

Grabka, Markus M.; Westermeier, Christian

Research Report

Editing and multiple imputation of item non-response in the wealth module of the German Socio-Economic Panel

SOEP Survey Papers, No. 272

Provided in Cooperation with:

German Institute for Economic Research (DIW Berlin)

Suggested Citation: Grabka, Markus M.; Westermeier, Christian (2015) : Editing and multiple imputation of item non-response in the wealth module of the German Socio-Economic Panel, SOEP Survey Papers, No. 272, Deutsches Institut für Wirtschaftsforschung (DIW), Berlin

This Version is available at:

<https://hdl.handle.net/10419/116767>

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.

SOEP Survey Papers

Series C – Data Documentation

SOEP – The German Socio-Economic Panel study at DIW Berlin

2015

Editing and Multiple Imputation of Item Non-response in the Wealth Module of the German Socio-Economic Panel

Markus M. Grabka and Christian Westermeier

Running since 1984, the German Socio-Economic Panel study (SOEP) is a wide-ranging representative longitudinal study of private households, located at the German Institute for Economic Research, DIW Berlin.

The aim of the SOEP Survey Papers Series is to thoroughly document the survey's data collection and data processing.

The SOEP Survey Papers is comprised of the following series:

Series A – Survey Instruments (Erhebungsinstrumente)

Series B – Survey Reports (Methodenberichte)

Series C – Data Documentation (Datendokumentationen)

Series D – Variable Descriptions and Coding

Series E – SOEPmonitors

Series F – SOEP Newsletters

Series G – General Issues and Teaching Materials

The SOEP Survey Papers are available at
<http://www.diw.de/soepsurveyspapers>

Editors:

Dr. Jan Goebel, DIW Berlin

Prof. Dr. Martin Kroh, DIW Berlin and Humboldt Universität Berlin

Prof. Dr. Carsten Schröder, DIW Berlin and Freie Universität Berlin

Prof. Dr. Jürgen Schupp, DIW Berlin and Freie Universität Berlin

Please cite this paper as follows:

Markus M. Grabka, Christian Westermeier. 2015. Editing and Multiple Imputation of Item Non-response in the Wealth Module of the German Socio-Economic Panel. SOEP Survey Papers 272: Series C. Berlin: DIW/SOEP

ISSN: 2193-5580 (online)

Contact: DIW Berlin
SOEP
Mohrenstr. 58
10117 Berlin

Email: soeppapers@diw.de

Editing and Multiple Imputation of Item Non-response in the Wealth Module of the German Socio-Economic Panel

Berlin, August 2015

Markus M. Grabka, Christian Westermeier*, DIW Berlin

1 Introduction

This documentation describes the preparation of the wealth data from the Socio-economic Panel Study (SOEP). Wealth information was collected on the individual level in 2002, 2007 and 2012. The objective of this documentation is to address questions concerning missing values and their editing and imputation.

Large-scale surveys are usually facing missing data, which poses problems for researchers and research infrastructure providers alike. Wealth is considered a sensitive information that is usually collected with rather high nonresponse rates compared to less sensitive questions such as pure demographic variables like age, sex, migration status (e.g. Riphahn and Serfling 2005, Frick et al. 2010). In longitudinal studies for some missing values might be past or future data points available. The question arises how to successfully transform this advantage into improved imputation strategies.

Having said that, 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 1987, 1996). The drawbacks of case-wise deletion strategies have been well documented (Little & Rubin 1987). Multiple imputation addresses this issue here.

Since the mid-1980s there was no consistent, complete micro data available on the wealth of private households in Germany, in particular on their private business equity. Furthermore, there was a lack of systematically collected data on the wealth of high-income earners. Since 2002 the German Socio-Economic Panel (SOEP) attempted to fill this gap by collecting information on private household wealth, providing new insight into this important issue.

In population surveys, assets are usually recorded at the household level. In this context, the SOEP methodology has a special feature since it records the individual assets of each respondent aged 17 or over. In contrast to only recording

* Corresponding author, contact at cwestermeier@diw.de, +49 30 89789-223.

household assets by a reference person, this approach can show differences within households and partnerships while it still allows the individual worth to be added to obtain a result for a particular household. Thanks to this change of observation unit, it is possible to analyze the distribution of assets and liabilities not only at the household level but also at the individual level, and thus also to look at the wealth distribution within households or between spouses or partners (see Frick, Grabka and Sierminska 2007). However, the wealth of children is not considered in the SOEP, thus the wealth aggregate will be slightly underestimated.

The 2002 SOEP questionnaire surveys nine asset and liability components. These include information on owner-occupied housing, other property, financial assets, business assets, tangible assets and private pensions (including life insurance). On the liability side mortgage debts for owner-occupied property, mortgage on other property and consumer credits are separately asked for. While in 2002 the item “private pensions and life insurances” included building loan contracts, the question was split up in subsequent waves, i.e. in 2007 and 2012 ten separate components are surveyed.

In addition to durable consumer goods (including vehicles), other types of assets are not recorded in the SOEP. These include cash, the value of livestock and crops, equipment, intangible fixed assets, claims against private health insurance companies, commercial loans and commercial holdings in residential buildings.

One additional major shortcoming is the lack of information on pension entitlements through both company pensions and the statutory German social pension fund (“Gesetzliche Rentenversicherung” for blue-collar and white-collar workers as well as the pension entitlements for civil servants), due largely to the difficulty of obtaining data on pension entitlements for individuals still in the labor force.¹

Like other population surveys, SOEP is affected by measurement error. This is especially true for questions on wealth. A typical type of measurement error is item non-response (INR), i.e., the failure to collect complete information on a specific item. Partial unit-non-response (PUNR) occurs in household surveys like SOEP when one or more members of a multi-person household do not take part in the survey while the rest do. An aggregation of wealth holdings across all members of a given household presumably leads to underestimation in the case of PUNR. Another problem arises from inconsistent information provided by members of the same household or couples sharing a specific wealth component: for example, couples who co-own their home. Here, the SOEP questionnaire asks for (an estimate of) the current market value of the home as well as the percentage share thereof owned by that individual. As such, the market value estimated by each of the two partners should coincide. Secondly, if the two partners are sole owners of the property, their respective shares should add up to 100%.

Any deviation from this must be considered measurement error and corrected through some form of “editing” as opposed to “imputation”, which is used for missing information due to item non-response. Not only does the SOEP conduct extensive consistency checks on the individual data, but it also uses multiple imputations to replace all missing asset and liability values. Due to the use of

¹ For the relevance of pension entitlements see Frick and Grabka (2010).

longitudinal data from the repeated wealth surveys in 2002, 2007, and 2012, the quality of the imputation can be improved in contrast to a single imputation.

After extrapolation and weighting factors are applied, the SOEP micro data give a representative picture of the sample in households and thus allow conclusions to be drawn about the entire population. Members of the population in institutions (for example, in nursing homes) were not taken into account. The weighting factors correct differences in the designs of the various SOEP samples² as well as the participation behavior of respondents after the first interview. The framework data of the German micro census is adjusted to increase its compatibility with official statistics.

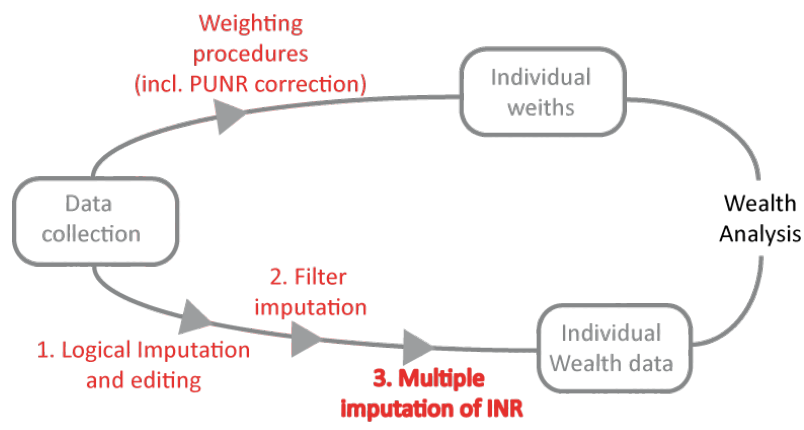


Figure 1 | Overview: From data collection to wealth analysis.

The Paper is organized as follows: Section 2 gives an overview of item nonresponse in the SOEP wealth data; Section 3 describes the surveyed wealth components using the 2012 questionnaire and the applied editing procedures and logical imputations in detail. Section 4 explains the filter imputations and multiple imputation methods for missing metric values. In Section 5 we compare various data sources with the SOEP wealth data and give an overview of restrictions to keep in mind, when using the data. Section 6 addresses differences due to a revision of weighting and imputation for SOEPv29. Section 7 concludes with a description of the data sets PWEALTH and HWEALTH as well as a list of variables included.

2 Editing and Imputation – an overview

The SOEP wealth module collects 10 different types assets and debts: value of owner-occupied and other property (and their respective mortgages), private insurances, building loan contracts, financial assets (such as savings accounts,

² In 2012 a new random subsample was drawn in the SOEP, however, all wealth questions were removed in order to not endanger the willingness to take part in the SOEP survey by asking rather sensitive questions in the very first wave. Thus, appropriate weighting factors are available for the relevant subsamples A through J.

bonds, shares), business assets, tangibles and consumer credits (see the original questions in the appendix).

In principle, at first 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 interviewee is the sole owner or, if the asset is shared, the individual share).

The first step in handling measurement errors in the SOEP wealth questionnaire is to check for the consistency and plausibility of information across household members. Information is harmonized, i.e. edited, using specific rules which will be presented in the following sections. In our context, editing means changing a non-missing value into a new value (possibly including values of zero), while imputation is used for correcting non-response. For selected components, however, imputation may be carried out by means of (single value) logical imputation, e.g., in the case of co-owner couples with one partner providing plausible information on his/her share of the wealth and the other providing none. In general, logical imputation involves an imputation derived from survey information given in the household questionnaire, by other household members or by valid information from other survey waves. The main imputation process starts with the replacement of item non-response on filter questions.

The filter variable indicates whether or not a given individual holds a specific wealth component. If this information is missing, it is imputed by a logistic regression, in each case using the specific set of covariates best suited for explaining variance in the missing filter information. Logistic regression is also applied for item non-response of the individual share of a specific wealth component because in the vast majority of cases, ownership is either 50% or 100%. The imputation of item non-response on all missing metric wealth data is addressed by a two-step procedure (see section 4).

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

Table 1 | Incidence of item non-response for individual wealth components in SOEP

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 home market value	-	-	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 home market value	-	-	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 home market value	-	-	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 Wealth Components of Private Households in the SOEP and Using the Data

3.1 Home market value

Perhaps the most important wealth component for private households in Germany is owner-occupied property. In the SOEP wave 2012, this component is surveyed as shown below, starting with a basic filter question followed (gross) market value, outstanding debts and the individual share of the property:

A Are you personally the owner of the house or apartment in which you live?

Yes..... ☐ 

No..... ☐ 

Value:
If you were to sell it today, how much would you receive for your house/apartment including land? euros

Loans:
If you are still paying off a loan on your house/apartment, how much is left to repay (excluding interest)? euros
All loans are paid off in full. .. ☐

Personal share of property:
Are you the sole owner (100%) or co-owner (e.g., with your spouse)? Sole owner ☐
If co-owner, what share do you own? Share in %.....

The editing process for owner-occupied property can be divided into three steps. First, the answers given in the individual questionnaire are checked for consistency with the individual questionnaires of the other household members. Second, the individual answers are checked for consistency with information from the household questionnaire. Third, missing values are imputed using logical imputation.

The SOEP makes it possible to link information on housing tenure from the regular household questionnaire to wealth information on owner-occupied property from the individual questionnaire. Consistency becomes an issue particularly in cases of co-ownership within one household. If inconsistent, data on the filter, personal share, and both metric values (market value and debt) may have to be edited. In the following we briefly describe the filter's intended purpose and the procedures used.

Filter of owner-occupied property

Aim: To accurately identify the owner (or holder or proprietor) within the household. A first preliminary consistency check among all household members is conducted to clarify the ownership status, especially between parents and children. A property can be owned by parents, by children or by both parents and children.

Market and debt value

Aim: To obtain consistent information on the market value of the property and the outstanding debts of each owner in the household. Several cases may arise:

- 1.) If the values for “market values” and/or “debts” given by co-owners of the same property differ by not more than 30%, the average value is applied to the respective individuals.
- 2.) Larger differences (interpreted as measurement error) arise from one co-owner giving an exact amount in euros, and the other co-owner basically stating the same value but in thousands of euros, i.e., dropping the last three digits of the same amount as mentioned by the first co-owner. In such (and similar) cases, the most plausible value is chosen on the basis of a case-wise check exploring regional information, size of housing unit, type of dwelling, etc.
- 3.) If the market value stated by one co-owner is about twice that stated by the other co-owner, in most cases, the higher value is chosen as the “correct” market value after performing case-by-case checks. Here, the basic assumption is that the smaller value relates to the person’s individual share instead of the total market value of the property.
- 4.) If the information provided by two co-owners (usually couples) on outstanding debts differs significantly, one of the two values is chosen and assigned to the other after performing case-by-case checks of occupancy, market value, income and monthly loan payments. If neither of the two values seems more plausible than the other, the average value is taken.
- 5.) If one co-owner states a positive value for the level of outstanding debts and his/her co-owner states “no debt”, the positive value is generally taken following case-by-case checks of other kinds of debts, monthly loan payments, occupancy etc.

Share of owner-occupied property

Aim: To prevent double-counting, that is, to ensure that the sum of the individual shares of one owner-occupied property does not exceed 100% within the same household.

- 1.) If both partners (or co-owners) claim to be sole owners (i.e., each owning 100%), or one partner claims to own 100% and the other states ownership of 50%, both shares are set to 50%.
- 2.) If one person declares to be the sole owner and the other states that his/her share is x (with $0 < x < 50\%$), the first person’s share is set to $100 - x$.
- 3.) If two persons in a household state the same share of more than 50%, it is assumed that this value gives the share both partners hold in common, and the remainder is owned by a third party not belonging to the household.
- 4.) If the overall household share is marginally less than 100% presumably due to rounding, the existing individual values are adjusted in order to achieve a sum of exactly 100% (e.g., 66% and 33% are changed to 67% and 33%, respectively).

Logical imputations

Before turning to the standard case of regression-based imputation of missing values, this section describes the logical imputation of market value for owner-occupied property and outstanding mortgage debt based on information given in the household questionnaire and other household members' individual questionnaires. We assume any valid information on owner-occupied housing given by other co-owners residing in the household affected by non-response to be superior to any other imputation routine, given that the information provided by (at least one of) the co-owners will consider the specific characteristics of the relevant property more explicitly than an imputation algorithm can do, the latter being subject to potential bias resulting from the restricted set of covariates (i.e., an omitted variable bias cannot be ruled out).

Imputation of missing filter information on owner-occupied property

Aim: To accurately define the proprietor within a family in case of INR and PUNR using information from the household questionnaire and information provided by other household members.

It should be noted that most of the cases dealt with in this section are affected by PUNR.

- 1.) Individuals with PUNR and those with INR on the central filter question asking for owner-occupied housing are set to "no owner" if the household is renting its home.
- 2.) If 100% ownership of the dwelling is claimed by another party, PUNR and INR on the filter question are coded to "no owner" assuming that there exist no other potential co-owners.
- 3.) If the sum of the co-owners' shares within the household is less than 100%, the filter of those individuals with PUNR/INR is imputed after case-wise checks of age and relationship to the head of household. It is assumed that very old and very young persons are not owners. If the filter is set to "owner", the individual share for this individual takes on the value of 100% minus x , with x being the cumulative share of the remaining co-owners.

Imputation of missing market values and outstanding mortgage debt

Aim: To achieve consistent information among all (co-)owners in the household on market value and amount of outstanding debts.

- 1.) If a plausible value is stated by only one co-owner, this value is also assigned to the remaining co-owners.
- 2.) If information on the level of outstanding mortgage debts is not given in the individual wealth questionnaire, information on monthly mortgage payments (e.g. variable SH32) from the household questionnaire is used to derive whether an individual is an outright owner. For the remaining individuals with PUNR or INR on the target variables, the level of outstanding mortgage debt still needs to be imputed (see Section 3.3).

3.) For owner-occupying households with no valid information on monthly mortgage payments (e.g. variable SH32) from the household questionnaire, the following logical imputations are carried out: if the household has inherited the dwelling or if occupancy has lasted for more than 25 years, it is assumed that the household is debt-free. In all other cases, the level of outstanding mortgage debt needs to be imputed (see Section 3.3.).

Imputation of missing personal shares of own property



Aim: To ensure that the sum of personal shares across all co-owners in a given household adds up to 100%. If the personal share of owned property is missing for at least one co-owner, we impute the missing information (as 100% minus the sum of all valid shares), i.e., we assume no ownership by parties living outside the household.

- 1.) In cases where two co-owners fail to give information on their respective share (INR), both individuals are assigned 50% of ownership.
- 2.) If one co-owner declares an individual share of x , the non-responding co-owner is assigned a share of 100% minus x .
- 3.) Owners living in single households are declared to be sole owners.

3.2 Other property

The second wealth component is “other property”. Other property refers to real estate that is owned by a given individual but not used as the principal residence. This set of variables encompasses the following information: the corresponding filter variable, the type of property, the number of other properties, the gross market value of thereof, the personal share, and the sum of outstanding debts related to this other property.

B Apart from the home you live in, do you own any other homes or land?

Yes.... ☐ 
 No.... ☐ 

Type and number of properties:	
What types of property are they?	
Single-family home / apartment (not used by you).....	<input type="checkbox"/>
Multi-family home / apartment building.....	<input type="checkbox"/>
Vacation home / weekend home	<input type="checkbox"/>
Undeveloped land	<input type="checkbox"/>
Other property	<input type="checkbox"/>
How many of these properties do you own (excluding the one where you live)?	Amount..... <input type="text"/>
Value:	
If you were to sell your property today (excluding the one where you live), how much would you receive?	euros <input type="text"/>
Personal share of property:	
Are you the sole owner (100%) or co-owner (e.g., with your spouse)?	Sole owner <input type="checkbox"/>
If co-owner, what share do you own?	Share in %..... <input type="text"/>
Loans:	
If you are still paying off a loan on your property, how much is left to repay (excluding interest)?	euros <input type="text"/>
	All loans are paid off in full. ... <input type="checkbox"/>

Internal checks of consistency and logical imputations

Aim: To achieve consistent wealth information in case of co-ownership of “other property”. Given the lack of information about other properties in the household questionnaire (which could have been used as an “external” benchmark) as well as the potential heterogeneity of the components included in this category, the philosophy for changing data is a rather conservative one. In other words, data is only edited if the basic information provided by co-owners living in the same household (mostly couples) with respect to type and amount of other property is not contradictory.

Editing and logical imputations of market value and debt has only been carried out, if deviations between the separate values within one household did not exceed one-third, as then the information likely covers the same estate. Two respondent’s differing market or debt values given for the same object then got replaced by the mean value. Is the deviation exceeding one-third, we check, whether there has been an error due to missing digits. In contrast to home market values the filter information have not been edited or logically imputed. Is information on the remaining debt missing, we check the household questionnaire for additional clues, the imputation of the values and remaining filters are addressed by (multiple) imputation.

3.3 Financial assets

Financial assets are the most prevalent wealth component in Germany, but given the large degree of heterogeneity in the potential components thereof one can assume higher non-response here than in the case of owner-occupied property, which mostly refers to a single object only. Given this and keeping in mind the difficulty of achieving a high response rate when collecting information on such complex issues, it was decided to ask only those individuals with a “significant” amount of wealth for information on their financial assets, setting the threshold at 2,500 euros in 2002. In the subsequent waves 2007 and 2012 this threshold was removed. The information gathered on this wealth component is the filter variable, the total value of the assets and the personal share of the assets held by the individual.

- D** Do you own financial assets in the form of a savings account, bonds (*Sparbrief* / *Pfandbrief*), shares, or investments?

Yes..... ☐ →
No..... ☐ ↓

Value:	
What do you estimate to be the value of your financial assets?	euros <input type="text"/>
Personal share of property:	
Are these financial assets in your name or do you share them through joint accounts, i.e., with your spouse?	Sole owner..... <input type="checkbox"/>
If shared, what share do you own?	Share in % <input type="text"/>

Checks of consistency

Consistency checks using information on financial assets from the household questionnaire are not applied. In 2002, in households that are less well-off

financially, the problem arises that although none of the individual household members owns above the 2,500 euro threshold used in the individual wealth questionnaire, it cannot be ruled out that overall financial assets aggregated across all household members exceeds this amount.

Logical imputations for co-owners

Aim: To ensure consistent information among co-owners. Changes are performed only if there is a strong indication that the head of household and spouse/partner share their financial assets equally (i.e., 50% each). After logical imputation, the value of financial assets is identical for all individuals who appear to own the same (set of) financial assets. Thus logical imputation is only conducted in very few cases.

3.4 Building loan contracts, private insurances, business assets, tangible assets, consumer credits

There is no sufficiently comparable (metric) information available in the household questionnaire on any of those components. Standard SOEP data includes qualitative information on the existence of various kinds of assets and the total amount of interest and dividends received from these investments. However, none of these components correspond perfectly with the wealth components defined in the individual wealth questionnaire.

Imputation of the personal share

For the four components considered here, respondents are asked to specify their personal share only in the case of business assets. Missing personal shares are imputed using a logistic regression model estimating the probability that someone is sole owner or co-owner of an enterprise, the latter being interpreted as a personal share of 50%.

Building loan contract

Ⓒ Do you personally have a building loan agreement (*Bausparvertrag*)?

Yes..... ☐ ➔
No..... ☐ ↓

Value:
What would you estimate the building loan credit balance (*Bausparguthaben*) to be, including interest and dividends?
If you have more than one such agreement, please add them together and state the total sum!

euros

Private insurances

Note that 2012 the phrase “including Riester or Rürup pensions” was added to the questionnaire, as those pensions are quite common and may be underestimated otherwise. For comparison’s sake the 2002 and 2007 questionnaires are in Appendix A and B.

- (E) Do you have a life insurance policy or private retirement plan (including Riester or Rürup pensions), with contributions paid either by you or by your employer?

Yes..... ☐ →
 No..... ☐ ↓

Value: What do you estimate the current cash surrender value or balance of these plans to be?	euros	<input type="text"/>
--	-------	----------------------

Business assets

- (F) Are you the owner of a commercial enterprise, i.e., a company, a shop, an office, a practice, or farm, or are you co-owner of one of these types of enterprises?

Yes..... ☐ →
 No..... ☐ ↓

Personal share of property:	
Are you the sole owner or co-owner of this enterprise, e.g., GBR, GmbH, or KG?	Sole owner..... <input type="checkbox"/> Co-owner..... <input type="checkbox"/>
Value: What do you estimate the current value of your enterprise or your share thereof to be? This is the price before taxes that you would receive on the sale of your enterprise or your share thereof, after deducting any remaining debts.	
euros	<input type="text"/>

Tangible assets

- G Do you own any tangible assets in the form of gold, jewelry, coins, or valuable collections?

Yes..... ☐ →
 No..... ☐ ↓

Value: If you could sell these assets, what would their total value be?	euros	<input type="text"/>
--	-------	----------------------

Consumer Credits

- (H) Leaving aside any mortgages on homes and property or building loans:
 Do you currently owe money on loans that were granted to you personally by a bank,
 other institution, or individual for which you personally are liable?
Not including mortgages or building loans!

Yes..... ☐ →
 No..... ☐ ↓

Loans: What are your outstanding debts?	euros	<input type="text"/>
---	-------	----------------------

Question 91
 next page!

4 Imputation Methods

The first step in every imputation procedure that accounts for INR in a 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). 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 will yield unbiased estimators, but will also result 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 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.

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 1987, 1996). The drawbacks of case-wise deletion strategies have been well documented (Little & Rubin 1987). Table 2 gives an overview of the impact of case-wise deletion strategy to address missing values in the SOEP wealth data.

Table 2 | Changes in net worth after editing and imputation

Net worth 2002	Valid observations	After editing and imputation	Change in %
Observations n	14,017	23,155	67.3
Mean	62,630	83,886	33.9
Median	5,000	15,000	200.0
Gini coefficient	0.83	0.79	-4.2
Share negative or zero net worth	40.45 %	28.03 %	-30.7

Net worth 2007	Valid observations	After editing and imputation	Change in %
Observations n	12,923	20,965	62.2
Mean	60,497	83,853	38.6
Median	3,500	14,425	312.1
Gini coefficient	0.88	0.81	-7.9
Share negative or zero net worth	40.88 %	28.07 %	-31.3

Net worth 2012	Valid observations	After editing and imputation	Change in %
Observations n	11,168	18,356	64.3
Mean	66,475	85,670	28.8
Median	5,000	16,840	236.8
Gini coefficient	0.85	0.78	-8,2
Share negative or zero net worth	39.54 %	27.7 %	-29,9

Source: SOEPv29, data set PWEALTH. Means over 5 implicates. Weighted results.

Wealth is considered a sensitive information that is usually collected with rather high nonresponse rates compared to less sensitive questions such as pure demographic variables like age, sex, migration status (e.g. Riphahn and Serfling 2005, Frick et al. 2010). In addition there is a rather high state-dependency in terms of ownership status of wealth components (e.g. owner occupied property), which facilitates the consideration of longitudinal information in the imputation process.

The most commonly used assumption about the nonresponse 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, 77). Hence, within the framework of the applied imputation methods we assume that the missing values do not depend on variables we did not include in our imputation models.

For the imputation of the SOEP wealth data v29 we opted for multiple imputation by chained equations (MICE, named after one of the first popular implementations, see Royston 2004). We additionally use a univariate imputation procedure for panel data known as the row-and-column method introduced by Little and Su (1989) if information from other waves is available for missing values.

Table 3 | Basic and fallback imputation methods for the imputation of item non-response for metric value

BASIC (for observations with missing values, information from other waves is available)	Row-and-column imputation (Little and Su 1989, L&S)
FALLBACK (for some observations with missing values, only cross-sectional information is available)	Multiple imputation by chained equations (MICE)

The reasoning behind the change of the imputation method compared to the 2002 and 2007 waves is as follows: In a simulation study we compared six combinations of cross-sectional and longitudinal imputation strategies by applying to the SOEP wealth data (Westermeier and Grabka 2014). The comparison was conducted creating simulation data sets by setting observed data points to missing based on three separate nonresponse generating mechanisms. The performance of imputation models was analyzed assuming the mechanisms are missing at random (MAR) or the data suffers by differential nonresponse at the bottom or the top of the wealth distribution. In principle, MICE was tested against the row-and-column method and the previously applied regression imputation with correction for sample selection, which was the standard imputation method for the SOEP wealth data in past survey waves. These three basic imputation methods have been tested in various combinations using the row-and-column-imputation as basic longitudinal imputation and MICE imputation or the regression approach as a fallback procedure. Overall, six different variants were analyzed.

The overall result did not yield that a single imputation method performs consistently better for all wealth types in a cross-sectional and longitudinal analysis. For the trend analyses, if the missing data are truly missing at random (MAR) the differences between the model based and the row-and-column approaches were generally rather small. However, for all three assets we considered, the combination of MICE and row-and-column imputation was at least among the best performing methods. Unexpectedly, this holds true independently of the level of state-dependency prevalent in the items.

Depending on the asset we applied the row-and-column method in roughly 50 percent of the cases. Due to higher panel attrition between 2007 and 2012 compared to 2002/2007 the row-and-column share is somewhat higher in 2007.

4.1 (Single) imputation of the filter variable

Missing filter data is imputed by means of logistic regression. For each component, there are separate models for INR and PUNR, each using individual information on sex and age as well as a wide range of covariates from the household level. The exact list of control variables, however, slightly differs for the various wealth components. In line with the procedures described above, predicted values below 0.5 are rounded down to zero, assuming that the person does not own the respective wealth component and vice versa for predicted values greater or equal to 0.5.

4.2 Multiple imputation of missing metric values

4.2.1 Basic method: row-and-column imputation technique

Little and Su (1989) proposed the row-and-column 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 there are three waves of wealth data, the column effects (for any wealth asset) are given by

$$c_t = \frac{(3 * \bar{y}_t)}{\sum_k \bar{y}_k} \quad (1)$$

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

$$r_i = \frac{1}{m_i} * \sum_j \frac{y_{it}}{c_j} \quad (2)$$

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 :

$$\dot{y}_{it} = r_i * c_t * \overset{\text{residual term}}{\left(\frac{y_{lt}}{r_l * c_t} \right)}. \quad (3)$$

\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 (3), we assign the residuals of the k nearest neighbors. Then terms (1) and (2) 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)). 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.

4.2.2 Fallback method: multiple imputation by chained equations (MICE)

Multiple imputation by chained equations (MICE) is an iterative and sequential regression approach that grew popular among researchers, because it demands very little technical preparation and is rather easy to use. 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. 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, \dots, Y_p and a set of predictors (without missing values) Z , then for iterations $n = 0, 1, \dots, N$, and with ϕ_j as the corresponding model parameters with uniform prior probability distribution, the missing values are drawn from

$$\begin{aligned} Y_1^{(n+1)} &\sim g_1(Y_1|Y_2^{(n)}, \dots, Y_p^{(n)}, Z, \phi_1) \\ Y_2^{(n+1)} &\sim g_2(Y_2|Y_1^{(n+1)}, Y_3^{(n)}, \dots, Y_p^{(n)}, Z, \phi_2) \\ &\vdots \\ Y_p^{(n+1)} &\sim g_p(Y_p|Y_1^{(n+1)}, Y_2^{(n+1)}, \dots, Y_{p-1}^{(n+1)}, Z, \phi_p), \end{aligned} \quad (1)$$

until convergence at $n = N$ is achieved. That is, in iteration $n + 1$ the dependent variables of each imputation model $g_j(\cdot)$ are updated with the corresponding imputed values of 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 $g_j(\cdot)$ 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 .³ 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 The choice of covariates differs only slightly between the separate imputation models. Regional variables, for instance, make the most sense for the imputation of dwellings and other properties and less so for consumer credits. The least variables were included for business assets, as models tend to be less stable. The covariates included in the imputation process cover *sex, age, region, household size, number of children, worries about own economic situation, number of years in education, civil servant dummy, unemployed dummy, household type (one family home, ...), district size/district type, interest and mortgage payments/other loan repayments, owner of shares, life insurances, businesses etc., owner occupier/tenant, mode of property acquisition, housing condition, household equipment, dwelling size, construction year, operating/maintenance costs, income from renting and leasing/interest income, household income*, and a few others.

(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 association 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 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 $g_j(\cdot)$ in (1) 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.

Implementing longitudinal information in MICE

The quality of any imputation is depending on the choice of method (Westerbeier and Grabka 2014) and the available information included in the imputation models. The row-and-column method on the one hand makes use of the rather high state-dependency in wealth data and directly draws from the information available from other waves on one specific missing item and observation. Using MICE as a fallback method the implementation of longitudinal information is less obvious. First, if information from other waves is available, the basic row-and-column imputation is used. Second, for all cases with missing information, already imputed information from other waves is a good predictor and one goal of the imputation procedure is to achieve consistency between the several waves.

In order to achieve consistency, we started the MICE imputation with a cross-sectional imputation of the 2007 wealth data. After this first imputation we implemented a cycle of imputations using the (already imputed) values from other waves, the panel information for the very same persons are not considered, as in those cases the basic row-and-column imputation method applies. This means, we iteratively specify the models for all waves conditioning on the respective asset values that have already been imputed until the models convergence.

4.2.3 Imputation of wealth below the threshold (2,500 euro) in 2002

In the survey year 2002 values for financial assets, tangible assets and consumer credits were not collected, if the respective amount was below a threshold of 2,500 euro. After a revision of the SOEP wealth module for the 2007 survey, the threshold was removed. Consequently, this decision led to the situation that incidences and aggregates are too low in comparison to later survey years. In order to correct micro and aggregate data for the 2002 wave, we used the information on those individuals and households holding low-valued assets—assuming that the characteristics of those individuals and households did not change between 2002 und 2007—, and carried out an additional imputation.

For all persons, that did not give information on financial assets, tangible assets or consumer credits in 2002, we run a logistic regression determining the probability that they owned that specific asset. The regression includes all persons that either does not hold the asset in 2007 or provided a value below 2,500 euro.⁴

The predicted probabilities from the logistic regressions are compared to random draws of a normal distribution (parameters: mean 0.5 and standard deviation 0.2). If the probability is below the randomly drawn threshold value, it is assumed that the person did not hold the component in 2002; if it is above the threshold value, it is assumed that the person did hold the component in 2002.

The value of all assets and debts are imputed using values below the threshold of 2,500 euro from survey year 2007 and randomly assigning them to the respective cases in 2002. The procedure is repeated five times in order to generate a multiply imputed data set that takes the uncertainty of the imputation into account.

⁴ Persons with missing metric values or missing filters in 2007 are not included.

5 Restrictions of the Data and Comparing with Other Data Sources

Not only does the national accounts approach face a number of methodological and statistical problems, but so too does the analysis of the distribution of wealth based on micro data representative of the population.

Neither approach takes into account—as is common the world over—the entitlements to statutory pension insurance. Accumulated pension-related claims are converted into personal earning points in Germany which do not unequivocally indicate social security assets and therefore are hardly directly ascertainable in a survey; this applies equally to occupational pension entitlements. However, since the majority of the German working population is subject to compulsory pension insurance or has pension-related claims, for example, in the form of training or childrearing periods, social security assets in the statutory pension scheme in particular can be assumed to represent the most frequent component in household net worth. Pension insurance data analyses have shown that 91 percent of men and 87 percent of women aged 65 or over have statutory pension entitlements. (In eastern Germany, the corresponding figures are even higher at 99 percent).

Other components of net worth are also commonly not addressed in population surveys since they are particularly difficult to record, such as household effects, including the value of vehicles. Neither of these two asset components flew into the concept of net worth in the SOEP. Thus, due to these limitations, in comparison to the national accounts approach, the net worth in the SOEP is, all other things being equal, underestimated. Note for example, a household's investment in a new car results in an increase of debt on liability side, which is not at least partially offset on the asset side, as cars or any other household goods are not collected.

A comparison of aggregated assets based on the SOEP and the sectoral and overall economic balance sheets of the German Federal Statistical Office (FSO) and Deutsche Bundesbank is complicated by a number of differences in distinctions and definitions. The following reasons for this are germane in this context. First, the FSO and Deutsche Bundesbank categorizes households together with private non-profit organizations. Second, in addition to durable consumer goods, other types of assets are also included in the national accounts which are not recorded in the SOEP, including cash, the value of livestock and crops, equipment, intangible fixed assets, claims against private health insurance companies, commercial loans, and commercial holdings in residential buildings.

Fourth, the SOEP generally records the current market value of real estate while the FSO calculates its replacement value. However, market value differs significantly from the replacement value of portfolio properties. As a result, the SOEP's 2002 calculation for net worth on this basis totaled almost 90 percent of the balance sheet figure arrived at by the FSO and Deutsche Bundesbank, but it was only 64 percent in 2012. In the case of real estate assets, the quantitatively most important asset component, the coverage rate fell from 129 percent in 2002 to slightly under 103 percent in 2012. Liabilities are recorded at 73 percent. With aggregate gross financial assets at 33 percent, the SOEP, as in most other wealth surveys worldwide, has significantly underestimated their value.

Table 4: Comparison of wealth aggregates in the system of national accounts (SNA) and SOEP.

	SNA			SOEP			Quota %		
in billion Euro	2002	2007	2012	2002	2007	2012	2002	2007	2012
Assets	8.052	9.818	11.348	6.929	7.278	7.425	86,1	74,1	65,4
Biological Ressources	3	5	/	---	---	---	---	---	---
Inventories	125	127	133	---	---	---	---	---	---
Produced Assets	4.339	5.203	6.261	4.780	4.868	5.142	110,2	93,6	82,1
Fixed assets	3.074	3.650	4.302	---	---	---	---	---	---
Dwellings	2.742	3.278	3.888	---	---	---	---	---	---
Other buildings and structures	331	371	414	---	---	---	---	---	---
Land & Intellectual property products	1.273	1.561	1.968	---	---	---	---	---	---
Financial assets	3.576	4.476	4.940	1.339	1.688	1.628	37,4	37,7	32,9
Currency and deposits	2.400	2.984	3.126	714	920	977	29,8	30,8	31,3
Insurance, pension and standardized guarantee schemes	1.176	1.492	1.813	1.321	1.397	1.237	112,3	93,6	68,2
Financial liabilities	1.553	1.547	1.566	1.126	1.162	1.146	72,5	75,1	73,2
Real estate loans	---	1.048	1.075	---	1.001	986	---	95,5	91,7
Consumer credits	---	200	194	---	161	161	---	80,8	82,7
Equity and investment fund shares/units, Other accounts payable	---	300	297	---	---	---	---	---	---
Net assets	6.499	8.271	9.782	5.802	6.115	6.278	89,3	73,9	64,2
Consumer durables	862	892	929	---	---	---	---	---	---
Net assets including consumer durables	7.361	9.163	10.710	---	---	---	---	---	---

Source: Statistisches Bundesamt and Deutsche Bundesbank (2013), SOEPv29. Households and non-profit institutions serving households/ SOEP private households only.

A comparison with the wealth survey conducted by the German Federal Bank in 2010/11 (Private Haushalte und ihre Finanzen, PHF) shows that the SOEP slightly underestimated per capita net worth at 86,000 euros, compared to the PHF's 95,000 euros. Here, it should also be taken into account that the PHF conducts a far more detailed survey of the asset situation, for example, also explicitly taking into account the value of vehicles.

Since 2002, the SOEP has included a subsample of "high-income households" in a concerted effort to counter the widespread problem in population surveys of not having a statistically significant subgroup of higher incomes and assets. In the context of high inequality in personal wealth distribution, this subsample and the sufficiently large number of wealthy households in the SOEP is especially important. In particular, the relationship between income and wealth distribution

for all groups, and above all for the group of high-income earners, can also be shown in greater detail, since assets, asset income, and savings depend to a large extent on disposable income. Nevertheless, despite this dedicated subsample, the problem remains that surveys such as the SOEP effectively do not contain top high net worth individuals. This applies in particular to billionaires as well as multi-millionaires with a net worth in the triple-digits million range. As a result, the true extent of wealth inequality is underestimated.⁵

Germany presently has no available external statistics, for instance, wealth tax statistics, to validate this potential underestimation. The need to provide fair market value of assets also presents such surveys with a fundamental problem. Estimating fair market value in a survey is difficult, especially when the object was inherited or purchased a long time ago and respondents do not have sufficient knowledge of the current market. As is well known, valuing business assets is also particularly difficult. In contrast to regular income, asset values can be very volatile and this further complicates their evaluation. Aside from the overall sensitivity of this issue, this in turn increasingly results in refusals to answer asset-related questions.

⁵ Westermeier, Christian and Grabka, Markus, (2015), Significant Statistical Uncertainty over Share of High Net Worth Households, DIW Economic Bulletin, 5, issue 14/15, p. 210-219.

6 Influence of the Data Revision on Asset Amounts and Distribution in 2002 and 2007

The asset data for survey years 2002 and 2007 (provided in 2010) deviates from current publications and data provided in SOEPv29 and onwards, because repeated revisions of weighting factors were required in the SOEP in the past and the imputation procedure has since undergone a fundamental reworking. Selected key figures are shown in the table 5 before and after revised weighting and improved imputation. There are no significant changes, i.e., the deviations between previous and revised data for 2002 and 2007 still fall within the usual fluctuation range of samples.

Table 5 | Influence of the Data Revision on Asset Amounts and Distribution in 2002 and 2007.

	SOEPv29			SOEPv24/v25		
	Lower threshold	Estimate 2002	Upper threshold	Lower threshold	Estimate 2002	Upper threshold
Gini	0.7737	0.7857	0.7977	0.735	0.7581	0.7811
p9050	13.3	14.0	14.7	9.1	11.3	13.5
Mean	79,163	83,783	88,403	74,460	79,263	84,065
p99	698,761	759,969	821,176	560,370	646,735	733,100
p95	311,660	323,722	335,784	296,100	316,095	336,090
p90	203,464	210,134	216,803	202,950	213,431	223,912
p75	94,046	98,130	102,214	99,757	105,650	111,542
p50	14,296	15,000	15,704	15,250	18,900	22,550
p25	0	0	0	0	0	0
p10	0	0	0	0	0	0
p5	-2,757	-1,610	-463	-3,045	-1,420	205
p1	-23,683	-20,360	-17,037	-34,592	-24,467	-14,341
Min	-6,387,866	-3,967,151	-1,546,435	-6,108,453	-1,973,343	2,161,766

People in households, individual assets; lower/upper thresholds represent a 95-percent confidence interval. Source: SOEP, without top coding.

	SOEPv29			SOEPv24/v25		
	Lower threshold	Estimate 2007	Upper threshold	Lower threshold	Estimate 2007	Upper threshold
Gini	0.7937	0.8065	0.8193	0.741	0.7659	0.7907
p9050	12.7	14.0	15.3	9.1	10.4	11.7
Mean	78,794	84,257	89,720	71,031	80,139	89,247
p99	700,282	787,500	874,718	516,211	731,860	947,509
p95	302,437	319,731	337,025	291,626	327,050	362,474
p90	199,062	207,695	216,327	188,343	206,005	223,667
p75	87,020	91,374	95,727	89,224	96,024	102,825
p50	13,409	14,818	16,228	17,508	19,833	22,158
p25	0	0	0	-310	97	504
p10	0	0	0	0	0	0
p5	-5,012	-4,000	-2,988	-4,285	-2,814	-1,343
p1	-36,299	-30,260	-24,221	-33,998	-25,880	-17,762
Min	-2,057,679	-1,500,000	-942,321	-1,788,368	-954,275	-120,181

Source: SOEP, without top coding. People in households, individual assets; lower/upper threshold represent a 95-percent confidence interval.

7 The Data Sets PWEALTH and HWEALTH

The generated SOEP wealth data is stored in two separate data files called PWEALTH for information at the individual level and HWEALTH for correspondingly aggregated data at the household level. Wealth-related variable names in the file PWEALTH consist of six digits. The first digit tells the user which wealth component is referred to, and the second to sixth digits provide more detailed information about possible filter information, the personal share, the gross amount, and the amount of any outstanding debt. In principle a digit is coded “1” if a given variable does indeed contain this specific piece of information and “0” otherwise. The code “2” indicates that this is a flag variable, showing whether or not the corresponding wealth information was imputed or edited. The wealth information in the SOEP questionnaire is surveyed at the individual level and thus also imputed or edited at the individual level (although checked against household information for consistency).

The first aggregation level is the individual level. It reports information on the share of a given wealth component the individual actually possesses. To obtain this “individual” information, a given market value referring to the object (e.g., house) needs to be multiplied by the individual percentage share operationalized by a value between zero and 100 in case of sole ownership. To give an example: the individual share of the market value of financial assets (variable F0101x with “x” referring to any of the five implicates (a, b, c, d and e)) results from the multiplication of the market value of financial assets (variable F0100x) by the individual percentage share (F00010) (see below).

The second aggregation level is the couple (legally married or cohabitating). Here, the information on the aggregate for a given wealth component held by a couple is reported. To give an example: the market value of financial assets (variable F0100x) reports all financial assets shared by the couple.

Finally, the third aggregation level is the household. Here, the amount of the total value of a given wealth component for all household members is given. To give an example: the household-level market value of financial assets (variable F010Hx) is the sum of all individual shares of financial assets (F0101x) across all household members. As such, in multi-person households with several co-owners, there is double counting in all variables carrying information on the market value of a given wealth component. Due to the additional consideration of the information on the share actually owned by an individual, there is no double counting problem in the household-level data.

For an easy identification of imputed or edited cases, all wealth variables in data sets PWEALTH (individual level) and HWEALTH (household level) have an additional flag variable. The flag is “0”, if the information is neither edited nor imputed, it is “1”, if the case was edited, it is “2” if a missing value was imputed.

HWEALTH contains all information on the household level; it is purely the result of aggregating the person-level information in PWEALTH. However, as mentioned above, for all persons with valid household level information that did refuse to respond to the personal questionnaire (partial unit non-response) imputations have been carried out and the results are included in HWEALTH.

The two data sets are provided in long format, i.e. for every observation a several row with data containing the wealth information, provided they responded in more than one wave. In order to separate between the survey years an additional variable for the survey year (SVYYEAR) is included that may as well be used to reshape the data into wide format.

7.1 Variable list at the individual level

Identifiers

PERSNR	Individual identifier
HHNRAKT	Wave specific household identifier
SVYYEAR	Survey year

Owner-occupied property

p10000	Filter information
p20000	Imputation flag for filter information
p0100x	Market value (x = implicate a,b,...,e)
p02000	Imputation flag for market value
p0010x	Debts (x = implicate a,b,...,e)
p00200	Imputation flag for debts
p00010	Individual share
p00020	Imputation flag for individual share
p0110x	Net market value ($p0100x - p0010x$; x = implicate a,b,...,e)
p02200	Imputation flag for net market value
p0101x	Individual share of market value ($p0100x * p00010/100$; x = implicate a,b,...,e)
p02020	Imputation flag for individual share of market value
p0011x	Individual share of debts ($p0010x * p00010/100$; x = implicate a,b,...,e)
p00220	Imputation flag for individual share of debts
p0111x	Individual share of net market value ($(p0100x - p0010x) * p00010/100$; x = implicate a,b,...,e)
p02220	Imputation flag for individual share of net market value

Other property

e10000	Filter information
e20000	Imputation flag for filter information
e0100x	Market value (x = implicate a,b,...,e)
e02000	Imputation flag for market value
e00010	Individual share
e00020	Imputation flag for individual share
e0010x	Debts (x = implicate a,b,...,e)
e00200	Imputation flag for debts
e0110x	Net market value ($e0100x - e0010x$; x = implicate a,b,...,e)
e02200	Imputation flag for net market value
e0101x	Individual share of market value ($e0100x * e00010/100$; x = implicate a,b,...,e)
e02020	Imputation flag for share of market value
e0011x	Individual share of debts ($e0010x * e00010/100$; x = implicate a,b,...,e)
e00220	Imputation flag for individual share
e0111x	Individual share of net market value ($(e0100x - e0010x) * e00010/100$; x = implicate a,b,...,e)
e02220	Imputation flag for individual share of net market value
e00001	Type: single-family house
e00002	Type: apartment building
e00003	Type: holiday home

e00004	Type: undeveloped real estate
e00005	Type: other property
e00007	Number of properties
e00026	Imputation flag for the type of property
e00027	Imputation flag for the Number of properties

Financial Assets

f10000	Filter information
f20000	Imputation flag for filter information
f0100x	Market value (x = implicate a,b,...,e)
f02000	Imputation flag for market value
f00010	Individual share
f00020	Imputation flag for individual share
f0101x	Individual share of market value ($f0100x \cdot f00010 / 100$; x = implicate a,b,...,e)
f02020	Imputation flag for individual share of market value

Building Loan Contract (available since 2007)

l10000	Filter information
l20000	Imputation flag for filter information
l0100x	Market value (x = implicate a,b,...,e)
l02000	Imputation flag for market value

Private Insurances (available since 2007)

h10000	Filter information
h20000	Imputation flag for filter information
h0100x	Market value (x = implicate a,b,...,e)
h02000	Imputation flag for market value

Private Insurances & Building Loan Contracts

i10000	Filter information
i20000	Imputation flag for filter information
i0100x	Market value (x = implicate a,b,...,e)
i02000	Imputation flag for market value

Business Assets

b10000	Filter information
b20000	Imputation flag for filter information
b0100x	Market value (x = implicate a,b,...,e)
b02000	Imputation flag for market value
b00001	Ownership status
b00002	Imputation flag for ownership status

Tangible Assets

t10000	Filter information
t20000	Imputation flag for filter information
t0100x	Market value (x = implicate a,b,...,e)
t02000	Imputation flag for market value

Consumer Debts

c10000	Filter information
c20000	Imputation flag for filter information
c0100x	Market value (x = implicate a,b,...,e)
c02000	Imputation flag for market value

Overall wealth

w0101x	Gross overall wealth ($p0101x + e0101x + f0101x + i0100x + b0100x + t0100x02$; x = implicate a,b,...,e)
w02020	Imputation flag for gross overall wealth
w0011x	Overall debts ($p0011x + e0011x + c0100x$; x = implicate a,b,...,e)
w00220	Imputation flag for overall debts

w0111x	Net overall wealth (w0101x - w0011x; x = implicate a,b,...,e)
w02220	Imputation flag for net overall wealth

7.2 Variable list at the household level

As a matter of principle, the wealth-related variable names at the household level carry a “H” at the fifth digit identifying the unit of analysis (household). This fifth digit at the individual level carries information on the individual share which, due to aggregation at the household level, is not a relevant piece of information as such. Imputation flag variables are also aggregated across household members, i.e., household-level wealth information is “imputed” if for at least one household member imputed data from the respective component is available.

Identifiers

HHNRAKT	Wave-specific household identifier
SVYYEAR	Survey year

Property, primary residence

p100h0	HH filter information (max of p10000 over all HH-members)
p200h0	HH imputation flag for filter information
p010hx	HH market value (sum of p0101x over all HH-members; x = implicate a,b,...,e)
p020h0	HH imputation flag for market value
p001hx	HH debts (sum of p0011x over all HH-members; x = implicate a,b,...,e)
p002h0	HH imputation flag for debts
p011hx	HH net value (p010Hx-p001Hx; x = implicate a,b,...,e)
p022h0	HH imputation flag for net value

Other property

e100h0	HH filter information (max of e10000 over all HH-members)
e200h0	HH imputation flag for filter information
e010hx	HH market value (sum of e0101x over all HH-members; x = implicate a,b,...,e)
e020h0	HH imputation flag for market value
e001hx	HH debts (sum of e0011x over all HH-members; x = implicate a,b,...,e)
e002h0	HH imputation flag for debts
e011hx	HH net value (e010Hx-e001Hx; x = implicate a,b,...,e)
e022h0	HH imputation flag for net value
e000h1	HH Type: single-family house
e000h2	HH Type: apartment building
e000h3	HH Type: holiday home
e000h4	HH Type: undeveloped real estate
e000h5	HH Type: other property
e000h7	HH Number of properties
e002h6	HH Imputation flag for the type of property
e002h7	HH Imputation flag for the Number of properties

Financial assets

f100h0	HH filter information (max of f10000 over all HH-members)
f200h0	HH imputation flag for filter information
f010hx	HH market value (sum of f0101x over all HH-members; x = implicate a,b,...,e)
f020h0	HH imputation flag for market value

Building Loan Contracts (for waves 2007 and 2012)

l100h0	HH filter information (max of i10000 over all HH-members)
l200h0	HH imputation flag for filter information
l010hx	HH market value (sum of i0100x over all HH-members; x = implicate a,b,...,e)

i020h0 HH imputation flag for market value

Private Insurances (for waves 2007 and 2012)

h100h0 HH filter information (max of i10000 over all HH-members)
 h200h0 HH imputation flag for filter information
 h010hx HH market value (sum of i0100x over all HH-members; x = implicate
 a,b,...,e)
 h020h0 HH imputation flag for market value

Private insurances and Building Loan Contracts (for wave 2002)

i100h0 HH filter information (max of i10000 over all HH-members)
 i200h0 HH imputation flag for filter information
 i010hx HH market value (sum of i0100x over all HH-members; x = implicate
 a,b,...,e)
 i020h0 HH imputation flag for market value

Business assets

b100h0 HH filter information (max of b10000 over all HH-members)
 b200h0 HH imputation flag for filter information
 b010hx HH market value (sum of b0100x over all HH-members; x = implicate
 a,b,...,e)
 b020h0 HH imputation flag for market value
 b000h1 HH business assets sole owner
 b000h2 HH business assets sole owner flag

Tangible assets

t100h0 HH filter information (max of t10000 over all HH-members)
 t200h0 HH imputation flag for filter information
 t010hx HH market value (sum of t0100x over all HH-members; x = implicate
 a,b,...,e)
 t020h0 HH imputation flag for market value

Consumer Debts

c100h0 HH filter information (max of c10000 over all HH-members)
 c200h0 HH imputation flag for filter information
 c010hx HH market value (sum of c0100x over all HH-members, x = implicate
 a,b,...,e)
 c020h0 HH imputation flag for market value

Overall wealth

w010hx HH gross overall wealth ($w010hx = p010Hx + e010Hx + f010Hx + i010Hx + b010Hx + t010Hx$; x = implicate a,b,...,e)
 w020h0 HH imputation flag for gross overall wealth
 w001hx HH overall debts: ($w001Hx = p001Hx + e001Hx + c010Hx$; x = implicate a,b,...,e)
 w002h0 HH imputation flag for overall debts
 w011hx HH net overall wealth ($w011Hx = w010Hx - w001Hx$; x = implicate a,b,...,e)
 w022h0 HH imputation flag for net overall wealth

8 Working with Multiply Imputed Wealth Data

The wealth information included in the SOEP use-files in the data sets PWEALTH and HWEALTH is multiply imputed, i.e. for every missing metric value five alternative imputed values are provided (indicated by a,b, ..., e). For all cases with validly observed values those 5 implicates will carry the same value.

In order to calculate mean or median the respective analysis has to be carried out five times for each of the five data sets. It is easiest to understand multiple imputation as carrying out the same exact analysis separately and then summing up the results in one final estimate (dividing the sum by the number of implicates.) The procedure is basically the same for all point estimates. However, multiple imputation was not invented to achieve more precise point estimates, rather its goal is that statistical testing, confidence intervals and standard errors reflect the uncertainty after missing values have been imputed.

The variance of an estimate has to be separated into a “within” and a “between” component. The “within” component of variance W of a coefficient β corresponds to the arithmetic mean of all implicates $R = 1, 2, \dots, 5$.

$$\overline{W} = \frac{1}{5} \sum_{R=1}^5 \text{var}(\beta')$$

The „between“ component of variance B is the variance of the estimated coefficients β for multiple imputation data sets $R = 1, 2, \dots, 5$.

$$B = \frac{1}{5-1} \sum_{R=1}^5 (\hat{\beta}' - \overline{\beta})^2$$

The total variance V for coefficient β is then given by combining “between” and “within” components (using 5 implicates in this case) as follows

$$\tilde{V}_{\beta} = \overline{W} + \left(1 + \frac{1}{5}\right) * B.$$

All up-to-date (statistical) analysis tools provide their users with the respective routines for multiply imputed data already implemented. For all new users it is best to consult the user manuals first, in order to correctly set up the data for the respective computational environment and calculate the correct means, standard errors and regressions.

References

- Allison, P. D. 1987. "Estimation of linear models with incomplete data."
- Barceló, C. 2006. "Imputation of the 2002 wave of the Spanish survey of household finances (EFF)." Banco de Espana. Documentos Ocasionales 0603. DOI: <http://dx.doi.org/10.2139/ssrn.901584>.
- European Central Bank (ECB). 2013a. "The Eurosystem Household Finance and Consumption Survey. Results from the first wave." Statistics Paper Series 2. Frankfurt/Main: ECB. Available at: <http://www.ecb.europa.eu/pub/scientific/stats/html/index.en.html> (accessed June 2014)
- European Central Bank (ECB). 2013b. "The Eurosystem Household Finance and Consumption Survey. Methodological report for the first wave." Statistics Paper Series 1. Frankfurt/Main: ECB. Available at: <http://www.ecb.europa.eu/pub/scientific/stats/html/index.en.html> (accessed June 2014)
- Frick, J. R.; Grabka, M.M. and Sierminska, E. 2007. "Representative wealth data for Germany from the German SOEP: The impact of methodological decisions around imputation and the choice of the aggregation unit", DIW Discussion paper #672, Berlin: German Institute for Economic Research (DIW).3
- Frick, J. R., Grabka, M.M., and J. Marcus. 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. (Available at <http://hdl.handle.net/10419/86162>)
- Frick, J.R. and Grabka, M.M. 2010. "Old-age pension entitlements mitigate inequality—but concentration of wealth remains high. DIW Weekly report, 8/2010.
- Frick, J.R., Grabka, M.M., and J. Marcus. 2010. "Editing und multiple Imputation der Vermögensinformation 2002 und 2007 im SOEP." DIW Berlin Data Documentation 51. Available at: <http://hdl.handle.net/10419/86163>.
- Heckman, J. J. 1979: "Sample Selection Bias as a Specification Error." *Econometrica* 47(1): 153-161.
- Frick, J.R., Grabka, M.M. and Sierminska, E.M. 2007. „Representative Wealth Data for Germany from the German SOEP: The Impact of Methodological Decisions around Imputation and the Choice of the Aggregation Unit." DIW discussion paper no. 562, Berlin, March.
- Kennickell, A. 1998. "Multiple Imputation in the Survey of Consumer Finances." Working paper, <http://www.federalreserve.gov/pubs/OSS/oss2/papers/impute98.pdf>
- Riphahn, R., and 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.
- Frick, J.R., Grabka, M.M. and Sierminska, E. 2010. Examining the gender wealth gap. In: *Oxford economic review* 62 , 669–690.
- 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* edited by D. Kasprzyk, G.J. Duncan, G. Kalton, and M.P. Singh, 400–425. New York: Wiley.
- Uhlig, N., Bryan, M., and S. Budd. 2012. "UKHLS Innovation Panel Household Wealth Questions: Preliminary Analysis." *Understanding Society Working Paper Series No. 2012 – 01*, January 2012. Available at: <https://www.iser.essex.ac.uk/publications/working-papers/understanding-society/2012-01> (accessed June 2014).
- 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.
- Westermeier, C. and Grabka, M. 2014. „Estimating the Impact of Alternative Multiple Imputation Methods on Longitudinal Wealth Data." Conference Paper, Available at: https://www.diw.de/documents/vortragsdokumente/220/diw_01.c.466448.de/v_2014_grabka_estimating_vfs.pdf (accessed August 2015)
- Westermeier, C. and Grabka, M. 2014. „Significant Statistical Uncertainty over Share of High Net Worth Households", *DIW Economic Bulletin*, 5, issue 14/15, p. 210–219.

Appendix A: Questionnaire 2002

Your personal assets and liabilities statement

Do you possess the following types of property or wealth?
If yes, then please estimate its current value.

(A) Are you personally the owner of the house or apartment in which you live?

Yes ☐ →

No ☐ ↓

Value:

If you were to sell today, how much would you receive
for your house/apartment including land?

EURO

Burden:

If you still have a loan taken out on your house/apartment,
how high is the remaining debt (excluding interest)?

EURO

Personal share of property:

Are you the sole owner (100%)
or co-owner (e.g. with your spouse)?

Sole Owner ☐

If the latter, how high is your own share?

Share in %

(B) Apart from the property you live in, do you possess another house or more land?

Yes ☐ →

No ☐ ↓

Type and number of properties:

What type of property is it?

One family house / Free hold flat (not used by yourself) ☐

Multiple family house / Apartment house ☐

Holiday home / Weekend home ☐

Undeveloped land ☐

Other property ☐

How many of this type of property do you have ?
(excluding the one used by yourself)

Amount

Value:

If you were to sell your property today
(excluding the one used by yourself),
how much would you receive?

EURO

Personal share of property:

Are you the sole owner (100%)
or co-owner (e.g. with your spouse)?

Sole Owner ☐

If the latter, how high is your own share?

Share in %

Burden:

If you still have a loan taken out on your property,
how high is the remaining debt
(excluding interest)?

EURO

(C) Do you possess financial assets over the value of 2500 EURO in the form of a savings balance, savings bonds, bonds, shares or investments?

Yes ☐ ➔
No ☐ ⬇

Value:

How high do you estimate the value of your financial assets?

EURO

Personal share of property:

Are these financial assets in your name or do they stretch over joint accounts, i.e. with your spouse?

Sole Owner ☐

If the latter, how high is your share?

Share in %

(D) Do you currently possess life insurance or a private pension plan or a buildings savings account?

Yes ☐ ➔
No ☐ ⬇

Value:

How high do you estimate the cash surrender value of these policies or financial assets to be?

EURO

(E) Are you the owner of a commercial enterprise, i.e. a company, a shop, an office, a practice or an agricultural enterprise, or are you involved in an enterprise such as the above forementioned?

Yes ☐ ➔
No ☐ ⬇

Personal share of property:

Are you the sole owner or co-owner of this enterprise, e.g. GBR, GmbH or KG?

Sole Owner ☐

Co-owner ☐

Value:

How high do you estimate the current value of your enterprise or of your share to be?
This is the price before tax, which you would receive at the sale of your enterprise or your share, taking into account any remaining financial burdens.

EURO

(F) Do you possess any tangible assets over 2,500 EURO (excluding motor vehicles) in the form of gold, jewellery, coins or valuable collections?

Yes ☐ ➔
No ☐ ⬇

Value:

If it were possible to estimate the value of these tangible assets: How high would the total value be?

EURO

**(G) Leaving aside any mortgages on house or property or house-building loan:
Do you at the present time have any debts relating to credit that you as a private individual have taken on at a bank or a similar institution or a another individual, for which you are accountable?
This is limited to debts greater than 2,500 EURO. This does not include mortgages or house-building loans!**

Yes ☐ ➔
No ☐ ⬇

Burden:

Current remaining debt (without interest):

EURO

**Question 86
next page!**

Appendix B: Questionnaire 2007

Do you possess the following types of property or wealth?
If yes, then please estimate its current value.

(A) Are you personally the owner of the house or apartment in which you live?

Yes ☐

No ☐

Value: If you were to sell today, how much would you receive for your house/apartment including land?	Euro	<input type="text"/>
Financial burden: If you still have a loan taken out on your house/apartment, how high is the remaining debt (excluding interest)?	Euro	<input type="text"/>
Personal share of property: Are you the sole owner (100%) or co-owner (e.g. with your spouse)?	Is paid off	<input type="checkbox"/>
	Sole Owner	<input type="checkbox"/>
If co-owner, how high is your own share?	Share in %	<input type="text"/>

(B) Apart from the property you live in, do you possess another house or more land?

Yes ☐

No ☐

Type and number of properties: What type of property is it?	
One family house / Free hold flat (not used by yourself)	<input type="checkbox"/>
Multiple family house / Apartment house	<input type="checkbox"/>
Holiday home / Weekend home	<input type="checkbox"/>
Undeveloped land	<input type="checkbox"/>
Other property	<input type="checkbox"/>
How many of this type of property do you have ? (excluding the one used by yourself)	Amount <input type="text"/>
Value: If you were to sell your property today (excluding the one used by yourself), how much would you receive?	Euro <input type="text"/>
Personal share of property: Are you the sole owner (100%) or co-owner (e.g. with your spouse)?	Sole Owner <input type="checkbox"/>
If co-owner, how high is your own share?	Share in % <input type="text"/>
Financial burden: If you still have a loan taken out on your property, how high is the remaining debt (excluding interest)?	Euro <input type="text"/>
	Is paid off <input type="checkbox"/>



(C) Do you personally have a savings contract with a building and loan association?

Yes ☐

No ☐



Value: What would you estimate your current building savings balance to be, including interest and dividends? If you have more than one such agreement, please add together and state the total sum!	Euro <input type="text"/>
--	---------------------------

- (D) Do you possess financial assets in the form of a savings balance, savings bonds, bonds, shares or investments?

Yes ☐ 
 No ☐ 



Value:
 How high do you estimate the value of your financial assets? Euro
 Personal share of property:
 Are these financial assets in your name or do they stretch over joint accounts, i.e. with your spouse? Sole Owner ☐
 If the latter, how high is your share? Share in %

- (E) Do you have a life insurance policy or private retirement insurance policy, purchased either by you or by your employer for you?

Yes ☐ 
 No ☐ 



Value:
 How high do you estimate the current cash surrender value of these policies to be? Euro

- (F) Are you the owner of a commercial enterprise, i.e. a company, a shop, an office, a practice or an agricultural enterprise, or are you involved in an enterprise such as the above forementioned?

Yes ☐ 
 No ☐ 



Personal share of property:
 Are you the sole owner or co-owner of this enterprise, e.g. GBR, GmbH or KG? Sole Owner ☐
 Co-owner ☐
 Value:
 How high do you estimate the current value of your enterprise or of your share to be?
 This is the price before tax, which you would receive at the sale of your enterprise or your share, taking into account any remaining financial burdens. Euro

- (G) Do you possess any tangible assets in the form of gold, jewellery, coins or valuable collections?

Yes ☐ 
 No ☐ 

Value:
 If it were possible to estimate the value of these tangible assets: How high would the total value be? Euro

- (H) Leaving aside any mortgages on house or property or house-building loan:
 Do you currently still own money on loans that you personally were granted by a bank, other organization, or private individual, and for which you personally are liable?
 Not including mortgages or building loans!

Yes ☐ 
 No ☐ 

Financial burden:
 How high are your outstanding debts? Euro

Question 127
 next page!