UNIVERSITÄT TRIER

Jana Emmenegger
Ralf Münnich
Jannik Schaller

Evaluating Data Fusion Methods to Improve Income Modelling

Research Papers in Economics
No. 3/22

Evaluating Data Fusion Methods to Improve Income Modelling

Jana Emmenegger[*], Ralf Münnich[†] and Jannik Schaller[‡]

Abstract

Income is an important economic indicator to measure living standards and individual well-being. In Germany, there exist different data sources that yield ambiguous evidence when analysing the income distribution. The Tax Statistics (TS) – an income register recording the total population of more than 40 million taxpayers in Germany for the year 2014 – contains the most reliable income information covering the full income distribution. However, it offers only a limited range of socio-demographic variables essential for income analysis. We tackle this challenge by enriching the tax data with information on education and working time from the Microcensus. For that purpose, we examine two types of data fusion methods that seem suited for the specific data fusion scenario of the Tax Statistics and the Microcensus: Missing-data methods on the one hand and performant prediction models on the other hand. We conduct a simulation study and provide an empirical application comparing the proposed data fusion methods, and our results indicate that Multinomial Regression and Random Forest are the most suitable methods for our data fusion scenario.

Keywords: Statistical Matching, Multi-source Estimation, Missing Data, Income Analysis, Statistical Learning

[*]Federal Statistical Office of Germany (DESTATIS), E-mail address: jana.emmenegger@destatis.de

[†]Trier University, E-mail address: muennich@uni-trier.de

[‡]Federal Statistical Office of Germany (DESTATIS), E-mail address: jannik.schaller@destatis.de

1 Introduction

Income is a crucial indicator to assess the individual well-being which has been modelled extensively ever since the pioneering works of Mincer (1958). Furthermore, questions related to economic inequality (Cowell 2000), poverty risks (Ravallion and Chen 1997) as well as the concentration of top incomes (Atkinson 2007, Piketty 2015) typically use income as the principal measure of interest. Consistent, high-quality research on the abovementioned issues requires an integrated data base, that includes reliable information on the full distribution of individual incomes alongside sound records of socio-demographic variables.

However, the income data situation is challenging. Different types of data sets provide pieces of the overall picture of the income distribution that need to be interpreted with respect to the data set's strengths and weaknesses (Deutscher Bundestag 2017). While most official statistics and academic analyses of inequality utilise household survey samples (BMAS 2017), certain parts of the literature and many policy evaluations rely on tax income records (Piketty 2015). The two data sources yield inconsistent estimates of the development of the income distribution, which causes uncertainty regarding the interpretation of academic results as well as related policy implications.

A perfect income data base does not exist in Germany. The Tax Statistics (TS) is an income register that covers the total population of more than 40 million taxpayers. The data set reveals the full income distribution from the bottom to the top providing the most reliable income information that is available in Germany. However, the administrative data lacks important covariates regarding the social disaggregation like education, family structures, occupation, and working hours. Due to the limited amount of explanatory variables, the distributional strength of the TS is not exploited. Most income analyses rather use survey data, such as the Microcensus or the SOEP. However, the income variables in survey data are known to face several drawbacks, i.e. self-response bias, top censoring, reports of classified or heaped data (Angel *et al.* 2019).

The aim of this paper is to evaluate methods that provide an integrated data base by enriching the reliable tax income details with socio-demographic variables from survey data. We choose to add education and working time, since it is essential to study the interdependencies of income with these two variables for an improved understanding of the distribution of incomes. The central reporting base for these two socio-demographic variables is the German Microcensus (MC), a representative 1% sample of the population. More specifically, our aim is to enhance the Tax Statistics (recipient study) with information on the socio-demographic variables education and working time, from the Microcensus (donor study). This requires sophisticated and performant data fusion methods, as the observation units of both data sets do not contain a unique identifier, which is why direct record linkage is not possible. The objective of a data fusion, also known as statistical matching, is to jointly analyse variables from (at least) two different data sources, where each of the data

sources originally served a different purpose (see e.g. Rässler 2002; D'Orazio *et al.* 2006a).

Our procedure consists of two steps: First, we select different data fusion methods that might be appropriate for the specific data fusion setting of TS and MC, and we evaluate their performance within a simulation study. In this context, we provide a strategy to adequately incorporate the rough income information from the Microcensus within the data fusion process to overcome the inherent problem of the Conditional Independence Assumption (CIA) (see Sims 1972) that is implicitly made in common data fusion techniques. Second, we conduct the real empirical application of all the considered data fusion methods and evaluate the empirical performance by comparing the regional conditional income medians of both data sets. To build common grounds for the comparison, we reweight the income distribution from the Microcensus data to match the frequencies observed in the tax data.

The remainder of the paper proceeds as follows: Section 2 introduces the two data sources with their specific strengths and weaknesses and explains the strategy to reconcile the two data sets as well as possible. While the relevant data fusion methods are presented in Part 3, the subsequent Section 4 explains the simulation design and the empirical study. The results are summarised in Section 5, and finally, Part 6 concludes.

2 Data

2.1 The Tax Income Statistics and the Microcensus

The majority of the existing income research is based on survey data. They offer a wide range of socio-demographic variables and provide information on the household context. However, income from survey data faces major challenges including small sample sizes and difficulties of measurement errors, such as under-coverage, sparseness or under-reporting of the tails of the income distribution. These problems affect all of the existing German household survey samples that are commonly used for income analyses, i.e. SILC, the SOEP, the Household Budget Survey as well as the German Microcensus.

The Microcensus is the largest survey sample in Germany based on a regional sampling design. It provides reliable information on education and working time for all respondents. Central official reports on the population's educational attainment are based on the Microcensus (Statistisches Bundesamt 2020a). In addition, the labour force survey – which is currently included in the Microcensus – builds the basis for regular official reports and international comparisons of individual working time (Statistisches Bundesamt 2020b).

Not suffering from self-response bias or sampling errors, the administrative data of individual tax records from the Tax Statistics provides the best available income information of high quality. It covers the full distribution of taxpayers' incomes and its large scale allows for small-scale regional income analysis. We use the Tax Statistics (TS) – an income register recording the total population of more than 40 million taxpayers in Germany – for the year 2014. Our aim is to build income models from the tax data to generate synthetic incomes

for regional analysis purposes. A severe drawback that arises during the modelling is the lack of important socio-demographic variables.

Increasing individual educational attainment or working time are two of the most important factors to raise personal earnings. The positive effects of years of schooling on the income distribution are evident and have extensively been studied, starting from the early works of Mincer (1958), focusing on the returns to education. He accounts for the importance of working time by analysing hourly wages. Many recent studies examine the ways in which working times and educational attainments affect the income distribution (Haughton and Khandker 2009, Atkinson and Bourguignon 2014), also on a regional level (Lee *et al.* 2016, Panori and Psycharis 2019).

We evaluate the potentials of data fusion methods to enrich the tax data with information on education and working time from the Microcensus. While a few studies have combined SOEP survey data with incomes from the taxpayer panel (Bach *et al.* 2009 and Bartels and Metzing 2019), no previous study has added socio-demographic variables to the full register of tax records, to our knowledge. The advantage of enriching the tax data is that the individual records of the tax microdata are maintained and can directly be used for subsequent joint analysis of income and the added variables. Table 1 provides an overview of the data sets and the variables relevant for our data fusion study. Following common data fusion notations, we refer to the variables observed in both studies as X, while we denote the specific income information from the recipient study (TS) as Y and the socio-demographic variables from the donor study (MC), education and working time, as Z. The education variable consists of the categories low, middle and high, as the largest parts of income deviations can be explained by these three categories. Since we are ultimately also interested in exploiting the regional variation in incomes from the tax data for statistical analysis, we evaluate the data fusion success at the regional level.

Table 1: Overview of Available Data Sets and Variables

	Tax Statistics (TS)	**Microcensus (MC)**
Source	tax authorities	official survey
Coverage	all taxpayers	1% of the population
Observed units	taxunit (individual or married couple)	individuals in households
Sample size	12.9 million single tax units 10.4 million married tax units	155,000 single tax units 99,200 married tax units
Specific variables (Y and Z)	Y: detailed, high quality income information	Z_1: education (low, middle, high) Z_2: working time (none, part-, full-time)
Common variables (X)	sex (male/female - only considered for individual tax units) age (16-65 years) employment status (employee, civil servant, self-employed, other) federal state (16 levels) tax unit's net income (continuous) family type (married, single, single parent) number of kids (0, 1, 2, ..., 9 or more)	

2.2 Harmonising the Two Data Sources

In preparation of the data fusion of the Tax Statistics and the Microcensus, it is important to reconcile the two data sets as well as possible. Table 2 summarises the advantages and shortcomings of the two data sets that are used in our study. Main differences are detected in the frames, reporting units and income concepts. The following subsection explains how we ensure that both data sets cover the same population, income-sharing units and income definitions.

Frame While the Microcensus provides a 1% sample of the population, the tax data covers the full population of taxpayers. Given that the tax data only covers taxpayers, our analysis does not provide insights for the full population in Germany, but rather for the sub-population of taxpayers. In order to harmonise the frames of both data sources, we define taxpayers in the Microcensus as all individuals whose main income source is reported as one of the following: (1) income from work, (2) assets, savings, dividends, renting or leasing, (3) pensions, (4) wage-replacement benefits.

Two main groups of people are identified as non-taxpayers, namely people with main income sources from either income from parents, spouses or other relatives and recipients of unemployment benefits according to Hartz IV. For our entire analysis, we exclude non-taxpayers from the Microcensus so that it is comparable to the tax data. Since our primary

Table 2: Advantages and Shortcomings of Both Data Sets

Data set	**Tax Statistics (TS)**	**Microcensus (MC)**
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
Spatial scale	+ fully exploitable at municipality level	− limited by sample size
Timeliness	− late availability (after 3 years)	+ yearly
Costs	+ automatic data transmission	− response burden

research interest is the comparison of different data fusion techniques, we restrict both data sets to working age individual taxpayers aged older than 15 and younger than 66 years for this analysis.

Reporting units Defined by the German personal income tax system, the reporting units in the tax data are either individuals or married couples, whereas in the survey data the different available reporting income-sharing units correspond to individuals, families and households[1].

We construct the reporting units from the tax data in the MC based on legal status. As household affiliations are unknown in the tax data, the opposite direction is not feasible, i.e. it is impossible to construct MC household compositions in the TS. In the reconciled MC data, all married couples are assumed to be one tax unit with joint taxation and all other individuals are defined as a single tax unit. This means that a household with an unmarried couple corresponds to two tax units. Married couples from both data sets are not considered for this methodological study. Bartels and Metzing (2019) choose a similar strategy to construct tax units in the SOEP.

Income Concept The income definitions differ in the survey and tax data. While the Microcensus contains self-reports of individual and households total net income during the last month provided in 24 disjunct income classes, the tax data captures yearly amounts

[1]For information on the family and household concept of the Microcensus, see e.g. Lengerer *et al.* (2005).

of income concepts defined by the German Income Tax Act, such as total earnings or taxable income, inter alia. The definition of these income variables reflect the administration of the German personal income tax system. Due to tax reliefs, allowances, and advertising costs, none of these income concepts is directly comparable to the net income reported in the survey data. We therefore reconstruct economic net income in the tax data based on the recorded individual earnings as well as the total income taxes of each tax unit, following a similar procedure as Bach *et al.* (2009). Table 4 in the Appendix shows the calculation scheme of economic net income in the tax data. The procedure first derives economic gross income and then transforms it into net income by subtracting taxes, paid alimony, transfers and social security contributions, where the latter are micro-simulated based on the Social Insurance Code. To ensure comparability with the survey concept, we exclude negative and zero incomes which are not reported in the Microcensus. We cannot calculate total earnings in the Microcensus because the data lacks information on individual income components, taxes and transfers.

Income interpolation In order to overcome problems of the Conditional Independence Assumption (CIA) that is required in data fusion situations, we aim to incorporate the income information from the Microcensus as a common variable within the data fusion process. The metric variable provides information advantages by quantifying income differences as opposed to classified income data originally included in the MC. However, the observed income information from the Microcensus faces several drawbacks: It is self-reported, provided in 24 disjunct income classes, and refers to the last month rather than an annual average. The Microcensus is an intra-year survey which implies that the household net incomes of all months of a year enter with approximately the same weight into the annual results. A common criticism is that the respondents do not indicate irregular income components and transfers very well (Hochgürtel 2019). This may lead to under-coverage of incomes examined hereafter in Table 3. The under-coverage is especially severe at the tails of the distribution. For this reason, Emmenegger and Münnich (forthcoming) correct the top incomes from the Microcensus based on the income distribution from the German tax records. Due to these quality concerns, we aim at analysing the income distribution from the tax data as the main variable of interest. The combined data allows to perform distributional analyses and to exploit the large variance of the highly reliable administrative income data conditional on socio-demographic variables which cannot be considered from the TS or MC alone. For instance, top incomes, tax burdens or the impact of different taxation policies can be analysed conditional on socio-demographic indicators.

To generate a continuous income value as a common variable, we interpolate the tax unit's income using the Generalised Pareto interpolation method developed by Blanchet *et al.* (2017). We estimate the Pareto parameters based on frequencies of observed incomes from the 24 MC income classes using the `R` package `gpinter` (Blanchet 2018). Compared

to linear interpolation, the Generalised Pareto interpolation yields a more realistic income distribution with a smooth, non-interrupted shape. The method provides the most appropriate picture of high incomes obtainable using only MC data. Nonetheless, the interpolated MC income data underestimates the upper tail of the income distribution compared to the tax data.

For single taxpayers, the individual interpolated net income is equal to the tax unit's net income. Given that the main interest of our paper is of methodological nature, we restrict our analysis to the subgroup of individual taxpayers. This reduces the complexity of tax unit composition and income aggregation due to joint taxation rules in Germany.

3 Methodological Framework

In this Section, we introduce our strategy in finding appropriate data fusion methods to match both studies, the Tax Statistics and the Microcensus. For this, we first introduce common data fusion scenarios and then highlight the opportunities and challenges of the specific data fusion situation of TS and MC. Subsequently, we describe the particular data fusion methods selected in more detail.

3.1 Introduction to Data Fusion Scenarios

Generally, common data fusion scenarios are defined as a specific missing-data pattern that occurs when we 'stack' two originally independent data sources A and B (see e.g. Rubin 1986; Rässler 2002, ch. 4; Meinfelder 2013; Meinfelder and Schaller 2020). Figure 1 displays this specific missing-by-design pattern where the blank parts are missing as the respective variables have not been part of the original study. Therefore, as already indicated before, we denote variables which are observed in both data files as common variables X in the following, while we denote the specific variables relevant for the analysis which are only observed in A (but not in B) as Y and, analogously, specific variables required for the analysis which are only observed in B (but not in A) as Z.

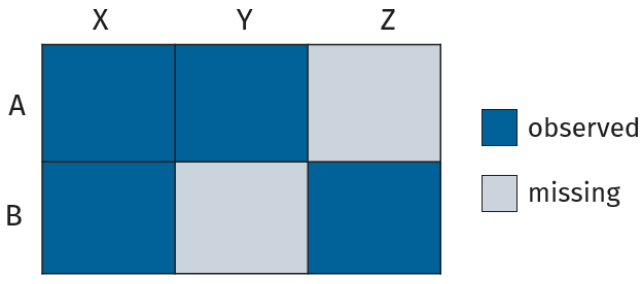

Figure 1: Common Data Fusion Scenario

A typical analysis objective is based on the specific variables Y and Z originally not jointly observed. Therefore, we need identifying assumptions regarding the joint distribu-

tion of Y and Z. In conventional data fusion methods, an implicit *Conditional Independence Assumption* (CIA) is made, which was first pointed out by Sims (1972) in a comment on Okner (1972). The CIA states that any association between Y and Z is a function of X, i.e. $f(Y|X,Z) = f(Y|X)$ and $f(Z|X,Y) = f(Z|X)$. This induces a correlation of zero between Y and Z if conditioned on X and we therefore assume that Y and Z are independent given X. Rodgers (1984) has discussed the shortcomings of the CIA in detail within a comprehensive simulation study. However, in recent years several publications have addressed the CIA issue by proposing to introduce auxiliary information (see e.g. Singh *et al.* 1993; Zhang 2015; Fosdick *et al.* 2016) or, if possible, by incorporating variables that are strongly related to Y or Z within the data fusion process (Donatiello *et al.* 2016). As our aim is to incorporate the rough income information from the Microcensus as a common matching variable through an appropriate method, our strategy is closely related to Donatiello *et al.* (2016).

Although the missing-data pattern displayed in Figure 1 suggests that both missing parts could be imputed within the stacked data set, in practical applications typically only one of both data sets is used for analysis. This study, say A in our case, is labeled as the *recipient* study, while file B represents the *donor* study that 'donates' its Z values to data file A (see e.g. van der Putten *et al.* 2002). In the following, we will introduce the specific data fusion scenario of matching TS and MC and elaborate the differences and challenges compared to traditional data fusion situations.

3.2 Specific Data Fusion Scenario of TS and MC

Since our objective is to enrich the TS with socio-demographic variables obtained by the Microcensus, it is apparent that TS represents the recipient study, whereas the MC serves as donor study. Figure 2 displays the specific data fusion scenario of TS and MC, where data file A equals TS with the common variables X and the observed income variables Y (without information on the socio-demographic variables), and data file B represents MC which consists of the common variables X and the observed socio-demographic variables Z (education and working time in our case).

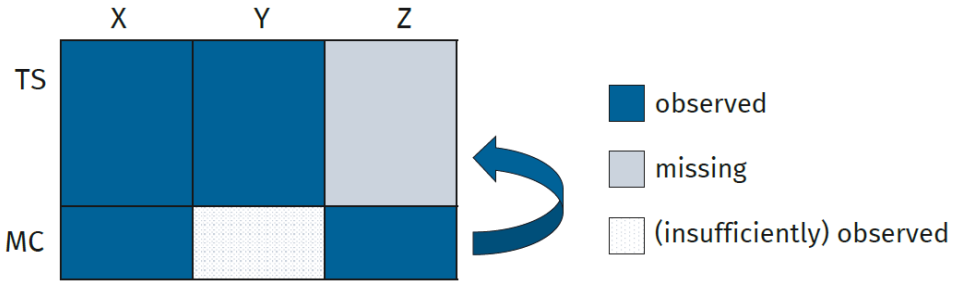

Figure 2: Specific Data Fusion Scenario for TS and MC

In common data fusion scenarios, usually the larger data file serves as donor study in order to ensure a sufficiently large donor pool that avoids donor scarcity for some cell

combinations in X (Andridge and Little 2010). However, as the TS is a full register sample consisting of all tax units in Germany and the MC is a 1% sample of the population, we face a recipient-donor-ratio of $\frac{n_{rec}}{n_{don}} \approx 0.01$ where the recipient study is considerably larger than the donor study. Therefore, one challenge consists of potential donor scarcity for some, especially rare cell combinations in X. In addition, while most data fusion research is based on matching two or more surveys (see e.g. Kamakura and Wedel 1997; Serafino and Tonkin 2017), we match a quite large survey sample (MC) to a full register sample (TS). Accordingly, one further challenge of the specific data fusion situation of TS and MC is that the data sets are quite large (see Tab. 1), which limits the available data fusion implementations due to computational restrictions.

Besides, another particularity of the data fusion scenario of TS and MC is that the income variables Y are not only observed in the TS, but also in MC. However, as stated in Section 2.1, their income concepts differ drastically. Yet, one potential advantage of income being observed in both data files is that it weakens the identification problem that requires a CIA because Y and Z are, in our case, jointly observed in both data sets, albeit with varying quality and measurements. We harmonise the different income variables in TS and MC according to the strategy explained in Section 2.2.

3.3 Relevant Data Fusion Algorithms

This section gives a short overview of traditional data fusion methods and illustrates potential matching techniques that are appropriate for the particular data fusion scenario of TS and MC.

In practical applications, many data fusion implementations are based on some form of non-parametric nearest neighbour hot deck matching, where recipient and donor records are 'matched' according to a given distance metric with regard to their common X characteristics, and the donor record 'donates' its observed Z value to the assigned recipient record (see e.g. Rodgers 1984; Koschnick 1995). Besides, also fully parametric methods (see e.g. Gilula *et al.* 2006) like regression based imputations, or semi-parametric approaches (see e.g. Little 1988; D'Orazio *et al.* 2006a, ch. 2.5) have been discussed in many studies, for example with regard to the ICW project, where Eurostat and NSIs pursue the goal to provide a joint data base consisting of 'income', 'consumption' and 'wealth' (ICW) (Leulescu and Agafitei 2013; Webber and Tonkin 2013; Serafino and Tonkin 2017; Meinfelder and Schaller 2020).

However, we deal with quite large data sets as described in Section 2. This restricts the potential and available data fusion algorithms as many implementations of data fusion methods have been primarily programmed for survey data that would require excessive (and absent) computing capacity as the memory size of the data sets increases. The `StatMatch` package (D'Orazio 2020), for example, is a useful package to match survey data by means of the widespread non-parametric nearest-neighbour approaches, but it generates memory-

and computationally-intensive distance matrices that would require excessive computational resources that are not available despite the high computational power at the Federal Statistical Office of Germany.

Therefore, our strategy in finding suited data fusion algorithms with regard to the specific scenario of matching TS and MC is twofold: First, as already indicated in Section 3.1, we embed our particular data fusion scenario into a broader missing-data context, as suggested for example by Rubin (1986) and Rässler (2002). Hence, we can consider any sophisticated missing-data technique in order to impute the missing education and working time values in TS. Therefore, we choose two imputation models, Multinomial Logistic Regression and Predictive Mean Matching (PMM). Furthermore, we consider the data fusion situation of TS and MC as a specific aspect of supervised Statistical Learning, aiming to impute the missing education and working time values in TS based on powerful prediction methods. In addition to the two aforementioned approaches, we use two forms of decision trees, Recursive Binary Partitioning and Random Forest. For data fusion scenarios, both Statistical Learning models are trained within the donor data file and the predictions are made on the recipient data file. Thus, the common variables X serve as predictor variables and the specific variables Z to be matched reflect the target variables. We briefly describe the proposed methods in the following.

- **Multinomial Logistic Regression (Multinom)**: This approach is based on computing Multinomial Regression models of Z (i.e., education and working time) on the common variables X within the donor data file and, subsequentely, predict the missing Z values in the recipient file. Since this method is purely model-based, it is equivalent to a fully parametric approach.

- **Predictive Mean Matching (PMM)**: PMM, however, is a semi-parametric approach that was first proposed by Rubin (1986) and Little (1988). The principle is to compare the predictive mean of the missing values with the predictive mean of the observed values and to choose the most similar unit as donor unit (non-parametric), while the computation of the predictive means is based on a (parametric) regression model of Z on X within the donor data file. However, PMM is actually an imputation method for metric variables only, but as a semi-parametric approach it might benefit from the particular advantages of non-parametric and fully parametric methods. Therefore, for PMM, we consider education and working time as a metric variable, which provides additional methodological insights on how PMM performs with regard to originally categorical Z variables. We conduct both the Multinomial Regression approach and PMM using the `mice` package (van Buuren 2020) in `R` and, thus, the imputations are based on Fully Conditional Specification (FCS) as described in van Buuren and Groothuis-Oudshoorn (2011).

- **Recursive Binary Partitioning (Rpart)**: This approach is based on classification or

regression trees (see Breiman *et al.* 1984), respectively, while we apply classification trees as we want to match categorical variables Z to the recipient data file. The general idea of classification trees is to split the observations into most optimal subgroups of the predictor variables X by a minimum error classification rate, while the predictions result from the mode of the observations within the resulted subgroups (see e.g. James *et al.* 2013, ch. 8). To do so, we use the `rpart` package (Therneau *et al.* 2019a) that builds classification trees based on recursive binary partitioning, where the predictor space X is split into most optimal binary subgroups (see Therneau *et al.* 2019b). Then, the binary partitioning is selectively continued within the resulting subgroups. Finally, the tree gets pruned by cost complexity pruning in order to avoid overfitting (see e.g. James *et al.* 2013, ch. 8.1).

- **Random Forest (RF)**: Random Forest (see Breiman 2001) computes a series of independent decision trees via bootstrapping. For each split in a tree, only a random sample of all predictor variables X is considered. This leads to a variance reduction by decorrelating, i.e. by avoiding a too strong dominance of a highly correlated predictor (see e.g. James *et al.* 2013, ch. 8.2.2). For this, we use the `ranger` package (Wright and Ziegler 2017), which is a fast implementation of the widely-used `randomForest` package (Breiman *et al.* 2018).

Note that it would also be possible to perform the Classification Tree and the Random Forest within the `mice()` function, since both methods are implemented there. However, due to the high-dimensional data sets of TS and MC, `mice` (and even `parlmice()` as a parallelised version of `mice()`) requires an excessive and absent computational capacity if *cart* or *rf* is selected as imputation method. Therefore, we use `rpart` and `ranger` to ensure a feasible data fusion implementation. In the next Section, we describe the research design to compare and evaluate the proposed data fusion methods.

4 Research Design

Our research design for finding a suited method to match TS and MC is twofold: First, we conduct a simulation study based on the MC where we know the true quantities of interest, which allows us to compare the performance of the proposed data fusion methods. According to the specific data fusion scenario of TS and MC described in Section 3.2, we conduct two simulation studies, one where we exclude income as a matching variable and one where income is included in the data fusion process. Second, we evaluate the results from the simulation study on empirical data, i.e. on the matched data file of TS and MC.

4.1 Simulation Design

Our Monte Carlo Simulation study (see Morris *et al.* 2019) is based only on the Microcensus, where all relevant variables (X,Y,Z) are observed, albeit with imprecise income information. Therefore, the Microcensus builds our surrogate population and, thus, represents the proxy for the Tax Statistics (TS) as the recipient study. From that surrogate TS study, we draw $k = 1,000$ small Microcensuses (SMC) according to the specific sampling procedure of the German Microcensus, which equals a one-stage cluster sampling where the clusters consist of combined information on the building sizes and regional location (Statistisches Bundesamt 2015). Within these clusters, 1% of the areas are drawn at random, and all households of the drawn areas are selected to participate in the Microcensus. We mimic this sampling procedure when drawing the $k = 1,000$ small Microcensuses (SMC) out of the complete Microcensus.

After the sampling step, for all $k = 1,000$ simulation draws, we artificially generate the specific data fusion scenario of TS and MC. For this, in each loop we delete the observed education and working time information within the TS substitute, i.e., the complete Microcensus, and impute the missing values by means of the four proposed data fusion methods. As already pointed out, we conduct two simulation studies. One where we exclude income as a matching variable and, thus, assume conditional independence of Y and Z given X. Therefore, for all $k = 1,000$ simulation runs where income is excluded, we also delete the income information from the drawn SMCs. With regard to the second simulation study, we include income as a common matching variable. For this purpose, we interpolate the 24 categorical income classes by means of the Generalised Pareto interpolation as described in Section 2.2.

In order to assess the performance of the proposed data fusion algorithms as realistically as possible, we base our evaluations on regression parameters β of education and working time derived from exemplary regression models on income. Besides education and working time, the regression models include age, sex, employment status, number of kids, family status, single parent and federal state as explanatory variables. Hence, we estimate the exemplary income models on the Microcensus, which is our surrogate TS population, in order to obtain 'true' quantities of interest, i.e. 'true' regression parameters β of education and working time that serve as benchmarks for evaluation purposes. Accordingly, after each simulation run, we compute the income models on the TS substitute where we include the education and working time variables that have been imputed by the four aforementioned data fusion methods. Thus, we can evaluate to what extent the proposed data fusion methods are able to reproduce the benchmark regression parameters resulting from possible income models.

4.2 Empirical Application

In addition to the simulation study based on the Microcensus, we apply all four data fusion methods to the real data fusion scenario. In other words, we add educational attainments and working times by means of all four methods. We evaluate this empirical application by comparing the tax data income medians conditional on the four different education and working time variables to the income medians from the Microcensus.

Note that this empirical evaluation step is typically impossible in common data fusion scenarios, as Y and Z are not jointly observed. In our specific data fusion scenario, however, income is observed in both the tax data and, albeit insufficiently, the Microcensus. Thus, the Generalised Pareto interpolated income within the MC allows us to obtain at least rough insights on how the joint distribution of income and the socio-demographic variables (education and working time) can be preserved within the matched data file using the four data fusion methods.

However, the level of the income medians differs between the tax and the survey data due to the aforementioned reasons. In order to design a more realistic basis for the evaluation of the data fusion methods, we intend to make the income distributions from the tax and the Microcensus data comparable. For that purpose, we apply reweighting methods to rebuild the income distribution from the tax data in the Microcensus.

4.3 Reweighting

This subsection describes the reweighting methods applied to account for the differences in the income distributions between the tax and the survey data. Table 3 compares the number of individuals in the 24 income classes (boundaries defined as in the Microcensus) between the tax data and the Microcensus in the status quo. Column 3 of Table 3 provides the benchmarks for the reweighting procedure, i.e. the number of individuals in the 24 income classes calculated from the full population tax data in the year 2014.

From Table 3 we can see that the bottom and the top are under-represented, while the middle of the income distribution between 300 and 2000 Euro is largely over-represented in the Microcensus. Compared to the incomes from the personal tax records, the bottom two classes, i.e. incomes of less than 150 Euro, are extremely under-represented in the Microcensus with a total coverage of 23%. The smallest deviations occur for incomes between 2000 and 3200 Euro.

We employ weight calibration techniques of Deville and Särndal (1992) to adjust the Microcensus weights to corrected income weights that represent the German taxpayer population in the best way using a "minimum-distance" criterion minimizing the sum of differences between original and corrected weights. Linear distance functions are used which are not restricted to a certain range. This well-studied calibration technique yields weights that are positive for all income recipients in the Microcensus. The calibrated weights range from

Table 3: Comparison of income frequencies in TS and MC 2014

Income class	Euro range			No. Obs. TS	No. Obs. MC	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

0.0014 to 1.1672 and do thus not require any further adjustment. Brzezinski *et al.* (2019) follow a similar empirical strategy for the Polish taxpayer population.

5 Results

In order to assess the four proposed data fusion methods and its potential to enhance income modeling, we first present the results from the simulation study that is only based on the MC, which enables us to compare the performance of the data fusion methods with regard to the reproduction of the 'true' regression parameters. Subsequently, we conduct the real data fusion of TS and MC by means of the four data fusion strategies to further evaluate the results of the simulation study based on empirical data.

5.1 Simulation Study

The results of the first MC simulation for individual taxpayers with income excluded as a common X variable and assuming conditional independence of Y and Z given X are illustrated in Figures 3 and 4. The vertical red line reflects the benchmark parameter which is the regression parameter of education and working time, respectively. As stated in Table 1, the Z variables education (low, middle, high) and working time (full, part-time, none-time) to be matched are three-scaled and, thus, we observe two regression parameters both for education and working time where low education and full working time, respectively, represent the reference category. The boxes, however, show the MC distributions of the β parameters resulting from all $k = 1{,}000$ simulation runs for each of the proposed data fusion methods.

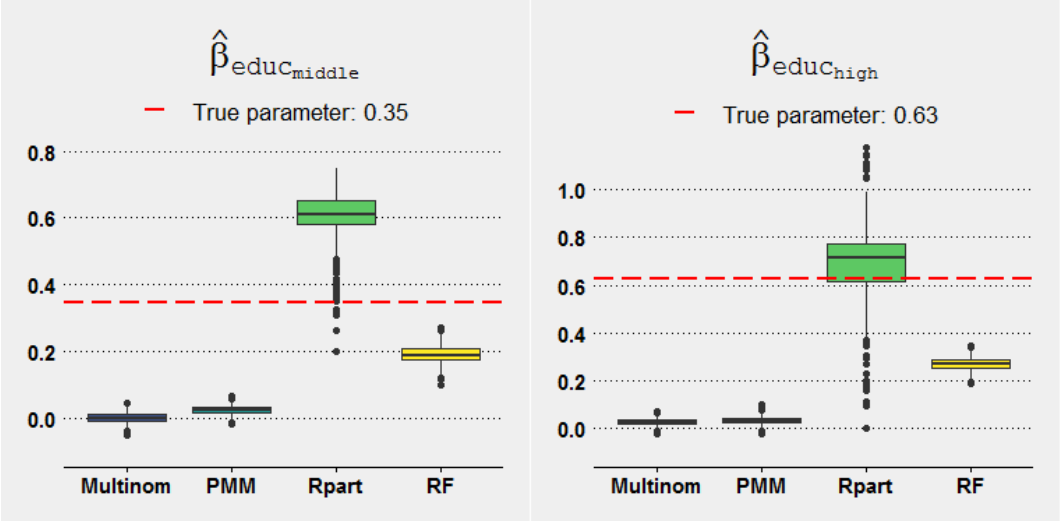

Figure 3: MC distributions of β_{educ} when income is excluded

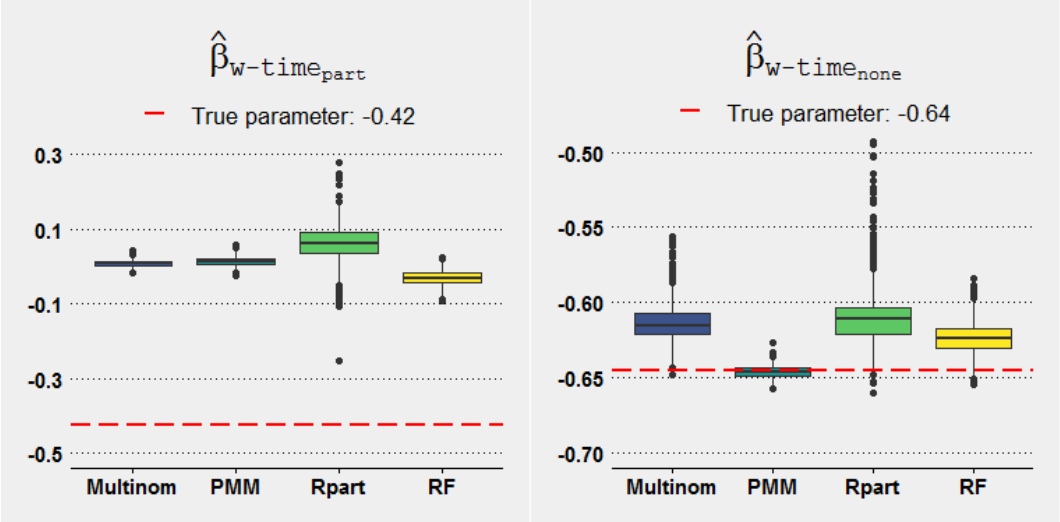

Figure 4: MC distributions of β_{w-time} when income is excluded

With regard to the education coefficients resulting from a data fusion process for single taxpayers where income is excluded, we see that only Rpart is able to reproduce the true quantity of interest of the high education regression parameter $\hat{\beta}_{educ_{high}}$, as illustrated in Figure 3. Concerning the working time parameters, we see in Figure 4 that the data fusion methods are only able to approximately preserve $\hat{\beta}_{w-time_{none}}$, with PMM and RF reproducing the true parameter most accurately here, while all methods perform poorly for $\hat{\beta}_{w-time_{part}}$. This is also evident in Figure 7, where all methods provide high Root Means Squared Error (RMSE) values for the $\hat{\beta}_{w-time_{part}}$ parameter. However, a suited data fusion method should be able to adequately reproduce all quantities of interest of a matched variable, i.e. middle and high education and none- and part-time, respectively, in order to adequately enhance income modeling. One significant problem could be the assumption of conditional independence that is implicitly made as all four data fusion methods produce conditional independence between Y and Z given X in the matched data file. However, it is quite unlikely that this assumption holds, which is why all four data fusion methods might suffer from the CIA and, thus, are not able to reproduce the benchmark coefficients.

To overcome the CIA problem, we incorporate income as a common X variable in the data fusion process, and the results of the simulation study when income is included as a matching variable are illustrated in Figures 5 and 6. Now we see, with regard to the education and working time parameters, that at least two methods, Multinomial Regression and Random Forest, are able to approximately reproduce both parameters of β_{educ} and β_{w-time}, respectively. In addition, it can be seen that RF tends to slightly overestimate the β parameters, whereas the Multinom method underestimates them except of $\beta_{w-time_{none}}$ that is overestimated by Multinom. The RMSEs illustrated in Figure 8 further indicate that Multinom reveals the lowest RMSE for the education parameters, whereas RF yields the lowest overall RMSE for the working time parameters. As the inclusion of the interpolated income variable as a common variable yields better results, we also incorporated income in the real data fusion process of TS and MC for the upcoming empirical application.

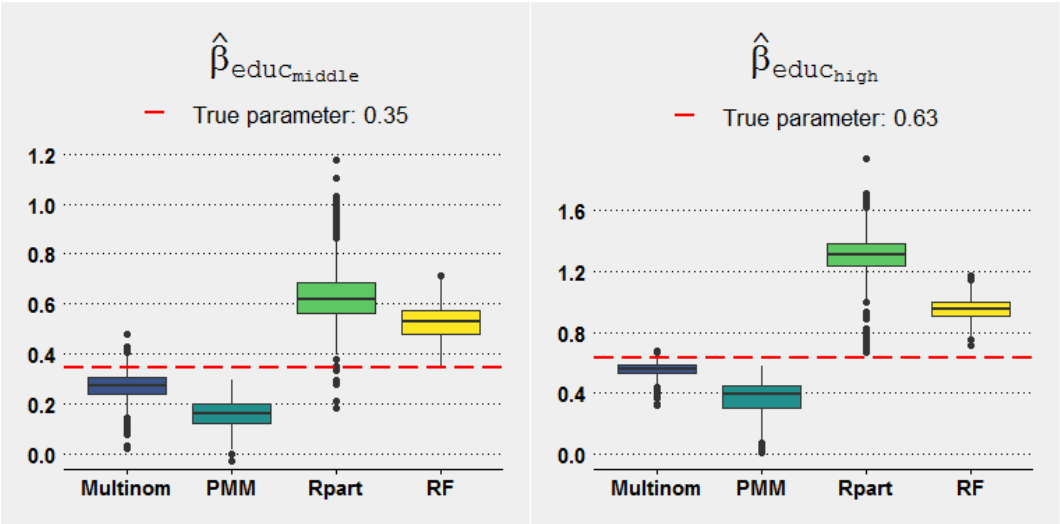

Figure 5: MC distributions for Singles of β_{educ} when income is included

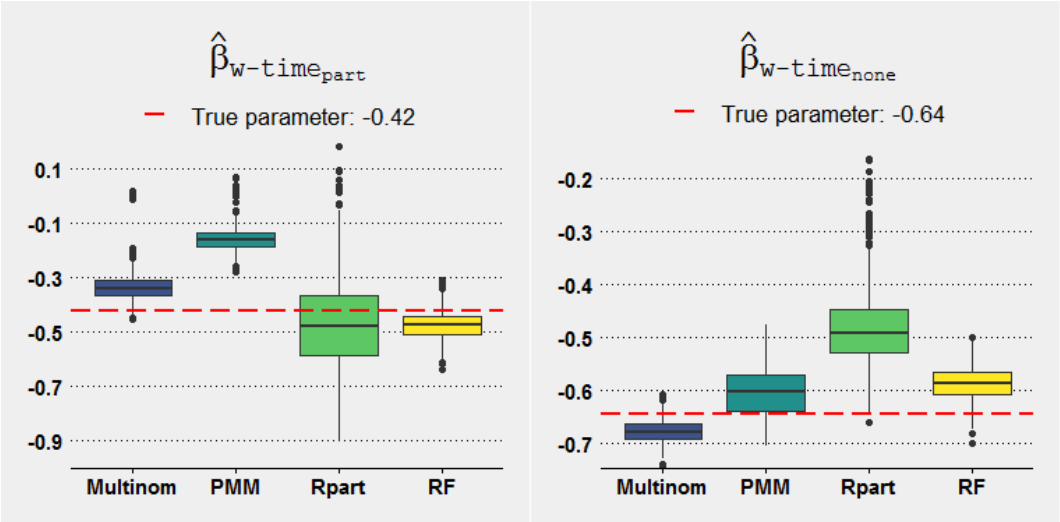

Figure 6: MC distributions for Singles of β_{w-time} when income is included

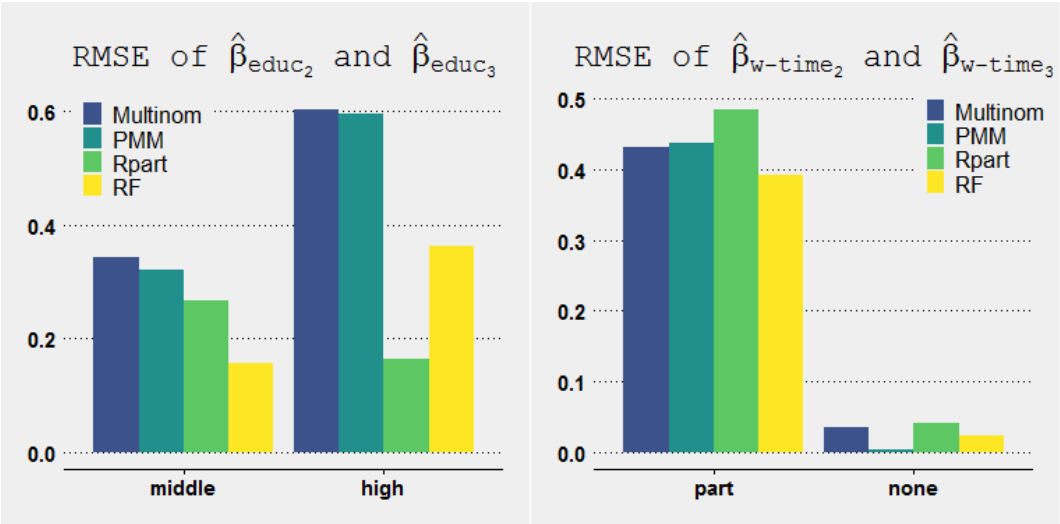

Figure 7: RMSE of β_{educ} and β_{w-time} when income is excluded

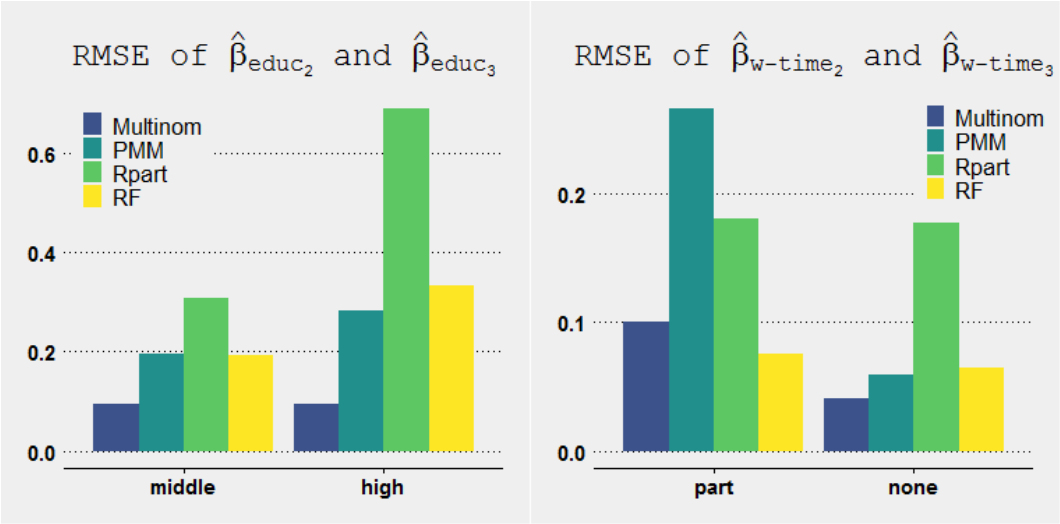

Figure 8: RMSE of β_{educ} and β_{w-time} when income is included

5.2 Empirical Evaluation

One advantage of the tax data is the possibility of regional analysis thanks to the large scale of the data and its seamless regional coverage. We therefore evaluate the results of our simulation study by looking at the regional conditional medians derived from the Microcensus and the tax data.

Figure 9 displays the conditional medians of economic net income on educational level for each federal state. It compares the values from the Microcensus (on the x-axis) to those from the enriched Tax Statistics (on the y-axis) when income was included in the data fusion process. The four panels of the plots correspond to the education variables imputed by the

four different methods. The colours indicate educational level, with darker colors reflecting lower education levels.

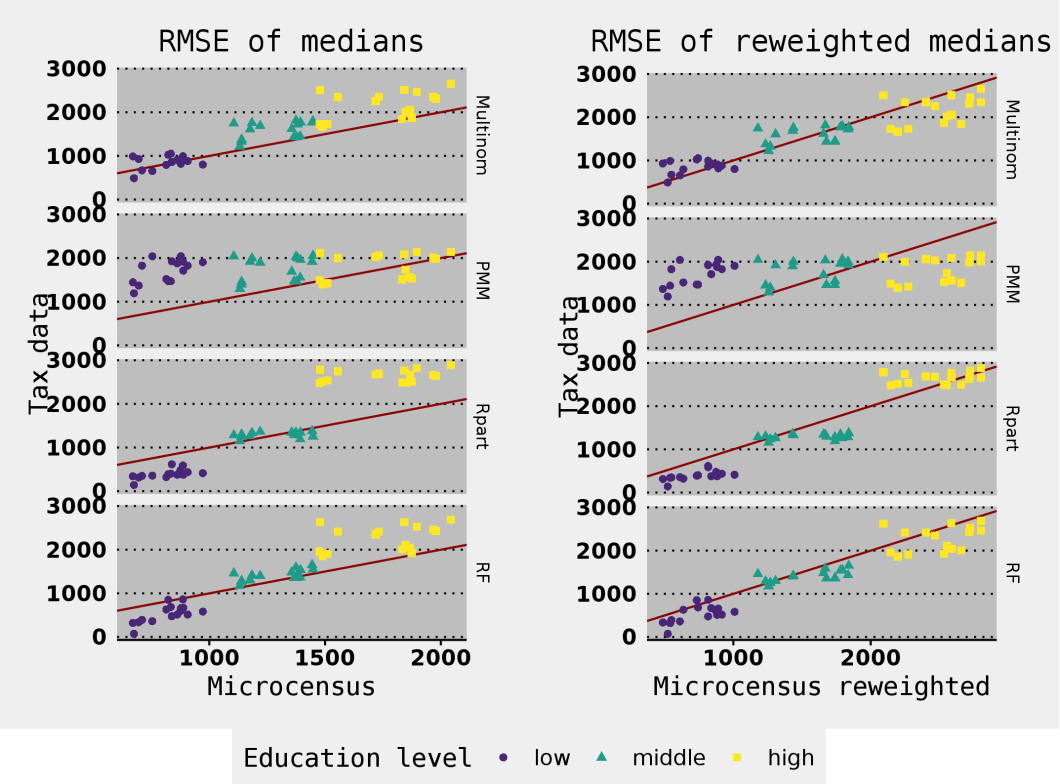

Figure 9: Tax and survey income medians of federal states by education level

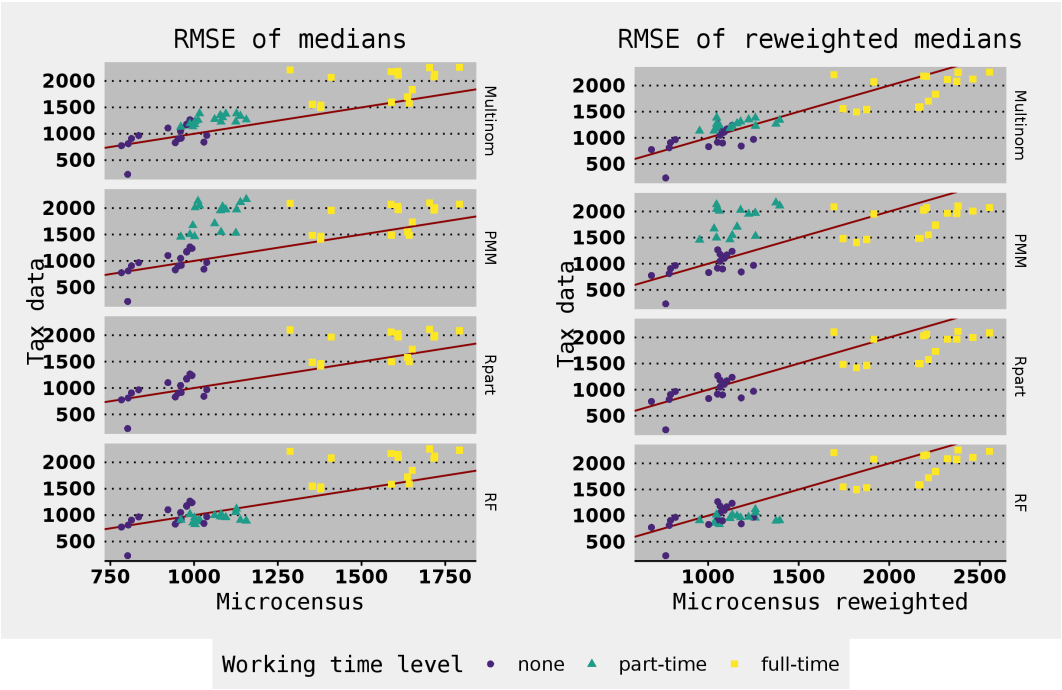

Figure 10: Tax and survey income medians of federal states by working time level

One central difficulty when comparing the conditional medians from the tax to the survey data is the disparity in the income distribution. The derivations from the bisector (red line) underline these differences, showing that incomes of those with low educational attainments are smaller in the Microcensus while the highly educated earn larger incomes in the tax data. The left panel shows the Microcensus results based on the standard weighting factor, whereas the right panel is based on the Microcensus results reweighted as explained in Section 4.3. The conditional reweighted medians are closer to the bisector. Thus, the differences between the medians based on the tax data and the Microcensus are reduced by the reweighting procedure.

Overall, the correlation structure between income and the educational attainments are relatively well maintained based on the three methods Multinom, Rpart, and Random Forest. This can be seen from the fact that the conditional median income rises with higher educational attainments which is in line with theoretical expectations. However, when considering the PMM based imputed education variable, the conditional income medians for all three educational levels are almost identical. Thus, the empirical results indicate that PMM is not well suited. However, it should be noted that PMM is an imputation method exclusively for numeric variables and, thus, we have 'misappropriated' this approach to categorical variables. Therefore, the results do not disqualify PMM as a suitable data fusion method in general, but rather imply that the advantages of PMM as a semi-parametric approach are not necessarily transferable to categorical variables.

Figure 11 sums up the empirical comparison of the four methods in one graph. It shows

the Root Mean Squared Error of the conditional income medians of the federal states in Germany with lower values indicating better performance. The left and right hand side show the results before and after the reweighting procedure, respectively. PMM shows the worst fit, which can potentially be explained by the severe violation of the underlying assumptions since this method is developed for continuous variables. Rpart and Random Forest are the most suitable methods for our data fusion scenario, especially in the reweighted case.

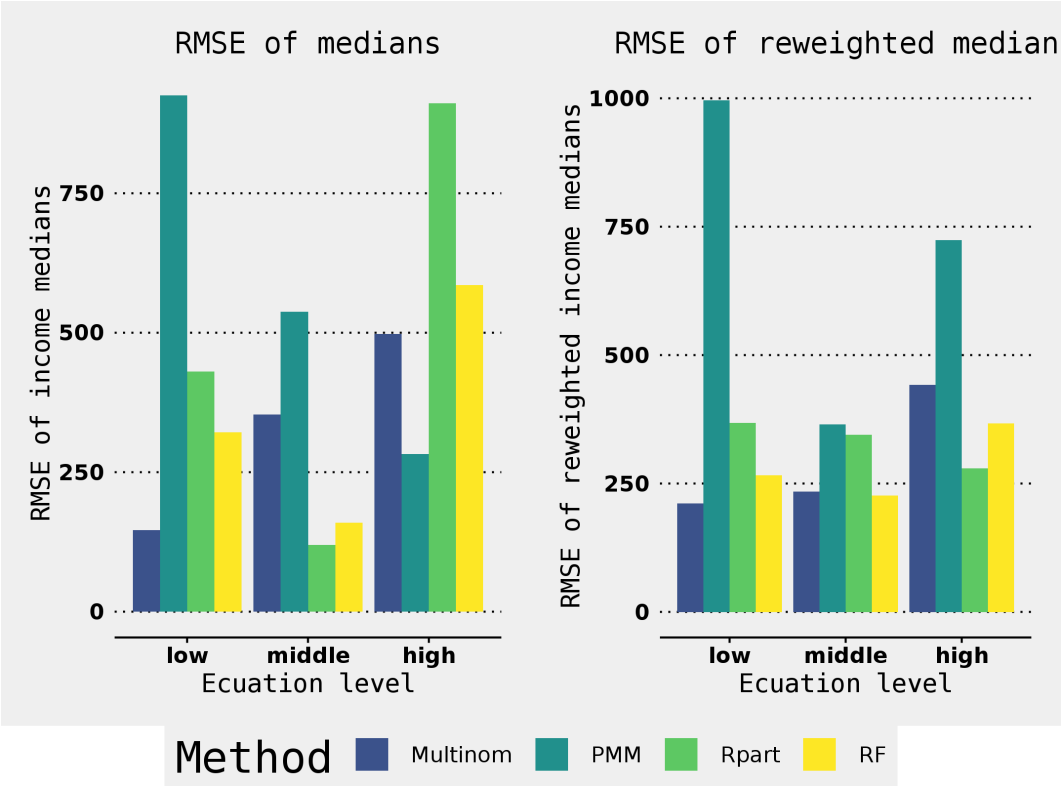

Figure 11: RMSE of income medians of federal states by education level

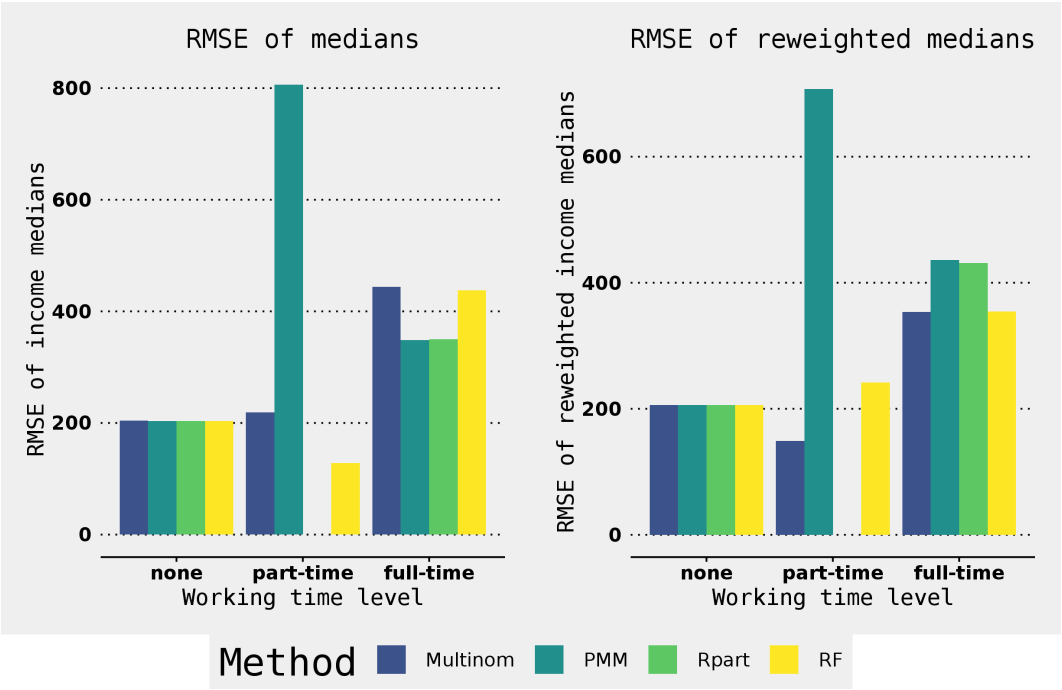

Figure 12: RMSE of income medians of federal states by working time level

With regard to the added working time variables, Figure 10 illustrate that Multinomial Regression and Random Forest can be identified as the most suitable methods in this context. Figure 12 supports this finding. Moreover, Rpart faces a particular challenge when it comes to the relatively sparse group of part-time workers. In deed, in the empirical application, no observations from the TS are chosen to be in the part-time group.

6 Conclusion

The objective of our research was to extend the Tax Statistics, which contains reliable income data on a small-scale regional level, by socio-demographic variables from the Microcensus to provide an integrated data base that support future income modeling. For that purpose, we examined four potential data fusion methods that seem suited for the underlying data fusion scenario. In addition to the atypical donor-recipient ratio and the large data sets, one particularity of the specific data fusion scenario of TS and MC is that the MC contains rough income information.

The results from the simulation study show that no satisfying data fusion outcome can be achieved if the Conditional Independence Assumption (CIA) must be presumed. Therefore, our strategy was to incorporate the rough MC income information within the data fusion process to mitigate the effects of the CIA. For that purpose, we harmonise the income information from the TS and MC by means of the Generalised Pareto interpolation in the MC and the conceptual definition of net income in the TS. As simulation studies can

never claim general validity, we further evaluated the data fusion methods within an empirical application. To concatenate our results from the simulation study and the empirical evaluation, we see that Multinomial Regression and Random Forest perform well in both the simulation study and the empirical evaluation for both regarded variables. However, while Multinomial Regression reveals tendencies of underestimating the regression parameters, Random Forest rather overestimates them. The empirical application indicates underestimations of incomes for the highly educated and larger overall RMSEs for this sub-group when applying Multinomial Regression, whereas Random Forest tends to underestimate the income of the part-time workers.

For the sake of simplicity, we only considered individual tax units. However, we believe we have demonstrated that the data fusion methods outlined, as well as our data fusion strategy, can be used to create an integrated database from the Tax Statistics and the Microcensus that contains reliable income information as well as relevant socio-demographic indicators to support high-quality future income analysis.

Acknowledgements

We appreciate funding of the German Research Foundation Research Group FOR 2559 MikroSim. This work was conducted at the Federal Statistical Office of Germany within the framework of an ongoing research cooperation on microsimulation with Trier university.

Appendix

Income details

Table 4: Construction of economic net income (ENI)

Income component	Tax reduction
+ Income from agriculture and forestry	− allowance for agriculture and forestry
+ Income from business activity	− allowance for business activity
+ Income from self-employment	− allowance for self-employment activity
+ Wage income	− (advertising costs + pension allowance)
+ Capital gains	− (advertising costs + savings allowance)
+ Income from renting and leasing	
+ Other incomes	− advertising costs lump sum
Sum of Income Components (SIC)	
− Exclude all capital gains	
− Exclude income from renting and leasing when smaller than -5000 EUR	
+ Allowances for agriculture, forestry, business, self-employment and pensions	
+ Advertising costs for employees and other incomes	
+ Tax-exempted foreign incomes	
+ Tax-exempted social transfers and child benefits	
Economic Gross Income (EGI)	
− (Income tax + church tax[2] + solidarity surcharge)	
− Social security contributions[3]	
− Paid alimony	
Economic Net Income (ENI)	

Note: *Church tax and social security contributions are partially simulated.*

References

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), 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*, vol. 1, oup Oxford, pp. 18–42.

— and BOURGUIGNON, F. (2014). *Handbook of income distribution*. Elsevier.

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), 303–330.

BARTELS, C. and METZING, M. (2019). An integrated approach for a top-corrected income distribution. *The Journal of Economic Inequality*, **17** (2), 125–143.

BLANCHET, T. (2018). Applying Generalized Pareto Interpolation with gpinter.

—, FOURNIER, J. and PIKETTY, T. (2017). Generalized Pareto Curves: Theory and Applications. *CEPR Discussion Paper*, (DP12404).

BMAS (2017). *Lebenslagen in Deutschland: Der Fünfte Armuts- und Reichtumsbericht der Bundesregierung*, vol. August.

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

—, CUTLER, A. F. O., LIAW, A. and WIENER, M. R. P. (2018). randomforest: Breiman and cutler's random forests for classification and regression: R package.

—, FRIEDMAN, J., STONE, C. J. and OLSHEN, R. A. (1984). *Classification and regression trees*. CRC press.

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 DP*, (12734).

COWELL, F. A. (2000). Measurement of inequality. *Handbook of income distribution*, **1**, 87–166.

DEUTSCHER BUNDESTAG (2017). Sachstand Einkommensungleichheit und Armutsrisikoquote: WD 6 - 3000 - 071/17.

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.

D'ORAZIO, M. (2020). Statmatch: Statistical matching or data fusion: R-package.

—, DI ZIO, M. and SCANU, M. (2006a). *Statistical Matching: Theory and Practice*. Wiley series in survey methodology, Chichester: John Wiley.

EMMENEGGER, J. and MÜNNICH, R. T. (forthcoming). *Localisation of the upper tail: Correcting regional top income distributions*. To be published at: `https://www.uni-trier.de/universitaet/fachbereiche-faecher/fachbereich-iv/faecher/volkswirtschaftslehre/professuren/empirische-wirtschaftsforschung/research-papers`, Universität Trier.

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.

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.

HAUGHTON, J. and KHANDKER, S. R. (2009). *Handbook on poverty+ inequality*. World Bank Publications.

HOCHGÜRTEL, T. (2019). Einkommensanalysen mit dem Mikrozensus. *WISTA*, (3).

JAMES, G., WITTEN, D. and HASTIE, T. (2013). *An Introduction to Statistical Learning: With Applications in R*. Springer Texts in Statistics.

KAMAKURA, W. A. and WEDEL, M. (1997). Statistical data fusion for cross-tabulation. *Journal of Marketing Research*, **34** (4), pp. 485–498.

KOSCHNICK, W. J. (1995). *Standard-Lexikon für Mediaplanung und Mediaforschung in Deutschland: Bd. 1.2*. München: Saur, 2nd edn.

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

LENGERER, A., BOHR, J. and JANSSEN, A. (2005). Haushalte, Familien und Lebensformen im Mikrozensus: Konzepte und Typisierungen. *ZUMA-Arbeitsbericht*, (05).

LEULESCU, A. and AGAFITEI, M. (2013). Statistical matching: A model based approach for data integration. **2013 Edition**.

LITTLE, R. J. A. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, **6** (3), pp. 287–296.

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.

— and SCHALLER, J. (2020). Data fusion for joining income and consumption information using different donor-recipient distance metrics. *arXiv preprint arXiv:2012.00081*.

MINCER, J. (1958). Investment in human capital and personal income distribution. *Journal of political economy*, **66** (4), 281–302.

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.

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), 101–126.

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

RÄSSLER, S. (2002). *Statistical matching: A frequentist theory, practical applications, and alternative Bayesian approaches*, *Lecture notes in statistics*, vol. 168. 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), 357–382.

RODGERS, W. L. (1984). An evaluation of statistical matching. *Journal of Business & Economic Statistics*, **2**, pp. 91–102.

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.

SERAFINO, P. and TONKIN, R. (2017). Statistical Matching of European Union Statistics on Income and Living Conditions (EU-SILC) and the Household Budget Survey. **2017 Edition**.

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

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.

STATISTISCHES BUNDESAMT (2015). *Qualitätsbericht. Mikrozensus 2014*. Statistisches Bundesamt (Destatis).

STATISTISCHES BUNDESAMT (2020a). Bildungsstand der Bevölkerung - Ergebnisse des Mikrozensus 2019.

STATISTISCHES BUNDESAMT (2020b). Qualitätsbericht - Mikrozensus - 2019.

THERNEAU, T., ATKINSON, B., RIPLEY, B. and RIPLEY, B. (2019a). Rpart: Recursive partitioning for classification,regression and survival trees. an implementation of most of thefunctionality of the 1984 book by breiman, friedman, olshen and stone: R package.

THERNEAU, T. M., ATKINSON, E. J. and FOUNDATION, M. (2019b). An introduction to recursive partitioning using the rpart routines.

VAN BUUREN, S. (2020). mice: Multivariate imputation by chained equations: R-package.

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*.

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. **2013 Edition**.

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).

ZHANG, L.-C. (2015). On proxy variables and categorical data fusion. *Journal of Official Statistics*, **31** (4), pp. 783–807.