Estimating the Impact of Alternative Multiple Imputation Methods on Longitudinal Wealth Data

Christian Westermeier[*], Markus M. Grabka, DIW Berlin[Φ]

Abstract

Statistical Analysis in surveys is often facing missing data. As case-wise deletion and single imputation prove to have undesired properties, multiple imputation remains as a measure to handle this problem. In a longitudinal study, where for some missing values past or future data points might be available, the question arises how to successfully transform this advantage into better imputation models. In a simulation study the authors compare six combinations of cross-sectional and longitudinal imputation strategies for German wealth panel data (SOEP wealth module). The authors create simulation data sets by blanking out observed data points: they induce item non response into the data by both missing at random (MAR) and two separate missing not at random (MNAR) mechanisms. We test the performance of multiple imputation using chained equations (MICE), an imputation procedure for panel data known as the row-and-columns method and a regression specification with correction for sample selection including a stochastic error term. The regression and MICE approaches serve as fallback methods when only cross-sectional data is available. Even though the regression approach omits certain stochastic components and estimators based on its result are likely to underestimate the uncertainty of the imputation procedure, it performs weak against the MICE set-up. The row-and-columns method, a univariate method, performs well considering both longitudinal and cross-sectional evaluation criteria. These results show that if the variables which ought to be imputed are assumed to exhibit high state dependency, univariate imputation techniques such as the row-and-columns imputation should not be dismissed beforehand.

[*] Corresponding author, contact at cwestermeier@diw.de, +49 30 89789-223.
[Φ] The authors gratefully acknowledge funding from the *Hans Böckler Foundation*.

1 Introduction

Large-scale surveys are usually facing missing data, which poses problems for researchers and research infrastructure provides alike. In a longitudinal study, where for some observations with missing values past or future valid information might be available, the question arises how to successfully transform this advantage into better imputation models. Single imputation proves to have undesired properties, because, as Rubin (1987, 1996) states, the uncertainty reflected by the respective parameters based on one single stochastic imputation is likely to be biased downwards, since the estimators treat the imputed values as if they were actually observed ones. Otherwise the drawbacks of case-deletion strategies have been well documented (Little & Rubin 1987). Multiple imputation remains as a measure to handle this problem. This study examines the performance of several multiple imputation methods for the adjustment for item-non response (INR) in panel wealth data. Wealth is considered as a sensitive information that is usually surveyed with rather high nonresponse rates (e.g. Riphahn and Serfling 2005, Frick et al. 2010) compared to less sensitive questions such as pure demographic variables like age, sex, migration status. In addition there is a rather high state-dependency in terms of ownership status of wealth components—at least in Germany—, which facilitates the consideration of longitudinal information in the imputation process.

In a simulation study the authors compare six combinations of cross-sectional and longitudinal imputation strategies for German wealth panel data collected for the German Socio-economic Panel Study (SOEP) in 2002, 2007 and 2012. The authors create simulation data sets by setting observed data points to missing based on three separate nonresponse generating mechanisms. We examine the performance of imputation models assuming the mechanisms are missing at random (MAR) or missing not at random (MNAR). We test the performance of multiple imputation by chained equations (MICE, for one of the first popular implementations see Royston 2004). MICE is an iterative and sequential regression approach that grew popular among researchers, because it demands very little technical preparation and is easy to use. We test an univariate imputation procedure for panel data known as the row-and-columns method introduced by Little and Su (1989). Additionally, we test a regression specification with correction for sample selection including a stochastic error term, which was the standard imputation method for the SOEP wealth data in past survey waves.

Prominent (household) panel surveys typically provide their users with imputed information. However, such surveys differ with respect to the imputation strategies applied to handle item-non response and also in the way how available longitudinal information was incorporated. In the following we present those panel surveys which collect wealth information and their imputation strategy as their consideration might give useful clues for the imputation of the SOEP wealth data in this study.

The recently established Eurosystem Household Finance and Consumption Survey (HFCS) is a household wealth survey conducted in 15 euro area countries and organized by the European Central Bank (ECB) (see ECB 2013a). This survey uses an iterative and sequential regression design for the imputation of missing data, similar to the sequential approach we evaluate in this paper (see chapter 4.2). The method used by the HFCS is adopted from similar surveys by the Federal Reserve Board and Banco de Espana (see Kennickel 1991 and 1998, and Barceló 2006). In most of the participating countries the HFCS will be continued as a panel study (ECB 2013b), however, the sequential approach the researchers are using has only been tried and tested in cross-sectional surveys so far. Thus, going forward, we argue that the evaluation of multiple imputation strategies for longitudinal wealth data will increase in relevance.

The Household, Income and Labour Dynamics in Australia Survey (HILDA) is a household-based panel study which collects information about economic and subjective well-being, labour market dynamics and family dynamics in Australia (see Watson and Wooden 2002). HILDA uses a combination of nearest neighbor regression imputation and the row-and-column imputation, depending on the availability of longitudinal information from other waves of the survey (Hayes and Watson 2009).

The US panel study of income dynamics (PSID) is the longest running household panel survey, it started in 1968. The PSID asks about nine broad wealth categories and impute INR using a single hot-deck imputation technique while for home equity a simple carry-forward method is applied (see PSID 2011). [1]

The German Socio-economic Panel Study (SOEP) is a longitudinal representative survey collecting socio-economic information on private households in Germany (Wagner et al. 2007). In contrast to other wealth surveys which survey only one household representative, the SOEP collects wealth information for all household members aged 17 and older in 2002, 2007 and 2012 individually. This survey strategy seems to be advantageous compared to a situation where wealth information is collected by a reference person only, given that accuracy and comparability to official statistics seem to perform better (Uhrig et al. 2012). One major drawback of this strategy is inconsistency on the household level given that wealth components held by several household members can deviate in terms of the stated metric value and result in an even higher share of INR. The major advantage is that,

[1] The rather new UK household longitudinal study „Understanding Society“ has not collected information about wealth yet, this is intended for wave 4 in 2013/14.

when compared to surveys that only ask one reference person, the risk to overlook wealth components of other household members is increased.[2]

The first wave of data was collected prior to the German reunification in 1984 with 12,245 respondents. The original sample was eventually supplemented by 10 additional samples to sustain a satisfactory number of observations and to control for panel effects. In 2002, an oversample of high-income earners was implemented (2,671 individuals), which is particularly relevant for the representation of high net worth individuals in the sample given that income and wealth is rather highly correlated. In 2012, more than 21,000 individuals were interviewed.

As wealth data in the SOEP was collected for the third time in 2012, we decided to compare our traditional approach for the imputation of missing values in wealth module with a few alternative methods (see section 4)—and, if necessary, revise it. The goal of this paper is to evaluate empirically the competing methods for imputing INR in wealth data collected in panel surveys. In many ways this work is a follow-up study to the evaluation study of single imputation methods for income panel data conducted by Watson and Starick (2011). For instance, we largely adopt their set of evaluation criteria with a few modifications, where we found it necessary. They conclude their study with a few remarks: future research should test the performance of imputation methods under different assumptions concerning the nonresponse mechanism, an issue that we are trying to address in this study. Furthermore, they focused on single imputation methods and left it to other researchers to evaluate the performance of multiple imputation methods, again, this something we are tackling with this study.

2 Incidence of INR in SOEP

The SOEP wealth module asks for 10 different asset and liability components: value of owner-occupied and other property (and their respective mortgages), private insurances, building loan contracts, financial assets (such as savings accounts, bonds, shares), business assets, tangibles and consumer credits.

A filter question is asked whether a certain wealth component is held by the respondent, then the market value is collected and finally information about the personal share of property is requested (determining whether the interviewee is the sole owner or, if the asset is shared, the individual share).

For the imputation of the wealth data, there are three steps involved (for more information see Frick et al. 2007, 2010): Firstly, the Filter imputation involves determining whether an individual does have a certain asset type in his or her portfolio. These variables are imputed using rather simple logit

[2] The SOEP wealth data is collected on the person level and not on the household level, however, the tested methods can easily applied to wealth data collected at the household level and we do not expect the results to be significantly different in a household level set-up.

regression models. Secondly, the metric values of the respective assets are imputed. And thirdly, a personal share is imputed again with a rather simple logit regression. In our simulation study we concentrate on INR for the metric values.[3]

Table 1 | Item nonresponse rates in SOEP Wealth Questions

Wave		Type of wealth question	missing filter information	share of missing filter	missing (metric) values*	share of missing values*
2002 (n = 23,892)	**gross wealth**	home market value	83	0.48 %	1,104	4.60 %
		other property	227	0.79 %	453	1.90 %
		financial assets	418	1.89 %	1,822	7.63 %
		building-loan contract	(in 2002 together with private insurances)			
		private insurances	333	1.53 %	3,308	13.85 %
		business assets	243	1.15 %	350	1.46 %
		tangible assets	373	1.70 %	592	2.48 %
	gross debt	debts owner-occupied property	-	-	63	0,26 %
		debts other property	-	-	6	0,00 %
		consumer credits	251	1.19 %	366	1.53 %
2007 (n = 20,886)	**gross wealth**	home market value	139	0.67 %	1,093	5.23 %
		other property	178	0.85 %	364	1.74 %
		financial assets	239	1.14 %	1,931	9.25 %
		building-loan contract	187	0.90 %	921	4.41 %
		private insurances	221	1.06 %	2,781	13.32 %
		business assets	177	0.85 %	290	1.39 %
		tangible assets	199	0.85 %	214	1.02 %
	gross debt	debts owner-occupied property	-	-	179	0.86 %
		debts other property	-	-	40	0.19 %
		consumer credits	180	0.86 %	212	1.02 %
2012 (n = 18,361)	**gross wealth**	home market value	308	1.68 %	958	5.22 %
		other property	350	1.91 %	341	1.81 %
		financial assets	470	2.56 %	1,469	8.00 %
		building-loan contract	349	1.90 %	812	4.42 %
		private insurances	390	2.12 %	2,385	12.99 %
		business assets	344	1.87 %	270	1.47 %
		tangible assets	402	2.19 %	196	1.07 %
	gross debt	debts owner-occupied property	-	-	276	1.50 %
		debts other property	-	-	53	0.29 %
		consumer credits	395	2.15 %	219	1.19 %

Source: SOEP v29; (*) Note that the absolute number of missing metric values, as well as the share, is determined by the sample members who did report that they are holding a certain asset type and could not provide a value, it excludes all members who did not report filter information which has yet to be determined in a separate pre-value imputation. That is why for some variables with a low incidence (such as business assets) the filter information is missing for more individuals than the metric value.

[3] Nevertheless, the applied imputation strategies in our simulation study could also applied to impute the filter and individual share information.

In table 1 we summarize the observed INR incidences for the SOEP wealth data 2002, 2007 and 2012 for the filter and the metric information. It is noteworthy, that the observed prevalence for INR in filter questions is–as expected–lower than for the metric values, given that the filter question is a not a very sensitive information. Here the respective shares on INR lay below three percent of the total population. Moreover, technical and theoretical limitations keep us from providing the data users with multiple imputations for missing information in filter questions, as the sample size for which a certain asset type has been observed (or imputed) would vary depending on the filter information. Thus, the uncertainty in the estimators induced by the imputation procedure is likely to be biased downwards, although presumably not by a large margin. With regard to the metric value, the respective share of INR varies between about zero for debts on other property and about 14 percent for private insurances.

For the above reasons, this study will focus on the performance of multiple imputation methods for item nonresponse in the metric values. Based on our experience, the inclusion of the filter imputation would have little effects on the results.

The Paper is organized as follows: Section 2 describes how we generate a simulation data set with missing values from completely observed cases from the SOEP. Section 3 explains in detail the evaluation set-up and criteria we are choosing to compare the imputation methods. In Section 4 we summarize the imputation methods and discuss their strengths and weaknesses. Section 5 details the performance of these methods using the SOEP wealth data. Section 6 concludes.

3 Simulating Nonresponse

The first step in every imputation procedure that accounts for INR in a given data set is to make an assumption concerning the nonresponse mechanism, which may be either explicitly formulated or implicitly derived from the imputation framework. The commonly used framework for missing data inference traces back to Rubin (1976), who differentiates the response mechanism for three assumptions: Missing Completely At Random (MCAR), Missing At Random (MAR) and Missing Not At Random (MNAR). When the observation is assumed to be MCAR the probability of an observation being missing does not depend on any observed or unobserved variables. With MCAR, excluding all observations with missing values will yield unbiased estimators, but that is, generally, a too strong assumption and of the estimators. Under MAR, given the observed data, the missing values do not depend on unobserved variables. That is, two units with the same observed values will share the same statistical behavior on other variables, whether observed or not. If neither of the two assumptions holds, the data is assumed to be MNAR: the response status is dependent on the outcome of

unobserved variables (e.g. the missing value itself) and cannot be accounted for by conditioning on observed variables.

Since it is necessary to justify a model based imputation procedure, the most commonly used assumption about the nonresponse mechanism is MAR. Besides, as Allison (1987) puts it, "As with other statistical assumptions, [...] the missing at random assumption may be a useful approximation even if it is believed to be false." Since simulating MCAR in a given data set amounts to randomly delete observed values, and therefore all estimators should still be unbiased, we will focus on the evaluation of the imputation methods described in Section 4 only under MAR and MNAR.

We opt to focus on three components of the asset portfolio covered by the SOEP: home value, financial assets and consumer credits. Home value is easily the most important component in the average wealth portfolio in Germany. Financial assets are subject to both a high nonresponse rate compared to other assets and a rather high incidence. Additionally, regression models for the home value tend to yield a good model fit, whereas models for financial assets tend to have relatively poor model fit (Frick et al. 2007). This is equally true for both prediction models of the asset values and modelling the nonresponse mechanism itself. We chose consumer credits as the third component to cover in this simulation study, because it exhibits a rather low incidence and tends to fare mediocre as far as modelling is concerned; the imputation cannot rely on a high number of sound covariates given that the SOEP does not collect additional information about this type of liability, when compared to owner-occupied property or financial assets.

Since there still remains a large pool of fully observed observations after blanking out all INR cases, this turns out to be useful for the creation of simulation data sets. Depending on component and wave there are between 2291 and 8103 values (see the sum of 'Number to be imputed' and 'Nonzero observations' in table 1). Since it is not possible to compare imputed values with the true ones in our imputation set-up, we need to generate a simulation data set. Basically, we estimate a set of logit regression models for the nonresponse mechanism from all cases fully observed in any of the three waves of the SOEP wealth data.

Variables included in the nonresponse model are the employment status und the total personal income, the interview mode, a set of socio-demographic variables (e.g. gender, age, number of children, years of schooling, region) and a rather small set of supplemental economic indicators (e.g. financial support received). Additionally, a set of dummies indicate nonresponse in other wealth components in the same survey wave and a lagged dummy variable indicates nonresponse of the same variable in the last wave (or the next wave in case of 2002) as state dependency is a matter for INR in subsequent waves (Frick and Grabka 2005). Those set of dummies covering the observed response behavior is among the most significant variables, when modelling the observed response behavior of the sample population. Their incorporation requires that we do not blank out observed values in our simulation data sets based

on a static prediction; we rather build a dynamic procedure that updates those predictions based on the response behavior in other waves and for the other two wealth components.

However, since the predicted probability that the value of a certain wealth component is highly dependent on whether the value has been observed in any of the two other waves, the share of observations in our simulation data sets with nonresponse in every wave was too high compared to the original dataset as the information whether INR already occurred in one of the other waves is the most important variable for the non-response process. Therefore we added a small stochastic component to the predictions to incorporate uncertainty. After the addition of this random error terms the share of observations for which information from the other two waves is available for longitudinal imputation is approximately the same as in the original datasets.

Table 2 | Descriptive statistics for observed and simulation data sets, goodness of fit for the logit regression models of the nonresponse mechanism (MAR)

INR assumption	Wave		McFadden R^2	Mean in Euro	Number to be imputed	Nonzero obser-vations	Coefficient of Variation
OBSERVED	**2002**	Home market value	-	243,769	-	7075	0.731
		Financial assets	-	39,798	-	8103	3.209
		Consumer Credits	-	26,544	-	2088	4.792
	2007	Home market value	-	237,508	-	6775	0.762
		Financial assets	-	40,114	-	8377	3.651
		Consumer Credits	-	17,935	-	2978	2.850
	2012	Home market value	-	230,613	-	6164	0.726
		Financial assets	-	44,740	-	7377	2.901
		Consumer Credits	-	16,866	-	2552	4.911
MAR	**2002**	Home market value	0.595	225,724	707	6368	0.773
		Financial assets	0.410	44,921	810	7293	2.026
		Consumer Credits	0.524	26,475	208	1880	1.733
	2007	Home market value	0.518	214,858	677	6098	0.746
		Financial assets	0.391	54,026	837	7540	6.060
		Consumer Credits	0.618	16,191	297	2681	2.048
	2012	Home market value	0.540	202,057	637	5527	0.789
		Financial assets	0.406	59,015	737	6640	3.010
		Consumer Credits	0.597	18,689	255	2297	1.871
MNAR 1	**2002**	Home market value	-	204,609	716	6359	0.634
		Financial assets	-	15,762	808	7295	1.894
		Consumer Credits	-	10,168	176	1912	1.801
	2007	Home market value	-	190,218	692	6083	0.756
		Financial assets	-	11,242	809	7568	2.917
		Consumer Credits	-	6,190	301	2677	2.304
	2012	Home market value	-	195,064	636	5528	0.873
		Financial assets	-	11,287	773	6604	2.306
		Consumer Credits	-	6,682	256	2296	1.871

MNAR 2	**2002**	Home market value	-	283,085	760	6315	0.705
		Financial assets	-	73,853	805	7298	2.253
		Consumer Credits	-	39,505	209	1879	1.748
	2007	Home market value	-	284,654	637	6138	0.800
		Financial assets	-	75,950	858	7519	2.690
		Consumer Credits	-	41,856	309	2669	2.334
	2012	Home market value	-	301,754	626	5538	0.924
		Financial assets	-	84,956	763	6614	2.629
		Consumer Credits	-	36,835	261	2291	6.917

Table 2 displays the McFadden R^2 for the nonresponse models under MAR, the number of observations with missing values and the number of nonzero observations for the simulation assets and waves. Note that the number to be imputed is fixed at around 10 percent of all valid nonzero observations, which is a rather high nonresponse incidence for home market value and consumer credits. The share of missing values for questions concerning the financial assets tends to be higher than 10 percent, however, since our performance criteria solely focus on the differences between imputed and observed values, this handicap has no relevance in practice.

However, as useful and necessary as MAR as an assumption for researchers to handle item nonresponse is, to assume the (non-)response mechanism is fully explained once we conditioned on observed variables (and dismiss any MNAR in the data as negligible) is simplistic. This is why we simulate additional response mechanisms under the MNAR assumption: in two different set-ups we assume that the probability to provide the value of a certain asset depends on the value itself. The empirically observed relationship between nonresponse incidence and the corresponding values tends to be U-shaped, which is better documented for income questions than it is for wealth questions: In fact, Frick and Grabka (2005) state that the incidence for nonresponse of a component of the post-government income for the lowest and highest income deciles is between 28 and 60 percent higher than for the fifth and sixth income deciles. Additionally, characteristics that are typically observed for low income and low wealth households, such as level of schooling and part time employment, have explanatory power in non-response models (Riphahn and Serfling 2005).

Under the assumption that wealth components share a similar nonresponse behavior, we assume in the MNAR1 data sets that the probability to provide a valid answer is the lower the higher true value is. In the NMAR2 data sets, we assume that probability for refusal is inversely proportional to the true value of the wealth components. Table 2 compares the effects on the mean and the coefficient of variation of the respective simulation data sets. Consequently the means for the remaining nonzero observations in the NMAR1 data sets are substantially lower, whereas in the NMAR2 data sets they are substantially higher.

4 Evaluation Criteria

Finding suitable evaluation criteria for a multiple Imputation project is not an easy task. The division into different evaluation categories traces back to Chambers (2001). As part of the EUREDIT project they propose five accuracy measures: ranking accuracy, imputation plausibility, predictive accuracy, distributional accuracy and estimation accuracy. Ranking accuracy refers to the preservation of order in the imputed values, while imputation plausibility refers to the plausibility of the imputed values considering the standard editing procedures.

Measures of ranking accuracy are already sufficiently covered by our criterions (see e.g. criterion 2, 3 and 8) and would not provide additional insight. Measures of plausibility of the imputations are part of the editing process and beyond the scope of this paper. Thus, we closely follow the evaluation framework laid down by Watson and Starick (2011) and focus on a set of 8 different instead of 11 criteria applied by the authors. Those criteria are divided into the three remaining categories: predictive accuracy, distributional accuracy and estimation accuracy. We opt to not include four criteria of Watson and Starick (2011) that we find do not add another dimension to the evaluation at hand and, thus, are redundant. This includes the preservation of skewness and kurtosis, since the preservation of the shape of the distribution is covered by the Kolmogorov-Smirnow distance (6) and the regression analysis for skewed distributions (3) in combination with the first two standardized moments, (4) and (5). Furthermore, unlike Watson and Starick (2011) we do not include Pearson correlations between two wealth types. There is basically no variation for this criterion and for the wealth types we choose for this study. For the same reasons, we choose to not include the Euclidian distance between observed and imputed set of variables. Finally, we add the relative bias to our set of evaluation criteria (see section 4.1).

The main purposes of the SOEP wealth data are divided in two components. Cross-sectional analyses focus on wealth distributions and trend analyses, therefore those should be adequately replicated by the imputation procedure, while longitudinal analyses focus on wealth mobility. Ultimately, the ideal imputation model would account for cross-sectional and longitudinal accuracy. We divided the evaluation criteria into two subsets, to account for the comparatively higher importance of wave-specific trend analyses (six criteria in section 4.1) when compared to analyses that make use of the panel components of the data (two additional longitudinal criteria in section 4.2).

4.1 Wave-specific evaluation criteria

Following Chambers (2000), the first of our eight criteria assesses the relative bias in the difference between the imputed and the observed values of the imputation data sets and separately for all wealth types. With y_i denoting the observed value of a wealth variable for individual i and $\dot{y}_{i,r}$ denoting the imputed value for individual i and in the imputation data set r $(r = 1, 2, \dots, R)$. The **relative Bias (1)** is calculated for each replicate r and averaged over all R replicates.

$$RBias = \frac{1}{R}\sum_{r=1}^{R} RBias_r$$
$$= \frac{1}{R}\sum_{r=1}^{R} \frac{\sum_{i\in S_r}(\dot{y}_{i,r} - y_i)}{\sum_{i\in S_r}\dot{y}_i(1 - R_i)}$$

As we already mentioned before, we set $R = 5$ in this study, because most surveys provide their users with 5 replicates of an imputation data set and the SOEP adopts this strategy.

The next criterion we use is the **Pearson correlation (2)** between observed and imputed variables, defined as

$$r_{\dot{y}y} = \frac{\sum_{i=1}^{n}(\dot{y}_i - \bar{\dot{y}})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(\dot{y}_i - \bar{\dot{y}})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

and is considered as an assessment of the predictive accuracy of an imputed variable. The third criterion, a **regression analysis for skewed distributions (3)**, is as well a measurement of predictive accuracy. We transform both true and imputed values by taking the natural logarithm and calculate a standard linear regression model with the transformed imputed variable as an independent and the transformed observed variable as dependent variable ($Y = \beta\dot{Y} + \varepsilon$). The idea is that the coefficient β should be as close to one as possible. By calculating a standard t-test for $\beta = 1$ we can easily compare the test statistics for each imputation method: the smaller t-test statistic the better the imputation method.

Chambers (2001) notes the imputation results should reproduce the lower order moments of the distribution of the true values. Given that we can directly compare the lower order moments between imputed and observed values, we chose to include the **absolute relative difference in means (4)** and the **absolute difference in the coefficient of variation (5)** for the assessment of estimation accuracy.

$$CR(4) = \left|\frac{(\bar{y} - \bar{\dot{y}})}{\bar{\dot{y}}}\right|$$

$$CR(5) = |\frac{\sigma}{\bar{y}} - \frac{\dot{\sigma}}{\dot{\bar{y}}}|$$

For the reasons we already mentioned above, unlike Watson and Starick (2011) we choose not to include the third and fourth standardized moments.

Distributional accuracy is achieved when the distribution of the true values is preserved by the imputed values. The **Kolmogorov-Smirnov distance (6)** is the higher the more the two tested empirical distributions of the imputed and the true values deviate from each other.

$$d_{KS} = \max_j(|\frac{1}{n}\sum_{i=1}^{n} I(y_i \leq x_j) - \frac{1}{n}\sum_{i=1}^{n} I(\dot{y}_i \leq x_j)|)$$

4.2 Additional longitudinal evaluation criteria

We apply two additional evaluation criteria that help to examine the performance in a longitudinal study. The first criterion assesses the distributional accuracy of wealth mobility between waves for specific components and includes all observations with a positive value for the specific wealth type in two waves simultaneously considered. Wealth mobility is defined by the change in wealth decile group membership in 2002 vs. 2007, 2007 vs. 2012 and 2002 vs. 2012. A standard Chi-square test for fit of the distributions is performed where the imputed cell frequencies are the observed ones and the expected cell frequencies are the true cell frequencies.

$$\chi^2 = \sum_{j=1}^{10}\sum_{i=1}^{10} \frac{(\dot{n}_{ij} - n_{ij})^2}{n_{ij}}$$

Thus, the higher the **Chi-square test statistic (7)** the worse the imputation method can replicate the observed mobility for the wealth component in consideration.

The second longitudinal criterion is the **cross-wave correlation (8)** for each wealth types separately: before and after the imputation procedure the differences of the correlations between each wealth type are compared and should be close to zero. The higher the deviation from zero the worse the performance of the imputation method.

$$r_{y_1y_2} - r_{\dot{y}_1\dot{y}_2} = \left| \frac{\sum_{i=1}^{n}(y_{i1}-\bar{y}_1)(y_{i2}-\bar{y}_2)}{\sqrt{\sum_{i=1}^{n}(y_{i1}-\bar{y}_1)^2 \sum_{i=1}^{n}(y_{i2}-\bar{y}_2)^2}} - \frac{\sum_{i=1}^{n}(\dot{y}_{i1}-\bar{y}_1)(\dot{y}_{i2}-\bar{y}_2)}{\sqrt{\sum_{i=1}^{n}(\dot{y}_{i1}-\bar{y}_1)^2 \sum_{i=1}^{n}(\dot{y}_{i2}-\bar{y}_2)^2}} \right|$$

5 Imputation Methods

The imputation methods which can be considered in our simulation study are limited by the fact that we choose to use multiple imputation techniques. We have to rule out all single imputation techniques beforehand. This includes for example all carryover methods which use valid values observed in the last or next wave of the survey (and variations thereof). This also includes, more generally, all imputation methods without a stochastic component. The methods we choose to examine are commonly used by researchers, practitioners and research infrastructure providers alike, as we already referenced in the first chapter.

We also refrain from considering (longitudinal) hotdeck imputation given that Watson and Starick (2011: 711) already present evidence in a simulation study that their applied hotdeck imputation method does "not perform particularly well on either cross-sectional or longitudinal accuracy".

5.1 Multiple imputation by chained equations (MICE)

We present the basic set-up for imputations using chained equations in this chapter, but for more detailed information we refer to van Buuren et al. (1999), Royston (2004), and van Buuren et al. (2006), among others. Multiple imputation by chained equations (MICE) is not an imputation model by itself, it is rather the expectation that by sequentially imputing the variables using univariate imputation models the analyst expects convergence between the imputation variables after a certain number of iterations. For each prediction equation all but the variable for which missing values ought to be imputed are included, that is, each prediction equation exhibits a fully conditional specification. It is necessary for the chained equations to be set up as an iterative process, because the estimated parameters of the model are possibly dependent of the imputed values. Formally, we have p wealth components $Y_1, Y_2, \dots, Y_p$ and a set of predictors (independent variables, which do not have missing values) Z, then what essentially happens can be described as follows. For iterations $n = 0,1, \dots N$, and with ϕ_j as the corresponding model parameters with a uniform prior probability distribution, the missing values are drawn from

$$Y_1^{(n+1)} \sim g_1(Y_1|Y_2^{(n)}, \dots, Y_p^{(n)}, Z, \phi_1)$$
$$Y_2^{(n+1)} \sim g_2(Y_2|Y_1^{(n+1)}, Y_3^{(n)}, \dots, Y_p^{(n)}, Z, \phi_2)$$
$$\dots$$
$$Y_p^{(n+1)} \sim g_p\left(Y_p\middle|Y_1^{(n+1)}, Y_2^{(n+1)}, \dots, Y_{p-1}^{(n+1)}, Z, \phi_p\right)$$

until convergence at $n = N$ is achieved. That is, in iteration $n + 1$ the dependent variables of each univariate imputation model $g_j(.)$ are updated with the corresponding imputed values determined by the last iteration n (or the ongoing iteration, if the dependent variable already has been imputed). One of the main advantages is that the univariate imputation models $g_j(.)$ may be chosen separately for each imputation variable, which is also why in spite of a theoretical justification for MICE it is widely used by researchers and practitioners. We did not make use of this specific feature at the project at hand, as all wealth variables exhibit similar statistical and distributional characteristics. However, we choose an adjusted set of additional independent variables Z_j for each imputation variable Y_j. In line with the experiences of other countries and surveys for the imputation of wealth data, the additional independent variables Z_j we choose are a set of (1) covariates determining the non-response (variables of the non-response model under the MAR assumption mentioned in section 4.1.), (2) covariates that are considered good predictors for the variable we want to impute (3) economic variables that are possibly related to the outcome variable (according to economic theory) and (4) variables that are good predictors of the covariates included in the rest the groups of variables. However, the last group is especially important in the first iterations and the more dependence between the imputation variables is expected. Nonetheless, we follow those guidelines for the independent variables in the prediction equations and refer to Barceló (2006) for an excellent overview on the reasoning behind the extensiveness of the set covariates and some examples. To give an example why we adjusted the set of independent variables for each imputation variables: e.g. regional information tends to have significant explanatory power for the imputation models of real estate but do not contribute to the estimated models for most of the remaining wealth components.

We specified the imputation models $g_j(.)$ using predictive mean matching (PMM) to account for the restricted range of the imputation variables and to circumvent the assumption that the normality of the underlying models holds true. Predictive mean matching (PMM) was introduced by Little (1988) and is a nearest-neighbor matching technique used in imputation models to replace the outcome of the imputation model for every missing value (a linear prediction) with an observed value. The set of observed values from which the imputed value is randomly drawn consists of (non-missing) values derived from the nearest neighbors which are closest to the linear prediction. Thus, the distribution of the observed values will be preserved for the imputed values.

5.2 Regression with Heckman correction for sample selection

When wealth data was collected for the second time in 2007 in the SOEP, the researchers opted for a regression design with Heckman correction for sample selection for the imputation of the missing (metric) values (cf. Frick et al. 2007, 2010). The first step involved a cross-sectional imputation of

missing values for 2002. These data were then used for a longitudinal imputation of the 2007 data using the lagged wealth data from 2002 as covariates. The third step was a re-imputation of 2002 wealth data using the now-completed longitudinal information from 2007, and starting a cycle of regression models with longitudinal info until convergence between 2002 and 2007 was achieved. The stochastic component in each step, which is necessary to generate multiple implicates, was added through the assignment of randomly drawn residuals derived from the respective regression models. As for this study, we decided to include this already deployed approach in our simulation to compare its performance with other multiple imputation methods.

With the 2012 wealth data and three available waves, the pool of available longitudinal information grew considerably. We decide to add the regression models for 2012 after convergence between 2002 and 2007 has been achieved, with 2007 now serving as the base year. Consequently, longitudinal information from the survey wave 2007 is used for the imputation of missing values in 2002 and 2012 alike.

The variables included in those models are mostly similar to the set of covariates used in the MICE approach (see Section 4.1). However, this regression approach is not sequentially adding updated imputed values from other wealth types; hence the models, predictions and imputed values are mostly calculated isolated, the prediction equation does not include the metric values of the other wealth types. There are a few exceptions: The regression model for home value (other property values) additionally includes the home debt (other property debt). The imputations for both these values are generated in an iterative process in itself, since both values have very high explanatory power in the respective models.

For now, we are including this former SOEP standard in our simulation set-up, even though we are well aware that it is not an adequate multiple imputation procedure in a very narrow sense.

5.3 Row-and-column imputation technique

Little and Su (1989) proposed the row-and-columns-imputation technique as a procedure for item nonresponse adjustment in panel surveys. It takes advantage of available cross-sectional as well as individual longitudinal information. It combines data available from the entire panel duration for every unit (row) and cross-sectional trend information (column) and adds a residual derived from a nearest neighbor matching, thereby attaching a stochastic component to an otherwise deterministic approach.

Since we have three waves of wealth data, the column effects (for any wealth asset) are given by

$$c_t = \frac{(3 * \bar{y}_t)}{\sum_k \bar{y}_k}$$

and is calculated for each wave separately. $\bar{y}_t$ is the sample mean wealth asset for t = 2002, 2007, 2012. The row effects are given by

$$r_i = \frac{1}{m_i} * \sum_j \frac{y_{it}}{c_j}$$

and are calculated for each member of the sample. y_{it} is the value of the wealth asset for individual i in wave t. m_i is the number of recorded waves in which the asset value of individual i has been observed.

Originally, the row-and-column-method was designed as a single imputation method. However, the last step—assigning the residual term from the nearest neighbor—may be modified in such a way that for every individual unit and wave multiple imputed values can be derived. After sorting the units by their row effects r_i, the residual effect of the nearest complete unit l in year j is used to calculate the imputed value for unit i:

$$\dot{y}_{it} = r_i * c_t * \frac{\overbrace{y_{lt}}^{\text{residual term}}}{r_l * c_t}.$$

$\dot{y}_{it}$ is the single imputed value using the residual effect from the nearest neighbor l. To generate multiple imputations we need only two additional steps. Instead of only assigning the residual of the nearest neighbor, we assign the residuals of the k nearest neighbors. Then $r_i * c_t$ is identical for every computation and n residual terms are used to generate k imputed values for every unit i and every year t. Since there is a tradeoff between the number of imputations and the distance to the "farthest" nearest neighbor, we reasoned that the generally agreed on number of five imputations would present a reasonable balance (e.g. the HFCS, other SOEP-variables, the Survey of Consumer Finances (SCF)). However, this decision is merely based on tradition and our expectations and has not been subject to an empirical analysis (yet). Also it is noteworthy, that the residual terms of the five nearest-neighbors have been randomly assigned to imputed values independently for every unit i in order to avoid any systematic differences of imputation accuracy in the five imputation data sets.

5.4 Row-and-columns imputation with age classes

When using the row-and column imputation the donor of the residual term (and the distance between donor and recipient) is solely depending on the sorting of the units by their row effects r_i. At the same

time, as Watson and Starick (2011) state, recipients and the respective donors should have similar characteristics, and those characteristics should be associated with the variable being imputed. They introduce an addition to the basic row-and-column imputation; the method is extended to take into account basic characteristics of the donors and recipients. For a comparison between the standard row-and-column imputation and an imputation with age classes (see figure 1) we match donors and recipients within longitudinal imputation classes defined by the following age classes (at the time, the survey was conducted) in the respective wave: 17–19, 20–24, 25–34, 35–44, 45–54, 55–64, 65 and older. Thereby it is guaranteed that donors will share their residual with recipients from the same age range.

As for the evaluation, we use a two-step approach. Firstly, we determine which row-and-columns imputation is superior. This question amounts to: If a row-and-columns imputation is used for observations that have valid information in other waves, do the addition of age classes improve the performance when compared to the standard row-and-columns imputation. The second and more important step determines which combination of basic and fallback methods yields the best results. Basic imputation method means the technique that is used for observations with missing values and information from other waves of that same individual has been observed. Fallback imputation method means that for some observations with missing values only cross-sectional information and variables are available and, therefore, only the two model based approaches can be applied.

Figure 1 | Basic and fallback imputation methods and evaluation set-up

Step 1	Determining the **superior row-and-column imputation**, given that this approach is used for observations with available longitudinal information			
BASIC (for observations with missing values, information from other waves is available)	Standard Row-and-columns imputation (Little and Su 1989)		Row-and-Columns imputation with age classes (Watson and Starick 2011)	
FALLBACK (for some observations with missing values, only cross-sectional information and variables are available)	Multiple imputation by chained equations	Regression model with Heckmann correction for sample selection	Multiple imputation by chained equations	Regression model with Heckmann correction for sample selection
acronym used in chapter 5	RC-MICE	RC-HECK	RCA-MICE	RCA-HECK

Step 2	Determining the **superior combination of BASIC and FALLBACK** imputation method including the result of step 1 as BASIC			
BASIC (for observations with missing values, information from other waves is available)	Row-and-columns imputation determined in step 1(Little and Su 1989)		Multiple imputation by chained equations	Regression model with Heckmann correction for sample selection
FALLBACK (for some observations with missing values, only cross-sectional information and variables are available)	Multiple imputation by chained equations	Regression model with Heckmann correction for sample selection	Multiple imputation by chained equations	Regression model with Heckmann correction for sample selection
acronym used in chapter 5	RC(A)-MICE	RC(A)-HECK	MICE	HECK

6 Results

As we illustrated in figure 1, we compare the performance of the six combinations of imputation methods in two steps using the eight evaluation criteria we introduced in chapter 3. We first consider the two variations of the row-and-column imputation (with or without additional age classes). Next we examine the best combination of basic and fallback methods to determine which imputation method yields the best results for the SOEP wealth data. The reason we examine the performances in two steps is that we are using ranks based on the evaluation criteria and when we started to sum up the results, it turned out that the row-and-column methods were very close for most parameters. In order to not potentially penalize the remaining imputation methods by our ranking procedure, we determine which row-and-columns method to use first. To give an example, let us assume that for one evaluation criterion the row-and-column method (both with or without age classes) performs best in combination with MICE as the fallback method, that would mean that the remaining combinations would have the ranks 3 through 6 (instead of 2 through 4) and are penalized disproportionately, which in our opinion is not justified based on the similarities between the row-and-columns methods.

As already mentioned, we rank the three chosen wealth items for three years, three assumed nonresponse mechanism and each evaluation criterion separately and compare the outcomes for the imputation methods. In the first step (section 5.1) we rank the imputations using the row-and-column imputation using age classes only against the standard row-and-column imputation, so we only assign the ranks 1 and 2 for the sake of comparison of those two methods. In the second step (section 5.2) we assign the ranks 1 to 4 in order to evaluate the remaining four imputation methods simultaneously. The evaluation criteria (1) – (6) are used for the wave-specific evaluations and the criteria (7) and (8) are additional criteria that solely can be calculated using the results of two waves (2002/07, 2007/12 and 2002/12), for these six longitudinal evaluation criteria we only present the overall rank average (see tables 3 - 8). For the wave-specific evaluation we present wave-specific rank average as well as the overall rank average. Since the outcome of step 2 present the final results, with step 1 only serving to determine, which of the row-and-column imputations we use in step 2, the readers may skip directly to section 5.2.

6.1 Assessment of row-and-column imputation techniques

In tables 3 we show wave-specific and overall rank averages for home market value, which means that the best a method can score is 1 and the worst a method can score is 2 in this first step of our evaluation. We compare RC-HECK versus RCA-HECK and RC-MICE versus RCA-MICE. Consequently the average ranks of these two respective measures add up to 3.

Table 3 | Performance of Home Market Value Imputation, Row-and-Column techniques

Home Market Value					
	Wave-Specific Evaluation				Longitudinal Evaluation
	2002	2007	2012	Overall Rank Average	Overall Rank Average
Assumption: Missing at Random					
RC-HECK	1.5	1.3	1.5	**1.44**	**1.33**
RCA-HECK	1.5	1.7	1.5	1.56	1.67
RC-MICE	1.3	1.5	1.3	**1.39**	1.50
RCA-MICE	1.7	1.5	1.7	1.61	1.50
Assumption: Missing Not at Random 1					
RC-HECK	1.5	1.7	1.3	1.50	1.67
RCA-HECK	1.5	1.3	1.7	1.50	**1.33**
RC-MICE	1.5	1.8	1.3	1.56	1.67
RCA-MICE	1.5	1.2	1.7	**1.44**	**1.33**
Assumption: Missing Not at Random 2					
RC-HECK	1.2	1.2	1.7	**1.33**	1.50
RCA-HECK	1.8	1.8	1.3	1.67	1.50
RC-MICE	1.5	1.5	1.7	1.56	1.50
RCA-MICE	1.5	1.5	1.3	**1.44**	1.50

The differences between the two methods are miniscule for both the overall wave-specific evaluations and the longitudinal evaluation. The detailed comparisons of wave-specific and longitudinal overall ranks for financial assets and consumer credits are buried in the Appendix (see tables A1 and A2), as it was difficult to identify a clear pattern for each of the considered wealth components.

Given that for each of the three wealth components no clear cut conclusion could be drawn, we summed up the results of the separate components into one overall 'sum of rank' (table 4). With regard to the wave-specific evaluation criteria the row-and-column imputation *without* age classes in combination with a simple regression with Heckman selection equation as the fallback method classes performs best when MNAR is assumed. In case of MAR still no distinct deduction can be drawn. When looking at the longitudinal evaluation criteria the pure row-and-column imputation *without* age classes yields the best results in case of MAR and MNAR1. Only in the MNAR2 scenario the combination of a row-&-column imputation *with* age classes and MICE as a fallback method performs best in comparison to the other methods. It appears that there is a small tendency that the row-and-column imputation *without* age classes is superior to the one with age classes. However, the differences are generally very small.

Table 4 | Performance of Home Market Value Imputation, Row-and-Column techniques

	Sum of wave-specific average ranks	**Sum of longitudinal average ranks**
Assumption: Missing at Random		
RC-HECK	4,55	**4,16**
RCA-HECK	**4,45**	4,84
RC-MICE	4,50	4,67
RCA-MICE	4,50	4,33
Assumption: Missing Not at Random 1		
RC-HECK	**4,28**	4,67
RCA-HECK	4,72	4,33
RC-MICE	4,39	**4,17**
RCA-MICE	4,61	4,83
Assumption: Missing Not at Random 2		
RC-HECK	**4,33**	4,83
RCA-HECK	4,67	**4,17**
RC-MICE	4,61	4,67
RCA-MICE	4,39	4,33

Watson and Starick (2011) identify an advantage for the performance of the row-and-column imputation with age classes. However, they use various income items in their research. One possible explanation, why we do not identify a similar advantage, is that the interactions between wealth and age are less strict for wealth items than they are for income items. Hence, the incorporation of age classes into the row-and-column imputation does not improve the performance similarly to income data. Given the unclear results with a small overall advantage for the imputation without age classes for further evaluations in section 5.2 we choose to use the standard row-and-columns imputation.

5.2 Evaluation of Imputation Methods

In this section we compare all imputation methods simultaneously but separately for each of the three wealth types and assumptions of the nonresponse mechanisms. Rank averages mean that the best a method can score is 1 (with each evaluation criterion scoring best) and the worst a method can score is 4 (with each evaluation criterion scoring worst). We compare the pure Regression with Heckman correction for sample selection as both basic and fallback method (HECK), the pure multiple imputations by chained equations as both basic and fallback method (MICE) and both methods when combined with the row-and-column imputation as basic imputation when longitudinal data is available (RC-HECK and RC-MICE).

Table 5 | Overall Performance of Home Market Value Imputation Methods

Home Market Value					
	Wave-Specific Evaluation				Longitudinal Evaluation
	2002	2007	2012	Overall Average Rank	Overall Average Rank
Assumption: Missing at Random					
HECK	3.33	2.67	3.17	3.06	2.67
RC-HECK	2.33	3.50	2.50	2.78	3.00
MICE	3.50	2.83	3.67	3.33	**1.50**
RC-MICE	2.50	3.17	2.33	**2.67**	2.83
Assumption: Missing Not at Random 1					
HECK	2.33	3.33	3.17	2.94	**2.33**
RC-HECK	2.67	1.67	2.67	**2.33**	2.50
MICE	3.67	4.17	2.67	3.50	**2.33**
RC-MICE	3.00	2.33	2.50	2.61	2.83
Assumption: Missing Not at Random 2					
HECK	2.83	2.83	4.00	3.22	2.83
RC-HECK	2.83	3.00	2.50	2.78	3.17
MICE	3.50	3.67	3.17	3.44	**1.33**
RC-MICE	2.50	2.17	2.00	**2.22**	2.67

If we would have considered only the home market value in this evaluation study, we would conclude that RC-MICE is better than HECK and MICE for the imputation of home values for the cross-sectional criteria: Equal weighting of nonresponse mechanisms and all wave-specific and the overall longitudinal performances yields that the overall average is 2.57 for RC-MICE with a standard deviation of 0.33 (HECK mean: 2.96, sd: 0.45; MICE mean: 3.00, sd: 0.86). One possible explanation

is that the home market values in Germany tend to be an asset type with a rather high state-dependency. The row-and-columns approach as univariate imputation technique, which solely considers future and past observed values and an overall trend effect, is closer to the original values than both model-based approaches that may incorporate the uncertainty of the imputation procedure but do underestimate the explanatory power of the lag (or lead) variable. As shown in table 6 this outcome is independent of the nonresponse mechanism that is assumed.

Table 6 | Overall Performance of Financial Assets Imputation Methods

Financial Assets					
	Wave-Specific Evaluation				Longitudinal Evaluation
	2002	2007	2012	Overall Average	Overall Average
Assumption: Missing at Random					
HECK	3.33	3.17	3.00	3.17	2.67
RC-HECK	2.83	2.33	2.83	2.67	2.83
MICE	2.33	2.17	2.00	2.17	**1.33**
RC-MICE	1.50	2.33	2.17	**2.00**	3.00
Assumption: Missing Not at Random 1					
HECK	2.00	3.50	2.17	2.56	2.67
RC-HECK	3.17	3.17	3.50	3.28	3.17
MICE	2.17	1.50	2.00	**1.89**	**1.67**
RC-MICE	2.67	1.83	2.33	2.28	2.50
Assumption: Missing Not at Random 2					
HECK	1.67	3.50	3.83	3.00	2.67
RC-HECK	2.67	2.17	1.50	**2.11**	2.83
MICE	2.67	2.17	1.50	2.44	**1.83**
RC-MICE	3.00	2.17	2.17	2.44	2.67

Considering the overall average ranks for financial assets reveals that is lowest for MICE: 1.94 (sd: 0.38), which is significantly better than both HECK (mean: 2.85, sd: 0.63) and RC-HECK (mean: 2.75, sd: 0.52). Generally, financial assets exhibit less state-dependency than home market values and regression models for both the imputation of the metric values and the nonresponse mechanism are mediocre when compared to other asset types. Thus, there is comparatively more uncertainty to consider by the imputation method and lagged or leaded variables are considerably less important. The higher uncertainty is better captured by the imputed values using the MICE procedure. HECK does not incorporate the uncertainty related to the estimation of the model parameters; hence the imputed values (the predictions plus a randomly drawn residual) may not reflect the uncertainty or be biased. A notable result is that the combination RC-MICE performs at least second best (mean: 2.36, sd: 0.42).

Table 7 | Overall Performance of Consumer Credits Imputation Methods

Consumer Credits					
	Wave-Specific Evaluation				Longitudinal Evaluation
	2002	2007	2012	Overall Average	Overall Average
Assumption: Missing at Random					
HECK	1.83	2.50	2.00	2.11	2.83
RC-HECK	1.83	2.33	1.83	**2.00**	2.50
MICE	3.17	3.00	3.33	3.17	3.00
RC-MICE	3.17	2.17	2.83	2.72	**1.67**
Assumption: Missing Not at Random 1					
HECK	1.83	2.17	2.33	2.11	2.67
RC-HECK	1.83	1.83	2.00	**1.89**	2.67
MICE	3.00	3.50	3.17	3.22	2.67
RC-MICE	3.33	2.50	2.50	2.78	**2.00**
Assumption: Missing Not at Random 2					
HECK	2.17	2.33	2.00	**2.17**	2.00
RC-HECK	2.00	2.50	2.33	2.28	3.33
MICE	2.83	2.83	2.67	2.78	**1.83**
RC-MICE	3.00	2.33	3.33	2.89	2.83

Consumer credits have the lowest state-dependency of the three wealth types we consider in this study. Note that the SOEP wealth data is collected in five-year intervals and credit periods for consumer credits are typically shorter. Following the same argumentation we already laid out for home market values und financial assets, we expect that the row-and-column imputation performs rather weak. As shown in table 8 this is not entirely the case. Both MICE (mean: 2.92, sd: 0.41) and RC-MICE (mean: 2.64, sd: 0.51) perform worse than the imputation with HECK (mean: 2.22, sd: 0.31) and RC-HECK (mean: 2.45, sd: 0.44). One possible explanation is that the imputation models for consumer credits are calculated with considerably less nonzero observations (as shown in table 2) and the standard errors of the respective parameters tend to be much larger than for wealth types with high amount nonzero observations (such as both financial assets and home market values). Thus, we assume that the performance of MICE (and consequently RC-MICE) is diminished by the small sample sizes.

Equal weighting of all nonresponse assumptions, all wealth types and all wave-specific and longitudinal evaluations (as each rank average consists of the mean of six rankings) reveals that the overall average is lowest for RC-MICE (2.52) and highest for HECK (2.68), with RC-HECK (2.56) and MICE (2.62) in between. The overall differences are miniscule and not significantly different in this evaluation set-up. However, the standard deviation of the average ranks may be interpreted as a measure of the performance stability over several waves, wealth types and nonresponse assumptions.

The stability of the performance is again the best for RC-MICE (0.44), it is the worst for MICE (0.76), RC-HECK is second (0.52) and HECK third (0.58). Based on those two overall performance indicators it is surprising, how variable the performance of multiple imputation using chained equations for both basic and fallback observations is, but how stable the combination with the row-and-column imputation proved to be. Despite this result, this evaluation study does not draw a final conclusion, as the outcome would heavily depend on the wealth type that is to be imputed, the survey wave considered, and the nonresponse mechanism assumed.

7 Conclusions

In an assessment of the performance of several imputation methods for longitudinal wealth data we used a set of eight evaluation criteria and three assumptions for the nonresponse generating mechanism. The overall result did not yield that an imputation method performs consistently better for all wealth types. However, in a first step we show that adding age classes to the standard row-and-column imputation as introduced by Little and Su (1989) does not improve the performance based on our criteria and the input data.

In a second step, we compare the standard row-and-column imputation for observations with available longitudinal data with two methods that rely purely on the prediction equations of regression models. We are well aware that the outcome of any evaluation study is a function of the evaluation criteria that are considered. However, in our analyses of the performance of the imputation methods we identified several effects the researcher has to consider for studies using multiple imputation and imputed data. If the data show high state-dependency (such as home market values) the univariate row-and-column imputation performs considerably better than imputation models using regression models, if longitudinal information is available. For wealth types with medium state dependency and high incidence (and consequently a large amount of nonzero observations for the imputation model estimation) such as financial assets the performance of multiple imputation using chained equations seems to yield better results than the regression approach using a Heckman correction for sample selection with a stochastic component. Low incidence and a low amount of nonzero observations for model estimation result in a comparatively poor performance of multiple imputations using chained equations. However, since the large standard errors of the imputed data are justified by the characteristics of the input data, we argue that this is not necessarily a clear disadvantage.

In order to draw a final conclusion for our specific imputation of the SOEP wealth data, we argue that the quantitatively the most important wealth types have a high state-dependency and a high incidence (home market value, other properties, private insurances and building loan contracts). Thus, based on

this study and our experience we would argue in favor of the row-and-column imputation with MICE as fallback imputation, when no longitudinal data has been observed. The overall effect of potentially poorer imputation performance of asset types with low incidence and low state-dependency on the (aggregated) net wealth and wealth inequality estimators should be negligible and reflected by a proper application of confidence intervals and significance tests.

One thing that remains to be addressed is that we refrained from including partial unit nonresponse in this simulation, e.g. individuals within households that choose to not respond, whereas the rest of the household did. The reason is that analyses with the SOEP wealth data focus on the individual level observation and PUNR observations would only affect household wealth estimators. However, we do not expect the results to be significantly different, had we considered PUNR observations. Potential extensions to this study could be the inclusion of additional wealth types, examining the effects of imputation methods on the total net worth and the aggregate net worth and additional imputation methods we did not consider for now.

8 References

Allison, P. D. (1987). Estimation of linear models with incomplete data. In C. Clogg (Ed.), Sociological methodology 1987. American Sociological Association. pp. 71–103.

Barceló, C. (2006), Imputation of the 2002 wave of the Spanish survey of household finances (EFF). Banco de Espana. Documentos Ocasionales. N.º 0603.

Chambers, R. L. (2001), Evaluation Criteria for Statistical Editing and Imputation. National Statistics Methodological Series No. 28, University of Southampton.

European Central Bank (ECB) (2013a), The Eurosystem Household Finance and Consumption Survey. Results from the first wave. Statistics Paper Series No 2.

European Central Bank (ECB) (2013b), The Eurosystem Household Finance and Consumption Survey. Methodological report for the first wave. Statistics Paper Series No 1.

Frick, J. R., Grabka, M.M., and Marcus, J. (2007), Editing and Multiple Imputation of Item-Non-Response in the 2002 Wealth Module of the German Socio-Economic Panel (SOEP). DIW Berlin Data Documentation 18.

Frick, J.R., and Grabka, M.M. (2005), 'Item nonresponse on income questions in panel surveys: incidence, imputation and the impact on inequality and mobility', in Allgemeines Statistisches Archiv, Vol.89, No.1, pp.49-61.

Frick, J.R., Grabka, M.M., and Marcus, J. (2010), Editing und multiple Imputation der Vermögensinformation 2002 und 2007 im SOEP. Data Documentation 51. German Institute for Economic Research, Berlin.

Hayes, C., and Watson, N. (2009), HILDA Imputation Methods. Hilda Project Technical Paper Series No.2/09. Melbourne Insitute of Applied Economic and Social Research.

Kennickel, A. (1991), Imputation of the 1989 Survey of Consumer Finances: Stochastic Relaxation and Multiple Imputation. Federal Reserve Working Paper Series. October 1991.

Kennickel, A. (1998), Multiple Imputation in the Survey of Consumer Finances. Federal Reserve Working Paper Series, September 1998.

Little, R. J. A. (1988), Missing-data adjustments in large surveys. Journal of Business and Economic Statistics 6: 287–296.

Little, R.J.A., and Su, H.L. (1989), 'Item nonresponse in Panel Surveys', in Panel Surveys, ed. D. Kasprzyk, G.J. Duncan, G. Kalton, and M.P. Singh, New York: Wiley.

Panel Study of Income Dynamics (PSID) (2011), Documentation for the 2007 PSID Supplemental Wealth File. Release 2: March, 2011 (http://psidonline.isr.umich.edu/Data/Documentation/wlth2007.pdf) accessed Februray 12th 2014.

Riphahn, R., and Serfling O. (2005), 'Item non-response in income and wealth questions', in Empirical Economics, Vol.30, No.2, pp. 521-538.

Royston, P. (2004.) Multiple Imputation of missing values. Stata Journal, Vol.4, No.3, pp. 227-241.

Rubin, D. B. (1976), Inference and missing data. Biometrika, Vol.63, No.3, pp. 581-592.

Rubin, D. B. (1986), Statistical matching using file concatenation with adjusted weights and multiple imputations. Journal of Business and Economic Statistics 4: 87–94.

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

Schenker, N., and Taylor, J. M. G. (1996), Partially parametric techniques for multiple imputation. Computational Statistics & Data Analysis 22: 425–446.

Uhrig, Noah, Mark Bryan and Sarah Budd (2012), UKHLS Innovation Panel Household Wealth Questions: Preliminary Analysis. Understanding Society Working Paper Series No. 2012 – 01, January 2012.

van Buuren, S., H. C. Boshuizen, and D. L. Knook (1999), Multiple imputation of missing blood pressure covariates in survival analysis. Statistics in Medicine 18: 681–694.

van Buuren, S., J. P. L. Brand, C. G. M. Groothuis-Oudshoorn, and D. B. Rubin. (2006), Fully conditional specification in multivariate imputation. Journal of Statistical Computation and Simulation 76: 1049–1064.

Wagner, G.G., Frick, J.R., and Schupp, J. (2007). The German Socio-Economic Panel Study (SOEP) – Scope, Evolution and Enhancements. Schmollers Jahrbuch. Journal of Applied Social Science Studies, 127, 161-191.

Watson, N. and Starick, C. (2011), Evaluation of Alternative Income Imputation Methods for a Longitudinal Survey. Nicole Watson, Rosslyn Starick. Journal of Official Statistics, Vol27, No.4, pp.693-715

Watson, N. and Mark, W. (2002), The Household, Income and Labour Dynamics in Australia (HILDA) Survey: Wave 1 Survey Methodology. HILDA Project technical papers series No. 1/02, May 2002 (Revised October 2002).

Appendix

A Comparison of Row-and-Column Imputations

Table A1 | Performance of Financial Assets Imputation, Row-and-Column techniques

Financial Assets					
	Wave-Specific Evaluation				Longitudinal Evaluation
	2002	2007	2012	Overall Average Rank	Overall Average Rank
Assumption: Missing at Random					
RC-HECK	1.33	1.83	1.33	1.67	**1.33**
RCA-HECK	1.67	1.17	1.67	**1.33**	1.67
RC-MICE	1.17	1.83	1.50	1.67	1.50
RCA-MICE	1.83	1.17	1.50	**1.33**	1.50
Assumption: Missing Not at Random 1					
RC-HECK	1.33	1.33	1.17	**1.28**	1.67
RCA-HECK	1.67	1.67	1.83	1.72	**1.33**
RC-MICE	1.33	1.33	1.33	**1.33**	**1.33**
RCA-MICE	1.67	1.67	1.67	1.67	1.67
Assumption: Missing Not at Random 2					
RC-HECK	1.50	1.50	1.33	**1.44**	1.50
RCA-HECK	1.50	1.50	1.67	1.56	1.50
RC-MICE	1.50	1.67	1.67	1.61	1.67
RCA-MICE	1.50	1.33	1.33	**1.39**	**1.33**

Table A2 | Performance of Consumer Credits Imputation, Row-and-Column techniques

Consumer Credits					
	Wave-Specific Evaluation				Longitudinal Evaluation
	2002	2007	2012	Overall Average Rank	Overall Average Rank
Assumption: Missing at Random					
RC-HECK	1.33	1.50	1.50	**1.44**	1.50
RCA-HECK	1.67	1.50	1.50	1.56	1.50
RC-MICE	1.17	1.67	1.50	**1.44**	1.67
RCA-MICE	1.83	1.33	1.50	1.56	**1.33**
Assumption: Missing Not at Random 1					
RC-HECK	1.17	1.67	1.67	1.50	**1.33**
RCA-HECK	1.83	1.33	1.33	1.50	1.67
RC-MICE	1.50	1.50	1.50	1.50	**1.17**
RCA-MICE	1.50	1.50	1.50	1.50	1.83
Assumption: Missing Not at Random 2					
RC-HECK	1.50	1.83	1.33	1.56	1.83
RCA-HECK	1.50	1.17	1.67	**1.44**	**1.17**
RC-MICE	1.50	1.50	1.33	**1.44**	1.50
RCA-MICE	1.50	1.50	1.67	1.56	1.50