ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Westermeier, Christian; Grabka, Markus M.

Article — Published Version

Longitudinal Wealth Data and Multiple Imputation: An Evaluation Study

Survey Research Methods

Provided in Cooperation with:
German Institute for Economic Research (DIW Berlin)

Suggested Citation: Westermeier, Christian; Grabka, Markus M. (2016) : Longitudinal Wealth Data and Multiple Imputation: An Evaluation Study, Survey Research Methods, ISSN 1864-3361, Universität Konstanz, Konstanz, Vol. 10, Iss. 3, pp. 237-252, https://doi.org/10.18148/srm/2016.v10i3.6387 , https://ojs.ub.uni-konstanz.de/srm/article/view/6387/6433

This Version is available at:
https://hdl.handle.net/10419/180907

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

Survey Research Methods (2016)
Vol. 10, No. 3, pp. 237-252
doi:10.18148/srm/2016.v10i3.6387

© European Survey Research Association
ISSN 1864-3361
http://www.surveymethods.org

Longitudinal Wealth Data and Multiple Imputation An Evaluation Study

Christian Westermeier
DIW Berlin, Germany

Markus M. Grabka
DIW Berlin, Germany

Statistical analysis in surveys is generally facing missing data. In longitudinal studies for some missing values there might be past or future data points available. The question arises how to successfully transform this advantage into improved imputation strategies. In a simulation study the authors compare six combinations of imputation strategies for German wealth panel data. The authors create simulation data sets by blanking out observed data points: they induce item non response by a missing at random (MAR) and two differential non-response (DNR) mechanisms. We test the performance of multiple imputation using chained equations (MICE), an imputation procedure for panel data known as the row-and-column method and a regression prediction with correction for sample selection. The regression and MICE approaches serve as fallback methods, when only cross-sectional data is available. The row-and-column method performs surprisingly well considering the cross-sectional evaluation criteria. For trend estimates and the measurement of inequality, combining MICE with the row-and-column technique regularly improves the results based on a catalogue of six evaluation criteria including three separate inequality indices. As for wealth mobility, two additional criteria show that a model based approach such as MICE might be the preferable choice. Overall the results show that if the variables, which ought to be imputed, are highly skewed, the row-and-column technique should not be dismissed beforehand.

Keywords: Panel data, SOEP survey, evaluation, simulation, missing at random, item non-response

1 Introduction

Large-scale surveys are usually facing missing data, which poses problems for researchers and research infrastructure providers alike. In longitudinal studies for some missing values there might be past or future data points available. The question arises how to successfully transform this advantage into improved imputation strategies. Single imputation proves to have undesired properties, because 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 (Rubin, 1986, 1987).[1] Multiple imputation addresses this issue. Our study examines the performance of several multiple imputation methods for the adjustment for item-non response (INR) in wealth panel data. Wealth is considered a sensitive information that is usually collected with rather high non-response rates compared to less sensitive questions such as demographic variables like age, sex, migration status (e. g. Frick, Grabka, & Marcus, 2010; Riphahn & Serfling, 2005). In addition, there is a rather high state-dependency in terms of ownership status of wealth components, which facilitates the consideration of longitudinal information in the imputation process.

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). They conclude their study with a few remarks: future research should test the performance of imputation methods under different assumptions concerning the non-response mechanism, an issue that we are trying to address in this study. Furthermore, they focus on single imputation methods and leave it to other researchers to evaluate the performance of multiple imputation methods. Again, this is something we are tackling with this study. In our simulation study we 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. We create simulation data sets by setting observed data points to missing based on three separate non-response generating mechanisms. We examine the performance of imputation models assuming the mechanisms are missing at random (MAR) or the data suffers by differential non-response

Corresponding author: Christian Westermeier, Mohrenstraße 58, 10117 Berlin, Germany (cwestermeier@diw.de)

[1] The drawbacks of case-wise deletion strategies have been well documented.

(DNR). We test the performance of multiple imputation by chained equations (MICE, named after one of the first popular implementations, see Royston, 2004). We test a univariate imputation procedure for panel data known as the row-and-column 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 survey waves 2002 and 2007.

The paper is organized as follows: Section 2 gives an overview of wealth surveys and their imputation strategies and of item non-response in the SOEP wealth data, Section 3 describes how we generate simulation data sets with missing values from observed cases. Section 4 explains the evaluation set-up in detail and the criteria we are choosing to compare the imputation methods. In Section 5 we summarize the imputation methods and discuss their strengths and weaknesses. Section 6 details the performance of these methods using our simulated wealth data derived from the SOEP. Section 7 concludes.

2 Wealth Surveys and Incidence of Item Non-Response in SOEP Wealth Data

Household panel surveys typically provide their users with imputed information. However, such surveys differ with respect to the imputation strategies applied to address item non-response and also in the way how available longitudinal information is incorporated. In the following we present panel surveys, which collect wealth information, and their imputation strategies. Their consideration might give useful clues for the imputation of wealth data in this study.

The recently established Eurosystem Household Finance and Consumption Survey (HFCS) is a household survey conducted in 15 euro area countries and organized by the European Central Bank (ECB) (see European Central Bank (ECB), 2013b). 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 section 4.2). The method used by the HFCS is adopted from similar surveys by the Federal Reserve Board and Banco de España (see Barceló, 2006; Kennickel, 1991, 1998). The number of implicates provided by the HFCS is five, which seems to be the generally agreed on number of imputations provided with survey data.[2] In most of the participating countries the HFCS will be continued as a panel study (European Central Bank (ECB), 2013a). However, the sequential approach the data providers are using has only been tried and tested in cross-sectional surveys thus far. We argue that the evaluation of multiple imputation strategies for longitudinal wealth data will increase in relevance in the future.

The Survey of Health, Aging and Retirement in Europe (SHARE) is a cross-national panel survey including more than 85,000 individuals from 20 European countries[3] aged 50 and older. SHARE also imputes data using a method that is similar to MICE (Christelis, 2011).

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, labor market dynamics and family dynamics in Australia (see Watson & Mark, 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 & 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; INR is imputed using a single hot-deck imputation technique, home equity is imputed using a simple carry-forward method (see Panel Study of Income Dynamics, 2011).

The German Socio-economic Panel Study (SOEP) – the survey used for this study – is a longitudinal representative survey collecting socio-economic information on private households in Germany (Wagner, Frick, & Schupp, 2007). In contrast to other wealth surveys that interview only one household representative, the SOEP collected wealth information separately for all household members (with age 17 or older) in 2002, 2007 and 2012. This survey strategy seems to be advantageous compared to collecting wealth information by one reference person per household only, given that accuracy and comparability to official statistics seem to perform better (Uhrig, Bryan, & Budd, 2012). One major drawback of this strategy is inconsistency on the household level. Given that asset values held by several household members can deviate from each other and may result in an even higher share of INR. The major disadvantage of surveys collecting the data solely interviewing one reference person is that the risk to overlook wealth, assets or debts of other household members increases. However, the methods we test in this evaluation study can be easily applied to wealth data collected at the household level, we do not expect the results to be significantly different in such a set-up.

The first wave of SOEP 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 additional sample 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 are highly correlated. In 2012, more than 21,000 individuals were interviewed.

The SOEP wealth module collects 10 different types of as-

[2] The same number of implicates is also provided by e. g. the SCF, the SOEP, and SHARE.

[3] http://www.share-project.org/home0/overview.html

Table 1
Item non-response rates in SOEP Wealth Questions

Type of wealth question	missing filter information	share of missing filter (%)	missing (metric) values[a]	share of missing values (%)[a]
2002 Wave (n=23892)				
gross wealth				
home market value	83	0.48	1104	4.60
other property	227	0.79	453	1.90
financial assets	418	1.89	1822	7.63
building-loan contract		(in 2002 together with private insurances)		
private insurances	333	1.53	3308	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 Wave (n = 20886)				
gross wealth				
home market value	139	0.67	1093	5.23
other property	178	0.85	364	1.74
financial assets	239	1.14	1931	9.25
building-loan contract	187	0.90	921	4.41
private insurances	221	1.06	2781	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 Wave (n = 18361)				
gross wealth				
home market value	308	1.68	958	5.22
other property	350	1.91	341	1.81
financial assets	470	2.56	1469	8.00
building-loan contract	349	1.90	812	4.42
private insurances	390	2.12	2385	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

[a] 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 or refuse to 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.

sets and debts: 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 asset 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 respondent is the sole owner or, if the asset is shared, the individual share).

The imputation of wealth data consists of three steps (for more information see Frick, Grabka, & Marcus, 2007, 2010): First, the filter imputation determines whether an individual has a certain asset type in his or her portfolio. These variables are imputed using logit regression models. Second, the metric asset values are imputed. And third, a personal share is imputed with logit regressions. In this simulation study we concentrate on item non-response (INR) for the metric asset values.[4]

In Table 1 we summarize the observed INR incidences for the SOEP wealth data 2002, 2007 and 2012 for the metric values and the filter variables. The respective share of INR varies between about zero for debts on other property and about 14 percent for private insurances.

3 Simulating Non-Response

The first step in every imputation procedure that accounts for INR in a data set is to make an assumption concerning the non-response 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). If 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 yields unbiased estimators, but also results in a loss of efficiency. Under MAR, given the observed data, the missing values do not depend on unobserved variables. That is, two units with the same observed values 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 value of unobserved variables (e. g. the missing value itself) and cannot be accounted for by conditioning on observed variables.

The most commonly used assumption about the non-response mechanism is MAR. However, "as with other statistical assumptions, [...] the missing at random assumption may be a useful approximation even if it is believed to be false" (Allison, 1987, p. 77). Thus, we focus on the evaluation of the imputation methods described in section 5 only assuming MAR and two variants of MNAR.

We focus on three components of the asset portfolio covered by the SOEP: home market value, financial assets and consumer credits. Home market value is easily the most important component in the average wealth portfolio in Germany. Financial assets are subject to both comparatively high non-response rates and rather high incidences. Additionally, regression models for the home market value tend to yield a good model fit, whereas models for financial assets tend to have a relatively poor model fit (Frick et al., 2007). We chose consumer credits as the third component to cover in this study, because it exhibits rather low incidences and modelling for both response and asset value tends to fare mediocre; the reason being that 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 in comparison to other assets.

A large pool of fully observed observations remains after blanking out all INR cases, which turns out to be useful for the creation of simulation data sets. Depending on component and wave there are between 2291 and 8103 nonzero asset 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 go one step back and create a simulation data set. Basically, we estimate a set of logit regression models for the non-response mechanism based on the full data set including all observations with empirically missing data.

The variables included in the non-response models 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 non-response in other wealth components in the same survey wave and a lag (or lead) dummy variable indicates non-response of the same variable in one of the other waves as state-dependency matters for INR in subsequent waves (Frick & Grabka, 2005). The simulation data sets, then, are generated by taking all complete cases of one wealth variable and one wave and predicting missingness based on the non-response models and conditional on non-response in other wealth variables and in other waves as already predicted. In order to fully generate the same patterns of missing values, depending on missingness in other variables and waves in the simulation data set, we need to update the prediction in a second sequence.

[4] (Partial) unit non-response and wave non-response – persons or households dropping out of the sample for a limited time or permanently – do not receive any imputation treatment in the person-level SOEP wealth data. Unit non-response generally is addressed by survey weighting procedures (see Kalton, 1986).

Table 2
Descriptive statistics for observed and simulated data (#1)

INR assumption	McFadden R^2	Mean in Euro	Number to be imputed	Nonzero observations	Coefficient of Variation
Observed					
2002 Home market value	-	243769	-	7075	0.731
2002 Financial assets	-	39798	-	8103	3.209
2002 Consumer Credits	-	26544	-	2088	4.792
2007 Home market value	-	237508	-	6775	0.762
2007 Financial assets	-	40114	-	8377	3.651
2007 Consumer Credits	-	17935	-	2978	2.850
2012 Home market value	-	230613	-	6164	0.726
2012 Financial assets	-	44740	-	7377	2.901
2012 Consumer Credits	-	16866	-	2552	4.911
MAR					
2002 Home market value	0.595	225724	707	6368	0.773
2002 Financial assets	0.410	44921	810	7293	2.026
2002 Consumer Credits	0.524	26475	208	1880	1.733
2007 Home market value	0.518	214858	677	6098	0.746
2007 Financial assets	0.391	54026	837	7540	6.060
2007 Consumer Credits	0.618	16191	297	2681	2.048
2012 Home market value	0.540	202057	637	5527	0.789
2012 Financial assets	0.406	59015	737	6640	3.010
2012 Consumer Credits	0.597	18689	255	2297	1.871
DNR I					
2002 Home market value	-	204609	716	6359	0.634
2002 Financial assets	-	15762	808	7295	1.894
2002 Consumer Credits	-	10168	176	1912	1.801
2007 Home market value	-	190218	692	6083	0.756
2007 Financial assets	-	11242	809	7568	2.917
2007 Consumer Credits	-	6190	301	2677	2.304
2012 Home market value	-	195064	636	5528	0.873
2012 Financial assets	-	11287	773	6604	2.306
2012 Consumer Credits	-	6682	256	2296	1.871
DNR II					
2002 Home market value	-	283085	760	6315	0.705
2002 Financial assets	-	73853	805	7298	2.253
2002 Consumer Credits	-	39505	209	1879	1.748
2007 Home market value	-	284654	637	6138	0.800
2007 Financial assets	-	75950	858	7519	2.690
2007 Consumer Credits	-	41856	309	2669	2.334
2012 Home market value	-	301754	626	5538	0.924
2012 Financial assets	-	84956	763	6614	2.629
2012 Consumer Credits	-	36835	261	2291	6.917

Source: SOEP v29
The number of observations to be imputed in the simulated data sets varies slightly around 10 percent of the nonzero observations in the observed data sets, as the exact number of missing values in each data set depends on a stochastic components under both MAR and DNR1 and DNR2. Likewise, the results for MAR, DNR1 and DNR2 are from #1 of randomly generated data sets.

However, since then the predicted probability that the value of a certain wealth component is missing 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 non-response in every wave was too high compared to the original dataset, as the information on the response status in other waves is the most important predictor. 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.[5]

Table 2 displays the McFadden R^2 for the non-response 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 non-response 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, the majority of our performance criteria are not affected by the share, as the focus is on the differences between imputed and observed data sets using only the respective imputed cases.

However, to assume the (non-)response mechanism is fully explained once we conditioned on observed variables may be putting things too simple. Thus, we simulate two additional response mechanisms under the assumption of differential non-response: 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 non-response 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 non-response 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 significant explanatory power in non-response models (Riphahn & Serfling, 2005). As Kennickel and Woodburn (1997) conclude with U.S. wealth data, the higher the household wealth is, the higher the probability that the household refuses to participate.[6]

Under the assumption that wealth components share a similar non-response behavior, we assume in the DNR1 data sets that the probability that a value is missing is the higher, the lower the true value is (i. e. differential non-response at the bottom of the distribution). In the DNR2 data sets, we assume the contrary, the higher the true value of the wealth the higher is the probability that the value is missing. Table 3 compares the effects on the mean and the coefficient of variation of one of the respective generated simulation data sets. Consequently, the means for the observations to be imputed in the DNR1 data sets are substantially lower, whereas in the DNR2 data sets they are substantially higher than in the data sets containing all observed cases.

As all non-response generating mechanisms have a stochastic component, we can easily repeat the steps involved for each assumption to generate 1000 simulation data sets per item non-response assumption. Those 1000 data sets are imputed separately using each of the six imputation methods presented in section 5, yielding in total $3 \cdot 6 \cdot 1000$ imputation procedures.

4 Evaluation Criteria

Our evaluation criteria differ from those of (Watson & Starick, 2011), we focus on a set of 8 instead of 11 criteria applied by the authors. We divide the main applications of wealth data into three sections. (I) Cross-sectional analyses focus on point estimates, trend and distributional analyses. (II) Inequality measurement focuses on the computation of the GINI coefficients and other inequality indices. (III) Longitudinal analyses focus on wealth mobility. (I) and (II) are rather closely related and should be adequately replicated by the imputation procedure. (III) is an additional focus, which we tackle in a separate evaluation. We divide the criteria into two subsets to account for the comparatively higher importance of wave-specific trend and inequality analyses (six criteria in section 4.1) compared to rare analyses that specifically make use of the panel structure of the data (two additional longitudinal criteria in section 4.2). Ultimately, an ideal imputation model would account for cross-sectional, longitudinal *and* inequality accuracy.

Generally, multiple imputation is supposed to yield valid inference as, in comparison to single imputation, the parameters calculated using imputed data do not exhibit biased standard errors. Thus, in the last step of this evaluation we assess the impact of the imputation methods on statistical inference. We compute the *relative bias of standard errors* (1) and compare the results by non-response assumption, method and as-

[5] Sequentially inducing non-response across several waves, assets and NR assumptions is a lengthy and complex exercise; the code for this section as well as sections 4 and 6 is available to researchers, we urge our readers to not hesitate to contact us, if anything is unclear. The code covering the data preparation and imputation is based on the imputations of waves 2002 and 2007 and even lengthier; as it would be a massive undertaking to provide it with decent commentary, it is available from the authors upon request.

[6] Vermeulen (2014) gives a comprehensive overview of the potential effects of differential non-response for high-net-worth-individuals on the measurement of inequality in the European HFCS survey data.

set.

$$\widehat{\mathrm{SE}}\left(\hat{\theta}\right) = \sum_{j=1}^{1000}\left(\frac{\widehat{\mathrm{SE}}\left(\hat{\theta}\right)_j - \mathrm{SE}\left(\hat{\theta}\right)}{\mathrm{SE}\left(\hat{\theta}\right)}\right) \tag{1}$$

$\mathrm{SE}(\hat{\theta})$ is the empirical standard error of the mean calculated using the originally observed data, $\widehat{\mathrm{SE}}(\hat{\theta})$ is the standard error of the mean calculated using the j-th replication of imputed data. Hoogland and Boomsma (1998) suggest that the bias shall not exceed 5%.

4.1 Wave-Specific Evaluation Criteria

Finding suitable evaluation criteria for multiple imputation is challenging. Most criteria applied by Watson and Starick (2011) are not applicable to the task at hand, as they would be heavily biased in favor of a replication of the observed value; for instance, an evaluation of the correlation between observed and imputed value does neglect the fact, that it is not the goal of multiple imputation to create a valid value for an individual missing item, but rather create a valid data set that takes the uncertainty of the imputation procedure into account. Hence, multiple imputation is best understood as simulating values for valid inference. In this study, we chose to evaluate trend, distributional and inequality accuracy jointly in a set of six evaluation criteria that take the overall data set into account instead of the replications of single values.

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 data sets, we chose to include the *absolute relative difference in means* (2) for the assessment of trend accuracy and the *absolute difference in the coefficient of variation* (3) as an indicator of distributional and inequality accuracy. Generally, the dot symbol indicates imputed values, whereas symbols without dots indicate observed values.

$$\mathrm{CR}\,(1) = \left|\frac{(\bar{y} - \bar{\dot{y}})}{\bar{\dot{y}}}\right| \tag{2}$$

$$\mathrm{CR}\,(2) = \left|\frac{\sigma}{\bar{y}} - \frac{\sigma}{\bar{\dot{y}}}\right| \tag{3}$$

Additionally, distributional accuracy is achieved when the distributional properties of the original data set is replicated by the imputed data sets. The *Kolmogorov-Smirnov distance* (4) is the higher the more the two tested empirical distributions of the imputed and the true values deviate from each other. Thus, the smaller the Kolmogorov-Smirnov distance is, the more accurate the imputation method.

$$d_{\mathrm{KS}} = \max_j\left(\left|\frac{1}{n}\sum_{i=1}^{n} I\left(y_i \le x_j\right) - \frac{1}{n}\sum_{i=1}^{n} I\left(\dot{y}_i \le x_j\right)\right|\right) \tag{4}$$

For the assessment of inequality we include three additional criteria. The *Gini coefficient* is especially sensitive against changes in the center of the distribution. The *mean log deviation* is sensitive for shifts at the bottom of the distribution. Those two criteria are complemented by an inequality measure for the top tail of the distribution, by using the *99/50 ratio of percentiles*.[7]

4.2 Additional Longitudinal Evaluation Criteria

We apply two additional evaluation criteria that help to examine the effects of the imputation on wealth mobility. 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. Here, 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.

$$\chi^2 = \sum_{j=1}^{10}\sum_{i=1}^{10}\frac{\left(\dot{n}_{ij} - n_{ij}\right)^2}{n_{ij}} \tag{5}$$

Thus, the higher the *Chi-square test statistic* (5) 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* (6) for each wealth type 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.[8]

$$r_{y_1y_2} - r_{\dot{y}_1\dot{y}_2} = \left|\frac{\sum_{i=1}^{n}(y_{i1} - \bar{y}_1)(y_{i2} - \bar{y}_2)}{\sqrt{\sum_{i=1}^{n}(y_{i1} - \bar{y}_1)^2 \sum_{i=1}^{n}(y_{i2} - \bar{y}_2)^2}} - \frac{\sum_{i=1}^{n}(\dot{y}_{i1} - \bar{\dot{y}}_1)(\dot{y}_{i2} - \bar{\dot{y}}_2)}{\sqrt{\sum_{i=1}^{n}(\dot{y}_{i1} - \bar{\dot{y}}_1)^2 \sum_{i=1}^{n}(\dot{y}_{i2} - \bar{\dot{y}}_2)^2}}\right| \tag{6}$$

[7] This indicator is not responsive to outliers – a relevant phenomenon in wealth analyses – compared to e. g. the half squared coefficient of variation (HSCV).

[8] For comparison's sake we need to mention that we opt to not include four criteria applied by 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 replication of the shape of the distribution is covered by the Kolmogorov-Smirnow distance (4). Furthermore, unlike Watson and Starick (2011) we do not include Pearson correlations between two wealth types. There is not enough covariation for this criterion to be applied for the asset types we choose for this study.

5 Imputation Methods

The imputation methods which can be considered in our simulation study are limited by the fact that we are interested to use multiple imputation techniques. We have to rule out all single imputation techniques beforehand. This includes all carryover methods, which use valid values observed in the last or next wave of the survey (and variations thereof, which have been applied in the PSID for home equity). This also excludes, more generally, all imputation methods without a stochastic component. The methods we choose to examine are commonly used by other important wealth surveys (section 2).

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

5.1 Multiple Imputation by Chained Equations (MICE)

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 present the basic set-up for imputations using chained equations in this section, but for more detailed information we refer to Royston (2004), van Buuren, Boshuizen, and Knook (1999, 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 separate univariate imputation models there will be convergence between the imputed 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 on the imputed values. Formally, we have p wealth components $Y_1, Y_2, \ldots, Y_p$ and a set of predictors (without missing values) Z, then for iterations $n = 0, 1, \ldots, N$, and with ϕ_j as the corresponding model parameters with uniform prior probability distribution, the missing values are drawn from

$$\begin{aligned} Y_1^{(n+1)} &\sim g_1\left(Y_1|Y_2^{(n)}, \ldots, Y_p^{(n)}, Z, \phi_1\right) \\ Y_2^{(n+1)} &\sim g_2\left(Y_2|Y_1^{(n+1)}, Y_3^{(n)}, \ldots, Y_p^{(n)}, Z, \phi_2\right) \\ &\vdots \\ Y_p^{(n+1)} &\sim g_p\left(Y_p|Y_1^{(n+1)}, Y_2^{(n+1)}, \ldots, Y_{p-1}^{(n+1)}, Z, \phi_2\right) \end{aligned} \tag{7}$$

until convergence at $n = N$ is achieved. That is, in iteration $n + 1$ the dependent variables of each imputation model $g_j(.)$ are updated with the corresponding imputed values of the last iteration n (or the ongoing iteration, if the dependent variable already has been imputed). The MICE imputation converges, once the distributions of $Y_1, Y_2, \ldots, Y_p$ all have become stationary conditional on the observed data and the other imputed wealth variables. 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 even though MICE lacks any theoretical justification, 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 barely adjusted the set of additional independent variables Z_j for each imputation variable Y_j. The most important variables among Z_j are the lag and lead variables of the respective assets value, which are drawn from the other waves. Additionally, and in line with the experiences of other countries and surveys for the imputation of wealth data, the independent variables Z_j we choose are in line with the framework laid out in Barceló (2006). We present a detailed overview and further explanations in online appendix C.

We specified the imputation models $g_j(.)$ in (7) 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 one randomly drawn out of the five nearest neighbors which are closest to the linear prediction.

5.2 Regression with Heckman Correction for Sample Selection

For the first two waves of wealth information in the SOEP, the researchers opted for a regression design with Heckman correction for sample selection for the imputation of the missing asset values (Frick et al., 2007, 2010). The first step involved a cross-sectional imputation of missing values for 2002. The 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. In total, Frick et al. (2010) repeat this cycle five times; as this study aims to replicate their approach, we conduct the same number of iterations. The stochastic component in each step, which is necessary to generate multiple implicates, is added through the assignment of randomly drawn

residuals derived from the respective regression models.

With the 2012 wealth data and three available waves, the pool of available longitudinal information grows considerably. We 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 similar to the set of covariates used in the MICE approach (see online appendix C). As in Frick et al. (2007, 2010) we use "life satisfaction" and a dummy for civil servants as selection instruments. However, generally in the Heckman regression the prediction equation does not include the metric values of the other wealth types, as they are not imputed yet.[9] All imputation models are specified separately. Additionally, in comparison to the MICE procedure, this regression model imputation does not include draws of the model parameters – the stochastic component is generated by draws from the residuals – , the uncertainty in the model estimation is not propagated in the imputations.

5.3 Row-and-Column Imputation Technique

Little and Su (1989) proposed the row-and-column imputation technique (RC) as a procedure for item non-response 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

$$c_t = \frac{(3 \cdot \bar{y}_t)}{\Sigma_k \bar{y}_k} \tag{8}$$

and are 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

$$r_i = \frac{1}{m_i} \cdot \Sigma_j \frac{y_{it}}{c_j} \tag{9}$$

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:

$$\dot{y}_{it} = r_i \cdot c_t \cdot \frac{\overbrace{y_{lt}}^{\text{residual term}}}{r_l \cdot c_t}. \tag{10}$$

$\dot{y}_{it}$ 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 in (10), we assign the residuals of the k nearest neighbors. Then terms (8) and (9) are 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 (see e. g. the HFCS, other SOEP-variables, the Survey of Consumer Finances (SCF)). However, this decision is merely based on our expectations and has not been subject to an empirical analysis. 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-Column Imputation with Age Classes

When using the row-and column imputation the donor of the residual term (and the distance between donor and recipient) in (4) is solely depending on the sorting of the units by their row effects r_i. Additionally, the trend component (2) is calculated using the complete sample. 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 (RCA) (see figure 2) 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 share their residual with recipients from the same age range. The column term (2) is calculated using observations from the respective age classes.

[9] 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.

Table 3
Basic and fallback imputation methods, and evaluation set-up

acronym used in section 5	Basic (for observations with missing values, information from other waves is available)	Fallback (for some observations with missing values, only cross-sectional information and variables are available)
MICE-RC	Standard Row-and-column imputation[a]	Multiple imputation by chained equations
REG-RC	Standard Row-and-column imputation[a]	Regression model with Heckmann correction for sample selection
MICE-RCA	Row-and-column imputation[a] using age classes	Multiple imputation by chained equations
REG-RCA	Row-and-column imputation[a] using age classes	Regression model with Heckmann correction for sample selection
MICE	Multiple imputation by chained equations	
REG	Regression model with Heckmann correction for sample selection	

[a] see Little and Su (1989).

One restriction of the Row-and-Column imputation is that it cannot be applied if no longitudinal information on the person level is available, thus we need a fallback method (e. g. the first wave of a respondent, or a specific wealth component is collected for the first time). As for the evaluation, we need a set-up that determines the superior combination of basic and fallback imputation methods simultaneously (see Table 3). The results of the evaluation should provide answers to several questions: (1) If a row-and-column imputation is used for observations that have valid information in other waves, does the addition of age classes improve the performance when compared to the standard row-and-column imputation? (2) 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 values from other waves of that same individual have been observed. Fallback imputation method means that for an observation with missing values only cross-sectional information and variables are available and, therefore, only either of the two model based approaches can be applied. Hence, in addition to the combinations using model based and row-and-column imputations, we test the performance of using a multiple imputation by chained equations as both basic and fallback method (MICE), and we proceed similarly with the regression with Heckman correction (REG).

6 Results

As we illustrated in Table 3, we compare the performance of the six combinations of prevalent imputation methods using the eight evaluation criteria we discussed in section 4. As we wanted to compare the performance of the methods on a metric scale, we refrain from any ranking of the results. Second, we favor the property that the punishment for large deviations is larger than for smaller deviations, which should depend on the overall variance of the outcomes considering the individual evaluation criteria. That means, if the overall variance is small, outliers are punished harder, and deviations that are close to each other are punished similarly. Again, this is a property that is not fulfilled by any ranking of the results. It is, however, fulfilled, if we choose a distance measure that shows the distance between a well-defined optimum and the respective values calculated with imputed data. The optimum is simple to define, as all criteria are either calculated in a way that zero is representing no deviations from the original data or may be transformed to have this respective property. As for the distance measure, using the Euclidian distance would either require a normative decision on a weighting matrix or, alternatively, all criteria would contribute similarly (after normalizing). In order to avoid normative weighting we choose the Mahalanobis distance measure, as it additionally accounts for the observed covariance structure (Mahalanobis, 1936), and thereby is removing any redundancy in our evaluation criteria.

Our evaluation shows the distance between the ideal imputation (all values are zero for all criteria) and the deviation of the imputed values from this ideal point after using the respective imputation method (all tables in section 6). Furthermore, this evaluation set-up allows us to compare the distances directly and interpret them on a metric scale, as the respective outcomes for the different methods are independent from each other (but depending on the overall variation and covariation of the evaluation criteria).

As already mentioned, we show the results for the three wealth items, the three years, and the three assumed nonresponse mechanisms separately and compare the outcomes for the imputation methods. The evaluation criteria (1) – (6) are used for the trend, distributional and inequality evalu-

Table 4
Performance of home market value imputation methods

	Wave-Specific Evaluation			
	2002	2007	2012	Overall Average Distance
Assumption: Missing at Random				
REG	4.93	5.64	5.46	5.34
REG-RC	5.23	5.86	5.82	5.64
REG-RCA	5.32	5.81	5.93	5.69
MICE	6.05	7.02	6.76	6.61
MICE-RC	**4.12**	4.94	**4.73**	**4.60**
MICE-RCA	4.16	**4.91**	**4.73**	**4.60**
Assumption: Differential Non-Response 1				
REG	5.79	6.32	5.77	5.96
REG-RC	6.50	6.25	6.46	6.40
REG-RCA	6.47	6.49	6.65	6.54
MICE	6.98	7.24	6.91	7.04
MICE-RC	5.61	**5.52**	**5.57**	**5.57**
MICE-RCA	**5.53**	5.72	5.71	5.65
Assumption: Differential Non-Response 2				
REG	6.45	5.91	6.06	6.14
REG-RC	6.34	4.94	5.38	5.55
REG-RCA	5.76	4.45	5.08	5.10
MICE	5.96	5.91	5.80	5.89
MICE-RC	5.59	4.42	4.68	4.90
MICE-RCA	**5.02**	**3.96**	**4.42**	**4.47**

Bold figures indicate the smallest average distance among the six imputation variants.

Table 5
Performance of financial assets imputation methods

	Wave-Specific Evaluation			
	2002	2007	2012	Overall Average Distance
Assumption: Missing at Random				
REG	5.82	6.37	5.46	5.88
REG-RC	**5.41**	5.78	5.19	5.46
REG-RCA	5.43	5.86	5.15	5.48
MICE	6.60	5.81	5.18	5.86
MICE-RC	5.49	**4.81**	4.89	**5.06**
MICE-RCA	5.55	4.91	**4.85**	5.10
Assumption: Differential Non-Response 1				
REG	6.28	6.89	6.07	6.41
REG-RC	5.80	6.84	6.09	6.24
REG-RCA	**5.68**	6.73	6.11	6.17
MICE	6.82	6.53	6.03	6.46
MICE-RC	6.17	6.18	**5.69**	6.01
MICE-RCA	6.12	**6.09**	5.70	**5.97**
Assumption: Differential Non-Response 2				
REG	**7.09**	6.51	6.59	6.73
REG-RC	7.38	6.26	**6.31**	**6.65**
REG-RCA	7.49	**6.24**	6.37	6.70
MICE	8.38	7.72	7.54	7.88
MICE-RC	7.22	6.49	6.44	6.72
MICE-RCA	7.35	6.44	6.40	6.73

Bold figures indicate the smallest average distance among the six imputation variants.

ations. The longitudinal criteria (7) and (8) are additional criteria, which can solely be computed using the joint results of two waves (2002/07, 2007/12 and 2002/12) as reported in section 6.2. In section 6.3 we present the results for the relative bias of standard errors.

6.1 Evaluation of Trend, Distributional and Inequality Accuracy

If we would have solely considered the home market value in this study (Table 4), we would conclude that the combination of MICE and the RC imputation yield better results than a pure MICE imputation: Only taking into account the average distances for the trend evaluation reveals that in all cases the MICE imputation performs worse than the combinations with the RC imputation with and without age classes. Looking at the performance for all single waves, in all cases the addition of the RC technique as basic imputation improves the performance of MICE. Combining REG with the RC imputation on the other hand does not regularly improve the results. What is even more surprising, even though the combination of MICE and RC technique seems to perform best overall, the pure MICE approach rarely performs better than the pure REG approach. A possible explanation for these findings is that the home market values tend to be an asset type with a rather high state-dependency. The RC approach as univariate imputation technique, which solely considers future and past observed values and an overall trend effect, is closer to the trend and inequality estimates based on the observed data sets than both model-based approaches that may incorporate the uncertainty of the imputation procedure. Note that these outcomes are basically independent of the non-response mechanism that is assumed.

Generally, financial assets exhibit less state-dependency than home market values and regression models for both the imputation of the metric values and the non-response mechanism are mediocre compared to other asset types (Table 5). Thus, there is comparatively more uncertainty to consider by the imputation method, and the lag or lead variables have, in theory, considerably less explanatory power. However, if the mechanism of missingness is MAR, combining MICE with the RC method, again, yields the best results. If the missing mechanism is differential non-response at the bottom of the

Table 6
Performance of consumer credits imputation methods

	Wave-Specific Evaluation			
	2002	2007	2012	Overall Average Distance
Assumption: Missing at Random				
REG	4.25	4.51	3.83	4.20
REG-RC	4.79	4.62	2.59	4.00
REG-RCA	**4.09**	4.31	**2.25**	**3.55**
MICE	5.35	4.65	4.63	4.88
MICE-RC	4.44	3.70	4.10	4.08
MICE-RCA	4.34	**3.48**	4.24	4.02
Assumption: Differential Non-Response 1				
REG	4.97	4.36	4.48	4.60
REG-RC	5.52	**3.90**	**3.44**	4.29
REG-RCA	4.39	3.95	3.84	**4.06**
MICE	5.30	5.26	4.97	5.18
MICE-RC	4.55	4.50	4.38	4.48
MICE-RCA	**4.22**	4.38	4.51	4.37
Assumption: Differential Non-Response 2				
REG	4.96	**4.56**	5.77	5.10
REG-RC	4.77	5.16	4.51	4.81
REG-RCA	4.85	4.86	**4.39**	4.70
MICE	5.07	4.85	4.63	4.85
MICE-RC	5.09	4.89	4.80	4.93
MICE-RCA	**4.41**	4.71	4.74	**4.62**

Bold figures indicate the smallest average distance among the six imputation variants.

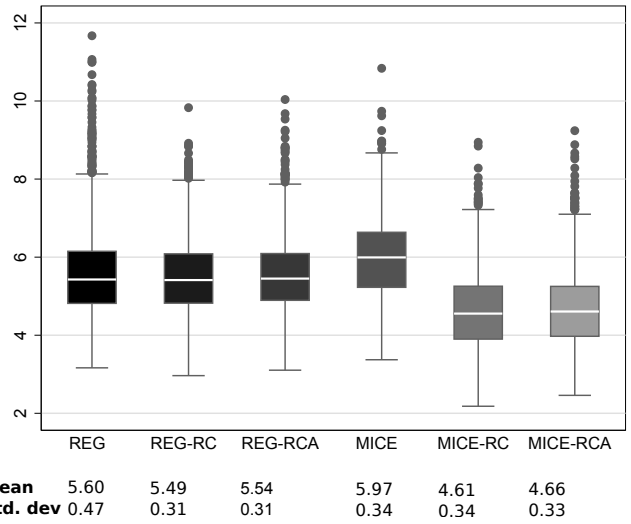

Figure 1. Boxplots for the distances to optimal imputations by imputation methods under Missing at Random (MAR)

distribution, MICE-RCA seems to yield the best results as well. Only if differential non-response at the top is assumed, it is equally viable to choose between any RC method including age classes. Interestingly, including age classes does not improve the results for the RC technique, the differences between RC and RCA seem to be random.

Interestingly, for the evaluation criteria that are considered in this study and for financial assets, it seems to be more viable to choose a pure REG approach over a pure MICE approach if there is differential non-response at the top. Combining REG and RC improves the results under MAR. However, it is notable that all combinations of MICE with the RC method again regularly perform better than both pure model based approaches under any non-response assumption, but considerably less so under DNR2.

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 RC imputation performs rather weak. The results of the evaluation prove us mostly wrong. As shown in Table 6, both RC methods perform oftentimes better if MAR, DNR1 or DNR2 is assumed. Additionally, RCA has an advantage as compared to RC. One possible explanation is that even if the overall state-dependency is much lower for consumer credits, the state-dependency at the bottom of the distribution may still be considerably high and the RC imputation might still yield more accurate imputed data sets in this case. Incorporating age classes seems to improve the results, because consumer credits are more prevalent among younger age classes, who are paying off their debts as they get older.

Comparing the distributions of the distances to the optimal imputations separately for the MAR assumption for all three waves and all assets jointly, confirms the conclusions we draw above (figure 1).

Including the RC imputation does improve the performance of MICE considerably and significantly. The distance between the optimal imputation and MICE versus both MICE-RC and MICE-RCA is roughly 1.3 units higher, the respective means and standard deviations are shown in figure 1 together with the boxplots of the distributions. Considering the performance of REG versus REG-RC and REG-RCA the differences are miniscule. Moreover, results for REG exhibit considerably more variance over the 1000 simulation data sets. Similar figures for DNR1 and DNR2 are presented in the online appendix.

Additionally, we observe that the incorporation of age classes in the RC imputation does not improve the overall imputation results. Watson and Starick (2011) report an advantage for the performance of the RC imputation with age classes for the imputation of income items. One possible explanation, why we do not identify a similar advantage, is

Table 7
Average performance on longitudinal evaluation criteria, all assets

	Home Market Value	Financial Assets	Consumer Credits	Overall Average Distance
Assumption: Missing at Random				
REG	1.51	2.11	**2.47**	2.03
REG-RC	2.53	2.25	2.56	2.45
REG-RCA	2.56	2.10	2.52	2.39
MICE	**0.71**	**0.94**	3.05	**1.57**
MICE-RC	1.79	1.37	3.00	2.05
MICE-RCA	1.77	1.33	2.94	2.01
Assumption: Differential Non-Response 1				
REG	1.77	2.77	**3.08**	2.54
REG-RC	2.77	2.52	3.19	2.83
REG-RCA	2.81	2.50	3.15	2.82
MICE	**1.14**	2.57	3.15	**2.29**
MICE-RC	2.30	**2.36**	3.13	2.60
MICE-RCA	2.35	2.37	3.12	2.61
Assumption: Differential Non-Response 2				
REG	**1.27**	2.35	**3.26**	2.29
REG-RC	2.23	2.31	3.53	2.69
REG-RCA	2.21	2.31	3.48	2.67
MICE	1.46	**0.77**	3.29	**1.84**
MICE-RC	1.72	1.40	3.62	2.25
MICE-RCA	1.63	1.42	3.60	2.22

Bold figures indicate the smallest average distance among the six imputation variants.

that there are less regular trends of increase and spend-down of asset values over the life cycle for home market value und financial assets as compared to income variables.

6.2 Evaluation of Wealth Mobility

As for the two additional longitudinal criteria, which focus on the changes in the observed mobility structures before and after imputations, the overall average distances include all pair-wise comparisons (2002/2007, 2007/2012, and 2002/2012) and are presented in Table 7. We expected that using RC imputations would overestimate the state-dependency for the wealth assets and undermine the actually observed mobility structures. This expectation gets confirmed to a certain extent.

Under MAR the pure MICE approach seems to perform better than the pure REG approach and all combinations with the RC method (at least for home market value and financial assets). This is to be expected, as the mobility seems to be severely reduced, once the only included variable is the lag or lead variable of the respective variable that is to be imputed. What is more surprising is that the REG approach performs considerably worse than MICE. One possible statistical explanation could be that the regression set-up is not taking into account one source of uncertainty, which the MICE procedure does take into account: the drawing of the respective model parameters. The only stochastic component in the REG approach is the drawing of a residual from the observed residuals, whereas MICE imputes values after drawing of the respective model parameters. Here, REG might underestimate the uncertainty of the imputation procedure and produce too less variation in the imputed values, thereby as well reducing mobility.

For differential non-response and especially consumer credits the results are less clear. MICE seems to reproduce the observed mobility structures slightly better than REG, in many cases the combination of MICE and the RC imputations yield satisfying results too, but generally distances to an optimal imputation seem to increase. We conclude that (1) a researcher interested in mobility structures would probably prefer the model based MICE approach to an univariate imputation procedure such as the RC method, and (2) even though REG yields imputed values using model prediction equations as well, the REG imputation performs worse than the MICE approach.

6.3 Evaluation of Standard Errors

Interestingly, inference seems to be affected differently for assets, and less for non-response assumption or imputation

Table 8
Relative bias of standard errors

	Home Market Value	Financial Assets	Consumer Credits	Overall Bias
Assumption: Missing at Random				
REG	−1.80	**5.88**	**−9.94**	−1.95
RC-REG	−0.85	0.14	**−10.98**	−3.90
RCA-REG	−1.00	0.27	**−10.98**	−3.90
MICE	3.04	**5.49**	**−6.10**	0.81
RC-MICE	1.22	1.52	**−7.60**	−1.62
RCA-MICE	1.13	1.67	**−7.73**	−1.64
Assumption: Differential Non-Response 1				
REG	−1.33	**9.00**	**19.01**	**8.89**
RC-REG	−1.17	3.66	**11.19**	4.56
RCA-REG	−1.32	3.77	**11.21**	4.55
MICE	1.72	2.16	4.28	2.72
RC-MICE	−0.16	−2.18	1.51	−0.28
RCA-MICE	−0.31	−2.20	1.50	−0.34
Assumption: Differential Non-Response 2				
REG	−0.74	−0.34	**−7.38**	−2.82
RC-REG	−0.22	−0.03	**−9.20**	−3.15
RCA-REG	−0.08	−0.01	**−9.04**	−3.04
MICE	1.24	0.59	**−7.56**	−1.91
RC-MICE	0.88	0.19	**−8.92**	−2.62
RCA-MICE	1.05	0.22	**−8.71**	−2.48

Bold figures indicate that the relative bias exceeds 5 percent.

method (Table 8).[10] Overall, the relative bias of standard errors is smallest for the imputation of home market values, it is slightly higher for financial assets under MAR and DNR1, and it is the highest for any of the imputation of consumer credits. The negative impact on standard errors by RC or RCA is not alarming in any of the cases analyzed here. Hoogland and Boomsma (1998) suggest that the relative bias of standard errors should not exceed 5 percent; here, only standard errors of the imputed values of consumer credits are showing worrisome results, but they do not indicate that a specific method yields considerably worse results. As for the intuition, why consumer credits are impacted the most, apparently once small liability values are missing (DNR1), imputation data sets tend to overstate the standard errors, and vice versa (DNR2 and MAR, see Table 2). This appears to be the result of generally poor imputation models, as the set of covariates with high explanatory power is smaller than for other assets in the SOEP as well as considerably less observations to rely on. Our experience with SOEP data shows that it is substantially more challenging to impute for missing liability values, which is reflected by the results of this study.

7 Conclusion

In an assessment of the performance of several imputation methods for longitudinal wealth data we use a set of eight evaluation criteria and three assumptions for the nonresponse mechanism. The overall result does not yield that a single imputation method performs consistently better for all wealth types in a cross-sectional and longitudinal analysis. We compare the row-and-column imputation (with or without age classes) for observations with available longitudinal data with two methods that rely on the prediction equations of regression models. 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.

As for the trend and inequality evaluation, if the missing data are truly missing at random (MAR), for all three assets we consider the combination of MICE and row-and-column imputation is at least among the best performing methods. Unexpectedly, this holds true independently of the level of state-dependency prevalent in the items. If the missing data

[10] As part of the online appendix we present the results in a less condensed form, allowing to differentiate for all three waves separately.

are missing not at random and instead are the result of differential non-response (DNR1 and DNR2) the combination of the row-and-column imputation with MICE does improve the performance in our evaluation study as well. This is the core outcome of this study: If the missing at random assumption is violated, the row-and-column imputation technique yields less biased overall imputation results for trend and inequality estimates. We like to stress that – based on this study and our experience with data imputation – this conclusion holds only true for variables that are highly skewed (such as assets, net worth or income variables). The imputation technique itself – and thus an improvement of the performance – is applicable to panel data only.

Furthermore, we find that adding age classes to the standard row-and-column imputation as introduced by Little and Su (1989) does not regularly improve the performance based on our criteria and the input data. However, there is an advantage for the imputation of consumer credits.

As for the wealth mobility criteria, the conclusions are less clear. Generally, MICE seems to reproduce the observed mobility structures better than the regression approach, in many cases the combinations of MICE and the row-and-column imputations yield satisfying results, too. However, it is clearly noticeable that for most assets and non-response assumptions the mobility is reduced, once the row-and-column imputation is applied. Hence, a data provider needs to weigh the options: for the SOEP we decided that the method of choice depends on data usage; as the data are mainly used for trend and inequality analyses and much less for mobility analyses, we opt for the combination of MICE and the row-and-column imputation.

One thing that remains to be addressed is that we refrained from including partial unit non-response (PUNR) in this simulation, e. g. individuals within households that choose not to 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.

References

Allison, P. D. (1987). Estimation of linear models with incomplete data. *Sociological Methodology*, *17*, 71–103.

Barceló, C. (2006). Imputation of the 2002 wave of the Spanish Survey of Household Finances (EFF). Banco de Espana, Documentos Ocasionales OP-0603. Retrieved from http://dx.doi.org/10.2139/ssrn.901584

Christelis, D. (2011). Imputation of missing data in waves 1 and 2 of share. Retrieved from http://dx.doi.org/10.2139/ssrn.1788248

European Central Bank (ECB). (2013a). The eurosystem household finance and consumption survey. methodological report for the first wave. Statistics Paper Series 1, Frankfurt/Main: ECB. Retrieved from http://www.ecb.europa.eu/pub/scientific/stats/html/index.en.html

European Central Bank (ECB). (2013b). The eurosystem household finance and consumption survey. results from the first wave. Statistics Paper Series 2, Frankfurt/Main: ECB. Retrieved from http://www.ecb.europa.eu/pub/scientific/stats/html/index.en.html

Frick, J. R. & Grabka, M. M. (2005). Item nonresponse on income questions in panel surveys: incidence, imputation and the impact on inequality and mobility. *Allgemeines Statistisches Archiv*, *89*, 49–61.

Frick, J. R., Grabka, M. M., & 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. Retrieved from http://hdl.handle.net/10419/86162

Frick, J. R., Grabka, M. M., & Marcus, J. (2010). Editing und Multiple Imputation der Vermögensinformation 2002 und 2007 im SOEP. SOEP Survey Papers, No. 146. Retrieved from http://hdl.handle.net/10419/86163

Hayes, C. & Watson, N. (2009). HILDA Imputation Methods. Hilda Project Technical Paper Series 2/09. Melbourne Institute of Applied Economic and Social Research. Retrieved from https://www.melbourneinstitute.com/hilda-research/Survey_Methods_and_Data.html

Hoogland, J. J. & Boomsma, A. (1998). Robustness studies in covariance structure modeling: an overview and meta-analysis. *Sociological Methods and Research*, *26*, 329–367.

Kalton, G. (1986). Handling wave nonresponse in panel surveys. *Journal of Official Statistics*, *3*, 303–314.

Kennickel, A. (1991). Imputation of the 1989 survey of consumer finances: stochastic relaxation and multiple imputation. Federal Reserve Working Paper Series, Washington DC.

Kennickel, A. (1998). Multiple imputation in the survey of consumer finances. Federal Reserve Working Paper Series, Washington DC.

Kennickel, A. & Woodburn, R. L. (1997). Consistent weight design for the 1989, 1992, and 1995 SCFs, and the distribution of wealth. Federal Reserve Board, Survey of Consumer Finances Working Papers.

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

Little, R. J. A. & Su, H. L. (1989). Item nonresponse in panel surveys. In D. Kasprzyk, G. Duncan, G. Kalton, & M. P. Singh (Eds.), *Panel surveys* (pp. 400–425). New York: Wiley.

Mahalanobis, P. C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, *2*, 49–55.

Panel Study of Income Dynamics. (2011). Documentation for the 2007 PSID supplemental wealth file. release 2: march 2011. Retrieved from http://psidonline.isr.umich.edu/Data/Documentation/wlth2007.pdf

Riphahn, R. & Serfling, O. (2005). Item non-response in income and wealth questions. *Empirical Economics*, *30*, 521–538.

Royston, P. (2004). Multiple imputation of missing values. *Stata Journal*, *4*, 227–241.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 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.

Uhrig, N., Bryan, M., & Budd, S. (2012). UKHLS Innovation Panel household wealth questions: preliminary analysis. Understanding Society Working Paper Series No. 2012 – 01, January 2012. Retrieved from https://www.iser.essex.ac.uk/publications/working-papers/understanding-society/2012-01

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

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

Vermeulen, P. (2014). How fat is the top tail of the wealth distribution? Working Paper Series 1692, European Central Bank.

Wagner, G. G., Frick, J. R., & Schupp, J. (2007). The German Socio-Economic Panel Study (SOEP) – scope, evolution and enhancements. *Journal of Applied Social Science Studies*, *127*, 161–191.

Watson, N. & 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). Retrieved from https://www.melbourneinstitute.com/hilda-research/Survey%5C_Methods%5C_and%5C_Data.html

Watson, N. & Starick, C. (2011). Evaluation of alternative income imputation methods for a longitudinal survey. *Journal of Official Statistics*, *27*, 693–715.