOCH, R. E. and ROSSI, P. E. (2006). A Direct Approach to Data Fusion. *Journal of Marketing Research*, **43** (1), pp. 73–83.

GOWER, J. C. (1971). A General Coefficient of Similarity and Some of Its Properties. *Biometrics*, **27** (4), pp. 857–871.

GREENE, W. H. (2020). *Econometric Analysis*. Harlow: Pearson, 8th edn.

HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J. H. and FRIEDMAN, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer, 2nd edn.

HAUGHTON, J. and KHANDKER, S. R. (2009). *Handbook on Poverty and Inequality*. Washington, DC: The World Bank.

HERZOG, T. H., SCHEUREN, F. and WINKLER, W. E. (2010). Record linkage. *Wiley Interdisciplinary Reviews: Computational Statistics*, **2** (5), pp. 535–543.

HO, T. K. (1995). Random decision forests. In *Proceedings of the Third International Conference on Document Analysis and Recognition*, Institute of Electrical and Electronics Engineers (IEEE), vol. 1, pp. 278–282.

HO, T. K. (1998). The Random Subspace Method for Constructing Decision Forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20** (8), pp. 832–844.

HOCHGÜRTEL, T. (2019). Einkommensanalysen mit dem Mikrozensus. *WISTA – Wirtschaft und Statistik*, **3/2019**, pp. 53–64.

HORNUNG, G. and SCHNABEL, C. (2009). Data protection in Germany I: The population census decision and the right to informational self-determination. *Computer Law & Security Review*, **25** (1), pp. 84–88.

JAMES, G., WITTEN, D., HASTIE, T. and TIBSHIRANI, R. (2021). *An Introduction to Statistical Learning: with Applications in R*. New York: Springer, 2nd edn.

KAMAKURA, W. A. and WEDEL, M. (1997). Statistical Data Fusion for Cross-Tabulation. *Journal of Marketing Research*, **34** (4), pp. 485–498.

KAMGAR, S., MEINFELDER, F., MÜNNICH, R. and NAVVABPOUR, H. (2020). Estimation within the new integrated system of household surveys in Germany. *Statistical Papers*, **61** (5), pp. 2091–2117.

KAUFMAN, L. and ROUSSEEUW, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken: John Wiley & Sons.

KIESL, H. and RÄSSLER, S. (2005). Techniken und Einsatzgebiete von Datenintegration und Datenfusion. In C. König, M. Stahl and E. Wiegand (eds.), *Datenfusion und Datenintegration: 6. Wissenschaftliche Tagung*, Tagungsberichte, Bonn: Informationszentrum Sozialwissenschaften, pp. 17–32.

KIESL, H. and RÄSSLER, S. (2006). *How Valid Can Data Fusion Be?* IAB Discussion Paper 15/2006, Institute for Employment Research of the Federal Employment Services (IAB).

KLEINKE, K. (2017). Multiple Imputation Under Violated Distributional Assumptions: A Systematic Evaluation of the Assumed Robustness of Predictive Mean Matching. *Journal of Educational and Behavioral Statistics*, **42** (4), pp. 371–404.

KOLLER-MEINFELDER, F. (2009). *Analysis of Incomplete Survey Data – Multiple Imputation via Bayesian Bootstrap Predictive Mean Matching*. Ph.D. thesis, University of Bamberg.

KOSCHNICK, W. J. (1995). *Standard-Lexikon für Mediaplanung und Mediaforschung in Deutschland*. München: K. G. Saur, 2nd edn.

LAMARCHE, P., OEHLER, F. and RIOBOO, I. (2020). European household's income, consumption and wealth. *Statistical Journal of the IAOS*, **36** (4), pp. 1175–1188.

LANDERMAN, L. R., LAND, K. C. and PIEPER, C. F. (1997). An Empirical Evaluation of the Predictive Mean Matching Method for Imputing Missing Values. *Sociological Methods & Research*, **26** (1), pp. 3–33.

LEE, N., SISSONS, P. and JONES, K. (2016). The Geography of Wage Inequality in British Cities. *Regional Studies*, **50** (10), pp. 1714–1727.

LEULESCU, A. and AGAFITEI, M. (2013). *Statistical matching: A model based approach for data integration*. Methodologies and Working papers KS-RA-13-020, Eurostat.

LI, J. and O'DONOGHUE, C. (2013). A survey of dynamic microsimulation models: Uses, model structure and methodology. *International Journal of Microsimulation*, **6** (2), pp. 3–55.

LITTLE, R. J. A. (1988). Missing-Data Adjustments in Large Surveys. *Journal of Business & Economic Statistics*, **6** (3), pp. 287–296.

LITTLE, R. J. A. and RUBIN, D. B. (2019). *Statistical Analysis with Missing Data*. Hoboken: John Wiley & Sons, 3rd edn.

LUMLEY, T. and MILLER, A. (2022). `leaps`: Regression Subset Selection. `R` package. URL: `https://cran.r-project.org/web/packages/leaps/leaps.pdf` (accessed October 2022).

MEINFELDER, F. (2013). Datenfusion: Theoretische Implikationen und praktische Umsetzung. In T. Riede, N. Ott, S. Bechthold, T. Schmidt, M. Eisele, B. Schimpl-Neimanns, F. Meinfelder, R. Münnich, J. P. Burgard and T. Zimmermann (eds.), *Weiterentwicklung der amtlichen Haushaltsstatistiken*, Berlin: Scivero, pp. 83–98.

MEINFELDER, F. and SCHALLER, J. (2022). Data Fusion for Joining Income and Consumption Information using Different Donor-Recipient Distance Metrics. *Journal of Official Statistics*, **38** (2), pp. 509–532.

MEINFELDER, F. and SCHNAPP, T. (2015). `BaBooN`: Bayesian Bootstrap Predictive Mean Matching – Multiple and Single Imputation for Discrete Data. `R` package. URL: `https://cran.r-project.org/web/packages/BaBooN/BaBooN.pdf` (accessed August 2022).

MENG, X.-L. (1994). Multiple-Imputation Inferences with Uncongenial Sources of Input. *Statistical Science*, **9** (4), pp. 538–558.

MICROCENSUS (2014). Microcensus, Germany.

MINCER, J. (1958). Investment in Human Capital and Personal Income Distribution. *Journal of Political Economy*, **66** (4), pp. 281–302.

MONTGOMERY, J. M. and OLIVELLA, S. (2018). Tree-Based Models for Political Science Data. *American Journal of Political Science*, **62** (3), pp. 729–744.

MORGAN, J. N. and SONQUIST, J. A. (1963). Problems in the Analysis of Survey Data, and a Proposal. *Journal of the American Statistical Association*, **58** (302), pp. 415–434.

MORIARITY, C. and SCHEUREN, F. (2001). Statistical Matching: A Paradigm for Assessing the Uncertainty in the Procedure. *Journal of Official Statistics*, **17** (3), pp. 407–422.

MORRIS, T. P., WHITE, I. R. and CROWTHER, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, **38** (11), pp. 2074–2102.

MÜNNICH, R., SCHNELL, R., BRENZEL, H., DIECKMANN, H., DRÄGER, S., EMMENEGGER, J., HÖCKER, P., KOPP, J., MERKLE, H., NEUFANG, K., OBERSNEIDER, M., REINHOLD, J., SCHALLER, J., SCHMAUS, S. and STEIN, P. (2021). A Population Based Regional Dynamic Microsimulation of Germany: The MikroSim Model. *methods, data, analyses*, **15** (2), pp. 241–264.

NELDER, J. A. and WEDDERBURN, R. W. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society: Series A (General)*, **135** (3), pp. 370–384.

OECD (2014). Indicator A7: What are the incentives to invest in education? In OECD (ed.), *Education at a Glance 2014: OECD Indicators*, OECD Publishing, pp. 150–170.

OKNER, B. (1972). Constructing a New Data Base from Existing Microdata Sets: The 1966 Merge File. In *Annals of Economic and Social Measurement, Volume 1, number 3*, National Bureau of Economic Research, Inc, pp. 325–362.

PANORI, A. and PSYCHARIS, Y. (2019). Exploring the Links Between Education and Income Inequality at the Municipal Level in Greece. *Applied Spatial Analysis and Policy*, **12** (1), pp. 101–126.

PAPADAKIS, M., TSAGRIS, M., DIMITRIADIS, M., FAFALIOS, S., TSAMARDINOS, I., FASIOLO, M., BORBOUDAKIS, G., BURKARDT, J., ZOU, C., LAKIOTAKI, K. and CHATZIPANTSIOU, C. (2022). `Rfast`: A Collection of Efficient and Extremely Fast R Functions. `R` package. URL: `https://cran.r-project.org/web/packages/Rfast/Rfast.pdf` (accessed October 2022).

PFEFFERMANN, D. and SIKOV, A. (2011). Imputation and Estimation under Nonignorable Nonresponse in Household Surveys with Missing Covariate Information. *Journal of Official Statistics*, **27** (2), pp. 181–209.

PIKETTY, T. (2015). About capital in the twenty-first century. *American Economic Review*, **105** (5), pp. 48–53.

QUINLAN, J. R. (1986). Induction of Decision Trees. *Machine Learning*, **1** (1), pp. 81–106.

R CORE TEAM (2022a). `parallel`. `R` package. URL: `https://stat.ethz.ch/R-manual/R-devel/library/parallel/doc/parallel.pdf` (accessed October 2022).

R CORE TEAM (2022b). `R`: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL: `https://www.R-project.org/` (accessed October 2022).

RAO, J. N. and SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, **79** (4), pp. 811–822.

RÄSSLER, S. (2002). *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*. New York: Springer.

RAVALLION, M. and CHEN, S. (1997). What Can New Survey Data Tell Us About Recent Changes in Distribution and Poverty? *The World Bank Economic Review*, **11** (2), pp. 357–382.

RIPLEY, B. and VENABLES, W. (2022). `nnet`: Feed-Forward Neural Networks and Multinomial Log-Linear Models. `R` package. URL: `https://cran.r-project.org/web/packages/nnet/nnet.pdf` (accessed October 2022).

RODGERS, W. L. (1984). An Evaluation of Statistical Matching. *Journal of Business & Economic Statistics*, **2** (1), pp. 91–102.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika*, **63** (3), pp. 581–592.

RUBIN, D. B. (1978). Multiple imputation in sample surveys – a phenomological bayesian approach to nonresponse. In *Proceedings of the Survey Research Method Section of the American Statistical Association*, pp. 20–40.

RUBIN, D. B. (1986). Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations. *Journal of Business & Economic Statistics*, **4** (1), pp. 87–94.

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

SERAFINO, P. and TONKIN, R. (2017). *Statistical matching of European Union statistics on income and living conditions (EU-SILC) and the household budget survey*. Statistical working papers KS-TC-16-026, Eurostat.

SIMS, C. A. (1972). Comments (on Okner 1972). *Annals of Economic and Social Measurement*, **1** (3), pp. 343–345.

SINGH, A., YADAV, A. and RANA, A. (2013). K-means with Three different Distance Metrics. *International Journal of Computer Applications*, **67** (10), pp. 13–17.

SINGH, A. C., MANTEL, H. J., KINACK M. D. and ROWE, G. (1993). Statistical Matching: Use of Auxiliary Information as an Alternative to the Conditional Independence Assumption. *Survey Methodology*, **19** (1), pp. 59–79.

SINGH, S. and GUPTA, P. (2014). Comparative Study ID3, Cart and C4.5 Decision Tree Algorithm: A Survey. *International Journal of Advanced Information Science and Technology (IJAIST)*, **27** (27), pp. 97–103.

SPAZIANI, M., FRATTAROLA, D. and D'ORAZIO, M. (2019). Integration of Survey Data in R Based on Machine Learning. *Romanian Statistical Review*, **3/2019**, pp. 5–16.

STATISTISCHE ÄMTER DER LÄNDER (2014). Schlüsselverzeichnis Mikrozensus 2014. URL: `https://www.forschungsdatenzentrum.de/sites/default/files/mz_2014_on-site_svz.pdf` (accessed October 2022).

STATISTISCHES BUNDESAMT (2015). Qualitätsbericht. Mikrozensus 2014. URL: `https://www.destatis.de/DE/Methoden/Qualitaet/Qualitaetsberichte/Bevoelkerung/mikrozensus-2014.pdf?__blob=publicationFile` (accessed October 2022).

STIGLITZ, J., SEN, A. and FITOUSSI, J. (2009). Report of the Commission on the Measurement of Economic Performance and Social Progress (CMEPSP). URL: `https://ec.europa.eu/eurostat/documents/8131721/8131772/Stiglitz-Sen-Fitoussi-Commission-report.pdf` (accessed October 2022).

TANG, F. and ISHWARAN, H. (2017). Random Forest Missing Data Algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, **10** (6), pp. 363–377.

TAX STATISTICS (2014). Wage and Income Tax Statistics, Germany.

THERNEAU, T., ATKINSON, B., RIPLEY, B. and RIPLEY, B. (2022). `Rpart`: Recursive partitioning for classification, regression and survival trees. An implementation of most of the functionality of the 1984 book by Breiman, Friedman, Olshen and Stone. `R` package. URL: `https://cran.r-project.org/web/packages/rpart/rpart.pdf` (accessed October 2022).

TOFAN, C. A. (2015). Decision tree method applied in cost-based decisions in an enterprise. *Procedia Economics and Finance*, **32**, pp. 1088–1092.

UÇAR, B., BETTI, G. *et al.* (2016). *Longitudinal statistical matching: transferring consumption expenditure from HBS to SILC panel survey*. Tech. Rep. 739, Department of Economics, University of Siena.

VAN BUUREN, S. (2018). *Flexible Imputation of Missing Data*. Boca Raton: CRC press, 2nd edn.

VAN BUUREN, S. (2022). `mice`: Multivariate Imputation by Chained Equations. `R` package. URL: `https://cran.r-project.org/web/packages/mice/mice.pdf` (accessed October 2022).

VAN BUUREN, S. and GROOTHUIS-OUDSHOORN, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, **45** (3), pp. 1–67.

VAN DER PUTTEN, P., KOK, J. N. and GUPTA, A. (2002). *Data Fusion Through Statistical Matching*. Working Paper 4342-02, MIT Sloan School of Management.

VAPNIK, V. N. (1999). An Overview of Statistical Learning Theory. *IEEE Transactions on Neural Networks*, **10** (5), pp. 988–999.

WEBBER, D. and TONKIN, R. (2013). *Statistical matching of EU-SILC and the Household Budget Survey to compare poverty estimates using income, expenditures and material deprivation*. Methodologies and Working papers KS-RA-13-007, Eurostat.

WHITE, I. R., DANIEL, R. and ROYSTON, P. (2010). Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables. *Computational Statistics & Data Analysis*, **54** (10), pp. 2267–2275.

WRIGHT, M. N., WAGER, S. and PROBST, P. (2022). `ranger`: A Fast Implementation of Random Forests. `R` package. URL: `https://cran.r-project.org/web/packages/ranger/ranger.pdf` (accessed October 2022).

WRIGHT, M. N. and ZIEGLER, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, **77** (1), pp. 1–17.

XIE, X. and MENG, X.-L. (2017). Dissecting multiple imputation from a multi-phase inference perspective: What happens when god's, imputer's and analyst's models are uncongenial? *Statistica Sinica*, **27** (4), pp. 1485–1545.

YANG, F., WANG, H.-Z., MI, H., LIN, C.-D. and CAI, W.-W. (2009). Using random forest for reliable classification and cost-sensitive learning for medical diagnosis. *BMC Bioinformatics*, **10** (1), pp. 1–14.

ZHANG, L.-C. (2015). On Proxy Variables and Categorical Data Fusion. *Journal of Official Statistics*, **31** (4), pp. 783–807.

Scientific Curriculum Vitae

Jannik Schaller
PhD candidate at Trier University

Scientific Career

11/2019 to 11/2022	**Research Associate and PhD Candidate** *Federal Statistical Office of Germany, Wiesbaden* Part of the research group 'Multi-sectoral Regional Microsimulation Model (MikroSim)', funded by the German Research Foundation
04/2017 to 09/2019	**Master of Science in Survey Statistics** **European Master in Official Statistics (EMOS)** *University of Bamberg*
07/2015 to 03/2019	**Scholarship** from the *Friedrich Ebert Foundation (FES)*
10/2013 to 03/2017	**Bachelor of Arts in Political Science** *University of Bamberg*

Scientific Awards

2020	Master Thesis: EMOS Master Thesis Award
2020	Master Thesis: Gerhard Fürst Award 2020 from the Federal Statistical Office of Germany in the 'Bachelor and Master Theses' category