MANNHEIM RESEARCH INSTITUTE FOR THE ECONOMICS OF AGING

DOCUMENTATION OF THE LOGICAL IMPUTATION USING THE PANEL STRUCTURE OF THE 2003-2008 GERMAN SAVE SURVEY

Michael Ziegelmeyer

173-2009

L13, 17_D-68131 Mannheim_Phone +49 621 181-2773/1862_Fax +49 621 181-1863_www.mea.uni-mannheim.de

Documentation of the logical imputation using the panel structure of the 2003-2008 German SAVE Survey[1]

Michael Ziegelmeyer[†]

February 2009

Abstract

This paper documents the implementation of a logical imputation based on the panel structure of the 2003 to 2008 waves of the German SAVE dataset. A new release of the waves 2003-2008 will be available from June 2009. The concept and the principles of the underlying logical panel imputation are described. Furthermore, the method applied to logically impute each variable is briefly commented. The logical panel imputation of the SAVE dataset reduces decisively the number of missing values for some variables. In some cases more than 50% of all missing values can be replaced by proper values.

Key words: Item-nonresponse, imputation, panel data, SAVE

JEL classification: C81

[1] The author would like to thank Michela Coppola for very helpful support.

[†]Mannheim Research Institute for the Economics of Aging (MEA); University of Mannheim; L13,17; 68131 Mannheim; Germany; Email: Ziegelmeyer@mea.uni-mannheim.de

1 INTRODUCTION

The problem of item-nonresponse is widespread in micro datasets. Households or individuals, who are not able or willing to respond to questions, leave the resulting dataset similar to a "rag rug". Researchers who want to analyse such datasets have therefore to deal with serious difficulties. Mainly two problems arise: First, if multivariate procedures are used to analyze certain effects, all the variables of each unit (household or individual) must be complete. If there is one missing value in a variable, the variable has to be dropped or the sample size has to be reduced by all units containing missing values. This observed-case analysis can lead to a serious reduction of the sample size and the connected loss of efficiency. Additionally, the sample size varies with the question investigated, since different variables are needed for the analysis. Second, the missing value of a variable might not be random and related to certain characteristics or the environment of the respondent, so that estimations based on only observed cases might lead to biased results.

There are different methods to deal with item-nonresponse. Rässler and Riphahn (2006) outline four approaches (complete case analysis, weighting, imputation, model-based procedures) and discuss their strengths and weaknesses. The authors conclude *"that a multiple imputation procedure seems to be the best alternative at hand to account for missingness and to exploit all available information* (Rässler and Riphahn 2006, p. 229)*."* This procedure was chosen for the SAVE dataset from 2003 on. Each year was imputed separately using a *"Markov Chain Monte Carlo multiple imputation procedure"* to fill the missing values with plausible substitutes. For a detailed description see Schunk (2008).

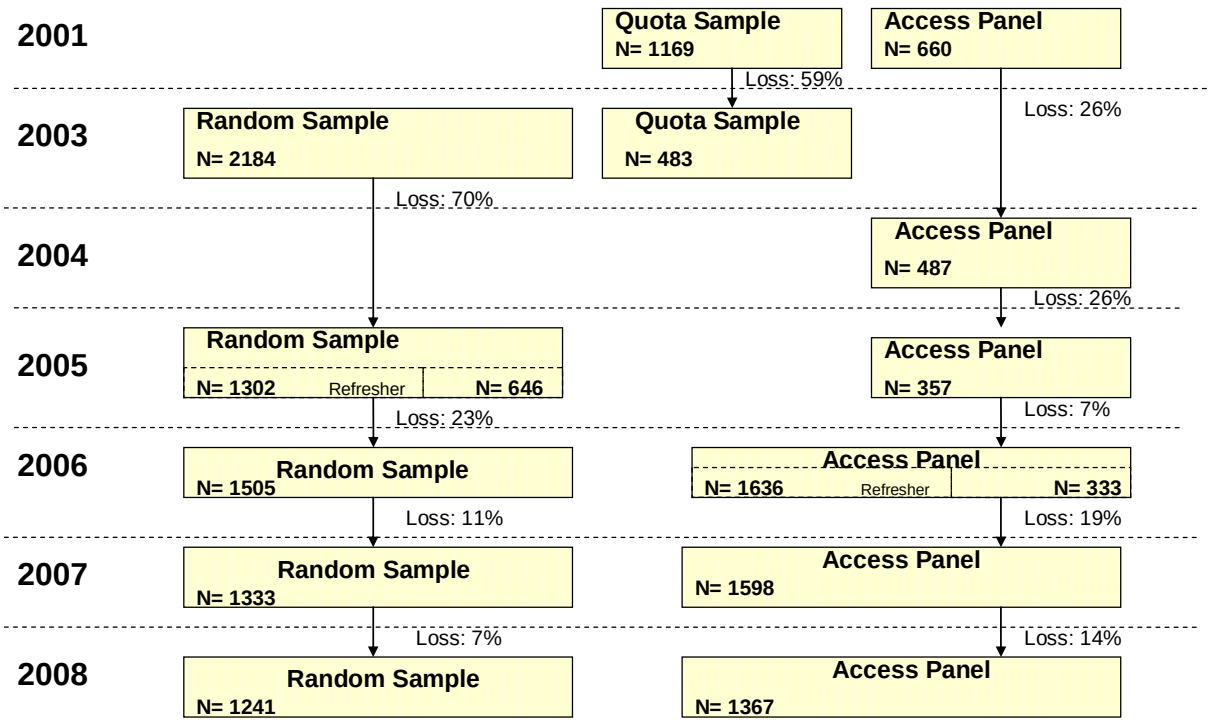
The first step of the complete imputation mechanism consists of a *logical imputation*. Logical imputation means that the true value of the missing value can be uniquely identified from within the dataset. The growing panel structure of the SAVE dataset offers new possibilities for logical imputation. This article documents the implementation of a logical imputation procedure for the SAVE data based on the waves from 2003 to 2008. The goal was to construct a transparent and traceable procedure, which allows the data user to evaluate the value added in the overall data accuracy. It should also demonstrate that data are not made-up.

The outline of this article is as follows: Section 2 gives a brief overview of the German SAVE Survey and its panel dimension. Section 3 explains the procedure using different examples and presents the principles of logical imputation adopted. Section 4 summarizes the implemented methods and presents the results and the achievements in term of data accuracy. Section 5 concludes and gives a perspective for the further improvement of the imputation methods of the SAVE dataset.

2 THE SAVE DATASET 2003-2008

The SAVE survey started in 2001. The first year was used to build up the optimal survey design for the following years. Since 2005 the survey has been repeated on a yearly basis (figure 1). The complete sample is split into two parts: a Random Route sample, which is a multiple stratified multistage random sample, and an Access Panel, which is a quota sample. SAVE was especially designed to better understand the various aspects of the saving behavior of German households. For a detailed description of scientific background, design, and results the reader is referred to Schunk (2006) and Börsch-Supan et al. (2008).

Figure 1: Sample design of SAVE



The key contributions of SAVE are the rich set of available control variables out of different areas like health, expectations, attitudes combined with detailed questioning about income, savings, debt and wealth. Moreover, SAVE is set up as a panel dataset and arrived at a fairly stable panel from 2006 on. Using a panel dataset, it is possible to distinguish between age and cohort effects, which is necessary for the empirical investigation of behavior over the life-cycle. The stable panel dimension allows for improving the data accuracy drastically. How this is done using the logical imputation is explained in the next section.

3 CONCEPT AND PRINCIPLES OF THE LOGICAL IMPUTATION

Logical imputation as the first step of a cross-sectional imputation procedure is a frequently applied technique, e.g. in the German Socio-Economic Panel (Frick & Grabka & Marcus, 2007) and the SCF in the US (Kennickell, 1991). Only a few datasets use panel imputation methods, e.g. the British Household Panel Survey (Buck et al., 2006), the German Socio-Economic Panel (Frick & Grabka, 2007) and the Household, Income and Labor Dynamics in Australia [Hilda] Survey (Starick, 2005): even these surveys, however, use panel estimation techniques only for specially chosen variables mainly out of the income section. In these cases, stochastic imputation procedures are generally applied. A logical longitudinal imputation was done in the case of the Canadian Survey of Labour and Income Dynamics (House, 2005). The questions about housing related content were imputed logically using the panel structure of the dataset if the postal code remained the same. In such cases it was assumed that the household did not move residence. The imputation method applied was a so-called “last value carried forward” method. For the SAVE dataset all sections were investigated for the application of a logical imputation using the panel structure, since the logical imputation allows replacing the missing values with a very high accuracy.

3.1 CONCEPT OF THE LOGICAL IMPUTATION

To provide the reader with a better understanding what a logical imputation procedure does, three examples are discussed: one example for the cross-sectional context and two for the panel context.
In the 2008 questionnaire, question 6 can be translated as follows:

"Do you live with a partner on a permanent basis? Yes or no."

In the survey 2008, this question presented 21 missing values. Before running complicated imputation procedures, it is probably worth to check first if these respondents, who did not answer question 6, reported somewhere else in the questionnaire useful information to fill in the missing value. Question 73, for example, provides, among others, the following response option: *"Does not apply, do not have a partner."* Choosing this option, the respondents implicitly report on their partner status, and the correct answer to question 6 can be reasonably derived from question 73. Indeed, using the information provided in this question, all the 21 missing values of question 6 can be filled up. In comparison with other imputation procedures, this way of proceeding has the advantage of relying only on the very mild assumption that the respondents consistently report the truth all over the questionnaire.

This way of filling up missing values in a cross-sectional context can be extended to a panel setting. In many cases, if the respondent provided the same answers in two years, but left a missing value in the year (or in the years) in between, it can be safely assumed that what he or she reported holds true also for the all years, so that the lacking data in one (or more) year(s) can be filled using the available answers. In some other questions, the structure of the possible answers allows to logically impute (at least part of) the missing values even when only a single observation is available. Two examples should help to clarify this concept.

As first example, a question out of the health section of the questionnaire is taken. All the respondents who report *not* to be currently smoking are asked the following question (number 30 of the 2008 questionnaire):

"Were you or your partner once a regular smoker? Yes or no."

If identical answers in at least two points in time are available, and given that the respondent reported not to smoke also in the year(s) between the two observations, it is quite safe to impute the missing value(s) between the two observed values by simply carrying on the

available answers. If in a certain year the respondent reported to have never regularly smoked in the past (so, if he or she answered "*No*" to the question), it is clear that this answer should be used to fill in the possible missings in the previous years (always conditioning on the current smoking behavior). Similarly, if in a certain year the respondent answered "*Yes*", the individual must be a regular smoker also in the future years (see table 1).

Table 1: Once a regular smoker – examples for logical imputation

	2003/04	**2005**	**2006**	**2007**	**2008**
Possible structure of the answers before the logical imputation:					
Are you currently smoking?					
Individual 1, 2, 3	*No*	*No*	*No*	*No*	*No*
Were you once a regular smoker?					
Individual 1	*Yes*	*Yes*	-	-	*Yes*
Individual 2	-	-	-	*No*	-
Individual 3	-	-	*Yes*	-	-
Structure of the answers after the logical imputation:					
Are you currently smoking?					
Individual 1, 2, 3	*No*	*No*	*No*	*No*	*No*
Were you once a regular smoker?					
Individual 1	*Yes*	*Yes*	***YES***	***YES***	*Yes*
Individual 2	***NO***	***NO***	***NO***	*No*	-
Individual 3	-	-	*Yes*	***YES***	***YES***

Note: Answers reported in capital, bold letters are meant to represent the logically imputed values. The dashes represent the missings.

The results for the household head and the corresponding partner are displayed in table 2. The overall missing rate can be reduced heavily. For the household head missing values can be reduced by around 45% and for the partner by around 44% over all years. In 2003 and 2004 there are no missing values available for the partner since this question was asked the first time in 2005.

Table 2: Once a regular smoker – result of logical imputation

	numbers of missing values in each year						
household head	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**sum**
before log. imputation	6	10	23	89	161	164	453
after log. imputation	6	5	8	34	71	124	248
information gain	0	5	15	55	90	40	**205**
partner							
before log. imputation	-	-	18	53	139	131	341
after log. imputation	-	-	9	28	59	96	192
information gain	-	-	9	25	80	35	**149**

The second example uses a question about the individual's unemployment history. The following question was asked in all years of the SAVE survey, (number 21 in the 2008 questionnaire):

> *"Have you or your partner ever been registered as unemployed at the German State labor agency? If yes, what was the longest continuous period for which you were unemployed?*
>
> *- Less than 1 month*
>
> *- Between 1 and 6 months*
>
> *- 6 to 12 months*
>
> *- 1 to 2 years*
>
> *- 2 years and more*
>
> *- No, I have never registered as unemployed."*

Again, the panel structure allows using the information given in one year to impute missing values in other years. If in a certain year, the respondent answered *"No, I have never been registered as unemployed"*, it can be reasonably argued that the same respondent should never have been registered as unemployed also in the years before: as in the example above, the answer can therefore be taken to impute possible missings in previous years (and it goes without saying that the information cannot be used to impute possible missings in *following* years). Similarly, if in a certain year the respondent answered *"2 years and more"* of registered unemployment, the highest category should be carried on to future years (but of course not on previous years!). Again, in case the respondent provided the same answer, e.g. *"6 to 12 months"*, in two years but left a missing value in the year (or years) in between, the lacking data can be filled with the available answer (see table 3).

Table 3: Registered unemployment – examples for logical imputation

	2003/04	**2005**	**2006**	**2007**	**2008**
Possible structure of the answers before the logical imputation:					
Individual 1	-	-	*Never unemployed*	*1 to 6 Months*	*1 to 6 Months*
Individual 2	*6 to 12 Months*	*1 to 2 Years*	-	*1 to 2 Years*	*1 to 2 Years*
Individual 3	*1 to 2 Years*	*2 or more Years*	*2 or more Years*	-	-

Structure of the answers after the logical imputation:					
Individual 1	***Never unemployed***	***Never unemployed***	*Never unemployed*	*1 to 6 Months*	*1 to 6 Months*
Individual 2	*6 to 12 Months*	*1 to 2 Years*	***1 to 2 Years***	*1 to 2 Years*	*1 to 2 Years*
Individual 3	*1 to 2 Years*	*2 or more Years*	*2 or more Years*	***2 or more Years***	***2 or more Years***

Note: Answers reported in capital, bold letters are meant to represent the logically imputed values. The dashes represent the missings.

Table 4: Registered unemployment – result of logical imputation

	numbers of missing values in each year						
household head	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**sum**
before log. imputation	10	11	28	81	6	2	138
after log. imputation	10	5	12	35	5	2	69
information gain	0	6	16	46	1	0	**69**
partner							
before log. imputation	46	9	44	71	12	10	192
after log. imputation	43	5	20	33	11	10	122
information gain	3	4	24	38	1	0	**70**

Table 4 displays the results of the logical imputation procedure outlined above. For the household head the overall number of missing values is halved. For the partner, the numbers of missing values are reduced by more than one third. Generally, the number of imputed missing values decreases with the years at the beginning (2003 and 2004) and the end (2008), and increases in the size of the panel component and the numbers of missings in each year.

The examples above illustrate the power of a proper logical imputation using the panel structure. The data quality can be improved drastically for some variables. Nevertheless, principles are needed to guide the implementation of the logical imputation to avoid the introduction of an excess of arbitrariness in the imputation procedure. These principles are discussed next.

3.2 PRINCIPLES OF THE LOGICAL IMPUTATION

Not all the questions in the survey are suitable for a logical imputation: questions about expectations, events during the last year, evaluation of current situations cannot be passively transferred across the years!

- The **first** step is therefore to identify the questions where a logical imputation using the panel structure can be implemented. For these questions, the logic of the

imputation has to be singled out: can a certain answer be transferred to future as well as to past waves or can it be used only in one temporal direction? Is it possible to logically impute the answers when only a single observation in time is available, or can we impute only those missings between two observed values? However, there are same questions, for which one cannot be absolutely sure that the imposed logic is true for all respondents. There might be exceptions one cannot control for with other variables. In such cases the alternative would be to use a stochastic imputation procedure such as a hierarchical hot-deck method. However, this stochastic imputation procedure, which is only based on the information available in each cross-section so far, increases the variance between the years and can thus bias estimation results based on panel estimation techniques. Thus, there is a trade-off between imputing this question logically and reducing the variance over the years or using a stochastic imputation procedure and increasing the variances sometimes drastically over the years. This is a "question to question" and "cases to case" consideration and involves a careful examination of the underlying data.

- In a **second** step, this logic has to be proved. In other words, we have to check if, among those who answered the question in all the years, the logic that we assumed is indeed obeyed. Back to the first example: we hypothesize that individuals who once report to have been smokers in the past, should then consistently report the same also in the future. Looking at the data we then should ask: is that really the case? As a matter of questionnaires, you will always find inconsistent answers (i.e. individual who report to have been smokers in the past and that a year later report to have never smoked). However, this number has to be *"reasonably"* small.
- The **third** step involves the implementation of the logical imputation. In the best case, one can uniquely identify the missing value with the information offered by the other years. To reduce the degree of arbitrariness to a minimum, the missing value is not logically imputed, if the answers of a certain household are inconsistent over the years. The reason is that one cannot decide which of the answers is the "true" one.

4 OVERVIEW OVER IMPLEMENTATION AND RESULTS

Since the logical imputation of each variable cannot be discussed in full length, table 5 summarizes the implementation in note form. The comments for each variable should give only a brief idea about the chosen procedure. The variable name corresponds to the variable name who is delivered with the datasets. Only the appendix *"_imp"* is missing since the variables refer to the still not imputed data. There are some exceptions mentioned in this table. These exceptions identify a possible violation of the assumed logic. Nevertheless, the logical imputation is done in these cases since the probability that these exception apply are found to be negligible small.[2] Moreover, there are cases in which the logic would have allowed logically imputing more values. However, if there are too many cases in which the observed data are in contrast to the imposed logic, the variables, for which these inconsistencies are observed in a serious way, are not logically imputed. The comments of table 5 do not discuss explicitly these cases. Table 6 displays the results.

The five multiple imputed SAVE datasets are always delivered with an indicator datasets. Before the logical panel imputation was done, each variable in the indicator dataset flagged with "1" implied a missing value and a variable flagged with "0" an observed value. After the logical panel imputation was done, the flag-dataset was updated: "0" indicates an observed value, "1" implies a stochastically imputed missing value and "2" a logically imputed value using the panel structure. This procedure allows the researcher identifying the missing values and the imputation procedure used.

[2] For a deeper investigation the do-files will be provided on request.

Table 5: Implementation

Var-Name	Label	Comment
B 2 Basic demographic information		
f06s	gender	No logical imputaion necessary.
f07o	year of birth	Logical imputation in every direction is done.
f08s	german nationality	If a respondent has the German nationality in a previous year, it is assumed that the respondent keeps the German nationality in the following years. In contrast foreigners can change their nationality to the German nationality.
f09s	marital status	Missing value = previous year = subsequent year.*
f11o	year of birth, partner	Missing value = previous year = subsequent year.* If only one year of birth is given over the complete panel period, the other years are set equal to the observed year only if no change in marital status occured (exception: a change of partners is possible maintaining the marital status).
f12s	Do you have any children?	Missing value = previous year = subsequent year.* If there are no children in the subsequent year and no change of the partner, there should be no children in previous years. If there are children in the previous year and no change of the partner, there should be children in subsequent years (exception: death of child).
f13o	number of children	Missing value = previous year = subsequent year.* Taking the information of the question "children yes or no" into account, the missing value of a previous year is equal to the number of children in the following year if both household head and his/ her partner are older than 50 years and there is no change in the maritual status (exception: adoption, child(ren) of the man 51 years or above born outside the partnership). In those few cases where the missing value is in one of the subsequent years, the missing values are imputed using the same schema (exception: death of child).
f15s	Do you have any grandchildren?	Missing value = previous year = subsequent year.* If there are no grandchildren in the following year, then there are no grandchildren in the previous year (exception: change of partner). If there are grandchildren in previous years, then there are grandchildren in the subsequent years (exception: death of grandchild).
f16o	number of grandchildren	Missing value = previous year = subsequent year.* Missing values are set to zero if the answer about "grandchildren yes or no" was no.
f20s1 f20s2	highest general school or college leaving certificate highest general school or college leaving certificate, partner	Missing value = previous year = subsequent year.* The missing in previous year is only logically imputed if the respondent (partner) has an elementary school leaving examination in the subsequent year.
f21s1 f21s2	complete professional training complete professional training, partner	Missing value = previous year = subsequent year.* If respondent (partner) is older than 35 years and there are at least two consistent answers about the highest qualification of a completed course of professional training, the missing values is filled with the qualification status of the following year. If there is no completed course of professional training in a subsequent year, there should be also no one the previous years.
f23s1 f23s2	background part-time employment background part-time employment, partner	Missing value = previous year = subsequent year.* If the age of the respondent (partner) is above 65 years and if the respondent (partner) is retired in all the following years, then the respondent (partner) should be retired in the previous year. If a respondent (partner) is retired in a previous year, he or she is retired in the subsequent years.
f24s1 f24s2	kind of employment kind of emloyment, partner	After considering the information of variable f23s1(f23s2), the missing value = previous year = subsequent year* for all cases excluding respondents (partners) who are currently not in paid employment. If the respondent (partner) is a civil servant the year before, then he or she is a civil servant in the subsequent years (exception: civil servants who change their employment status).
f25s1 f25s2	permanent or temporary position permanent or temporary position, partner	If wage earner or salaried employee, missing value = previous year = subsequent year.*
f26s1 f26s2	longest continuous period unemployed longest continuous period unemployed, partner	Missing value = previous year = subsequent year.* If the respondent (partner) was never unemployed in later years, he or she was never unemployed the years before. If the respondent (partner) was more than two years unemployed in previous years, he or she was more than two years unemployed in the subsequent years.
B 4 Health		
fg2s1 fg2s2	long-term health problems long-term health problems, partner	Considering only consistent answers, missing value = previous year = subsequent year* is imputed.
fg3m1_a fg3m1_b fg3m1_c fg3m1_d	heart disease high-blood pressure high cholesterol level stroke or circulatory problems affecting the brain	Missing value = previous year = subsequent year.* If the respondent does not have this long-term health problem, illness or disability in a subsequent year, he or she does not have this long-term health problem, illness or disability in the previous years. The importance to do this procedure for overall consistent answers should be emphasised since many inconsistency are found related to this question.
fg3m1_e fg3m1_f	chronic lung disease asthma	There is no logical imputaion possible, since it cannot be distinguished between chronic lung disease and asthma in subsequent years.
fg3m1_ef fg3m1_g fg3m1_h fg3m1_m fg3m1_l fg3m1_i	chronic lung disease, asthma cancer or malignant tumours excluding minor cases of skin cancer stomach ulcers, duodenal ulcer chronic backache mental illness other illnesses that are not listed here	Missing value = previous year = subsequent year.* If the respondent does not have this long-term health problem, illness or disability in a subsequent year, he or she does not have this long-term health problem, illness or disability in the previous years. The importance to do this procedure for overall consistent answers should be emphasised since many inconsistency are found related to this question.
fg3m1_j	none of the illnesses listed here	If the respondent does not have any long-term health problem, illness or disability, the missing value is one. If the respondent has at least one long-term health problem, illness or disability, the missing value is set to zero.

* The missing value is equal to the observed value of the previous year, if and only if the observed value of the previous year is equal to an observed value of a subsequent year.

fg3m2_a ... fg3m2_j	heart disease, partner none of the illnesses listed here, partner	The same procedure as for the respondent is applied to the partner.
f95s	regular smoker once?	Missing value = previous year = subsequent year.* If the respondent was a regular smoker once in a previous year, he or she was a regular smoker once in the subsequent years taking the actual smoking behavior into account. If the respondent was no regular smoker in a subsequent year, then he or she was no regular smoker in the previous years.
f94s	regular smoker?	Logical imputation uses only the information gained out of the question "once a regular smoker".
f95s2 f94s2	regular smoker once? - partner regular smoker? - partner	The same procedure as for the respondent is applied to the partner.
C 1 Savings		
fes1s	refused credit or not granted credit	Missing value = previous year = subsequent year (exception: five year horizon).* If the category "not applicable, I have never asked for credit" is chosen in a subsequent year, this status must also apply for the previous years.
fes2s	not applied for credit because of believing that it would be refused	Missing value = previous year = subsequent year (exception: five year horizon).*
f48s	record of household expenditure, parents	Missing value = previous year = subsequent year.*
C 3 Income		
f53m1_k f53m1_n	pension from the public retirement insurance system civil service pension	Missing value = previous year = subsequent year.* If the respondent gets this pension in a previous year, he or she gets this pension also in the subsequent years. If the respondent does not get this pension in a subsequent year, he or she does not get this pension in the previous years. This logical imputation was only done for consistent answers over the years (exception: orphan's pension).
f53m1_l f53m1_m f53m1_o f53m1_p f53m1_q f53m1_r f53m1_s	additional provision from civil service scheme company pension agricultural pension scheme occupational pension schemes pension deriving from a life insurance policy pension from private pension policies other pensions	Missing value = previous year = subsequent year.* If the respondent does not get this pension in a subsequent year, he or she does not get this pension in the previous years. This is only done for overall consistent answers.
f53m1_t	no, none of these - no independent income	If all types of current income sources are not paid, the missing value is one. If at least one type of retirement income is paid, the missing value is set to zero.
f53m2_k ... f53m2_t	pension from the public retirement insurance system, partner no, none of these - no independent income, partner	The same procedure as for the respondent is applied to the partner.
D 2 Old-age provision		
f60s	single retired?	If the respondent is not retired in a subsequent year, then the respondent is not retired in the previous years.
f60o	single retired, since?	Logical imputation in every direction possible considering consistent answer in all years the respondent answered this question.
f61s	couple retired?	If a couple is not retired in a subsequent year, then the couple is not retired in the previous years.
f61o1 f61o2 f61o3 f61o4	couple, both retired: since? - interviewed person couple, both retired: since? - partner couple, only interviewed person retired: since? couple, only partner retired: since?	Logical imputation in every direction possible considering consistent answer in all years the respondent answered this question.
f64m1_a f64m1_b f64m1_c f64m1_d f64m1_e f64m1_f f64m1_g f64m1_h f64m1_i	pension from the state pension insurance scheme additional provision from civil service scheme company pension civil service pension agricultural pension scheme occupational pension schemes for self-employed people pension deriving from a life insurance policy pension from private pension policies other pensions	Missing value = previous year = subsequent year.*
f64m1_j	none of these - no independent income	If all types of retirement income are not paid, the missing value is one. If at least one type of retirement income is paid, the missing value is set to zero.
f64m2_a ... f64m2_j	pension from the state pension insurance scheme - partner none of these - no independent income - partner	The same procedure as for the respondent is applied to the partner.
f72s_10	owner Rister- or Rürup-Pension	Variable frr1s (question 100 in the 2008 questionnaire) is used to logically impute this question. However, guestion 100 refers only to the Riester-Pension. Therefore, the imputation is additional conditioned on current profession. The logical imputation does not take place for self-employed or freelancer (exception: not only self-employed or freelancers could posses a Rürup-Pension).

* The missing value is equal to the observed value of the previous year, if and only if the observed value of the previous year is equal to an observed value of a subsequent year.

Table 6: Results

Var-Name	Label	missing values before log. Imputation 03	04	05	06	07	08	information gain abs	in %	missing values after log. Imputation 03	04	05	06	07	08
B 2 Basic demographic information															
f06s	gender	0	0	0	0	0	0	**0**	**0%**	0	0	0	0	0	0
f07o	year of birth	0	0	16	1	0	0	**11**	**65%**	0	0	6	0	0	0
f08s	german nationality	1	0	9	3	1	0	**10**	**71%**	1	0	0	2	1	0
f09s	marital status	3	0	2	11	23	21	**23**	**38%**	3	0	2	6	5	21
f11o	year of birth, partner	8	0	8	6	13	6	**14**	**34%**	7	0	6	3	5	6
f12s	Do you have any children?	4	1	9	58	62	22	**107**	**69%**	4	0	3	13	8	21
f13o	number of children	8	5	12	99	129	62	**178**	**57%**	8	0	3	30	42	54
f15s	Do you have any grandchildren?	4	1	13	94	125	68	**206**	**68%**	4	0	3	20	30	42
f16o	number of grandchildren	6	1	18	101	137	75	**155**	**46%**	6	0	7	35	60	75
f20s1	highest general school or college leaving certificate	11	0	16	8	3	3	**6**	**15%**	9	0	13	7	3	3
f20s2	highest general school or college leaving certificate, partner	42	1	14	15	7	7	**4**	**5%**	41	1	14	13	6	7
f21s1	complete professional training	3	3	153	9	1	0	**91**	**54%**	3	2	72	1	0	0
f21s2	complete professional training, partner	34	1	94	16	5	6	**45**	**29%**	34	1	52	13	5	6
f23s1	background part-time employment	41	15	68	143	84	59	**82**	**20%**	41	11	58	121	51	46
f23s2	background part-time employment, partner	31	29	64	198	92	66	**109**	**23%**	31	28	49	149	62	52
f24s1	kind of employment	5	7	65	146	68	69	**36**	**10%**	5	7	62	123	61	66
f24s2	kind of emloyment, partner	9	5	31	122	52	52	**28**	**10%**	9	5	28	106	43	52
f25s1	permanent or temporary position	7	8	80	199	88	88	**45**	**10%**	7	8	75	174	76	85
f25s2	permanent or temporary position, partner	11	6	37	138	67	67	**37**	**11%**	11	6	33	120	52	67
f26s1	longest continuous period unemployed	10	11	28	81	6	2	**69**	**50%**	10	5	12	35	5	2
f26s2	longest continuous period unemployed, partner	46	9	44	71	12	10	**70**	**36%**	43	5	20	33	11	10
B 4 Health															
fg2s1	long-term health problems	-	-	19	65	61	72	**45**	**21%**	-	-	19	53	28	72
fg2s2	long-term health problems, partner	-	-	21	53	47	56	**26**	**15%**	-	-	21	42	32	56
fg3m1_a	heart disease	-	-	53	374	197	256	**442**	**50%**	-	-	15	94	73	256
fg3m1_b	high-blood pressure	-	-	53	374	197	256	**422**	**48%**	-	-	19	102	81	256
fg3m1_c	high cholesterol level	-	-	53	374	197	256	**435**	**49%**	-	-	15	97	77	256
fg3m1_d	stroke or circulatory problems affecting the brain	-	-	53	374	197	256	**454**	**52%**	-	-	14	86	70	256
fg3m1_e	chronic lung disease	-	-	53	-	-	-	**0**	**0%**	-	-	53	-	-	-
fg3m1_f	asthma	-	-	53	-	-	-	**0**	**0%**	-	-	53	-	-	-
fg3m1_ef	chronic lung disease, asthma	-	-	-	374	197	256	**387**	**47%**	-	-	-	103	81	256
fg3m1_g	cancer or malignant tumours excluding minor cases of skin cancer	-	-	53	374	197	256	**454**	**52%**	-	-	14	86	70	256
fg3m1_h	stomach ulcers, duodenal ulcer	-	-	53	374	197	256	**454**	**52%**	-	-	14	85	71	256
fg3m1_m	chronic backache	-	-	-	374	197	256	**387**	**47%**	-	-	-	103	81	256
fg3m1_l	mental illness	-	-	-	374	197	256	**407**	**49%**	-	-	-	93	71	256
fg3m1_i	other illnesses that are not listed here	-	-	53	374	197	256	**380**	**43%**	-	-	25	129	90	256
fg3m1_j	none of the illnesses listed here	-	-	53	374	197	256	**347**	**39%**	-	-	38	138	101	256
fg3m2_a	heart disease, partner	-	-	919	290	184	206	**907**	**57%**	-	-	310	93	83	206
fg3m2_b	high-blood pressure, partner	-	-	919	290	184	206	**755**	**47%**	-	-	440	110	88	206
fg3m2_c	high cholesterol level, partner	-	-	919	290	184	206	**823**	**51%**	-	-	376	106	88	206
fg3m2_d	stroke or circulatory problems affecting the brain, partner	-	-	919	290	184	206	**931**	**58%**	-	-	288	91	83	206
fg3m2_e	chronic lung disease, partner	-	-	919	-	-	-	**0**	**0%**	-	-	919	-	-	-
fg3m2_f	asthma, partner	-	-	919	-	-	-	**0**	**0%**	-	-	919	-	-	-
fg3m2_ef	chronic lung disease, asthma, partner	-	-	-	290	184	206	**298**	**44%**	-	-	-	93	83	206
fg3m2_g	cancer or malignant tumours excluding minor cases of skin cancer, partner	-	-	919	290	184	206	**910**	**57%**	-	-	308	92	83	206
fg3m2_h	stomach ulcers, duodenal ulcer, partner	-	-	919	290	184	206	**922**	**58%**	-	-	295	93	83	206
fg3m2_m	chronic backache, partner	-	-	-	290	184	206	**271**	**40%**	-	-	-	105	98	206
fg3m2_l	mental illness, partner	-	-	-	290	184	206	**294**	**43%**	-	-	-	92	88	206
fg3m2_i	other illnesses that are not listed here, partner	-	-	919	290	184	206	**706**	**44%**	-	-	465	117	105	206
fg3m2_j	none of the illnesses listed here, partner	-	-	919	290	184	206	**673**	**42%**	-	-	511	106	103	206

f95s	regular smoker once?	6	10	23	89	161	164	**205**	**45%**	6	5	8	34	71	124
f95s2	regular smoker once? - partner	-	-	18	53	139	131	**149**	**44%**	-	-	9	28	59	96
f94s	regular smoker?	3	2	6	25	26	23	**41**	**48%**	3	1	4	9	12	15
f94s2	regular smoker? - partner	-	-	7	25	16	25	**34**	**47%**	-	-	6	9	10	14
C 1 Savings																
fes1s	refused credit or not granted credit	-	-	39	99	67	74	**95**	**34%**	-	-	25	48	37	74
fes2s	not applied for credit because of believing that it would be refused	-	-	45	131	131	98	**120**	**30%**	-	-	45	78	64	98
f48s	record of household expenditure, parents	121	3	69	98	99	83	**73**	**15%**	121	3	59	71	63	83
C 3 Income																
f53m1_k	pension from the public retirement insurance system	120	5	56	52	43	38	**123**	**39%**	101	3	23	18	14	32
f53m1_l	additional provision from civil service scheme	120	5	56	52	43	38	**132**	**42%**	98	2	16	13	15	38
f53m1_m	company pension	120	5	56	52	43	38	**134**	**43%**	97	2	14	13	16	38
f53m1_n	civil service pension	120	5	56	52	43	38	**138**	**44%**	95	2	14	13	15	37
f53m1_o	agricultural pension scheme	120	5	56	52	43	38	**136**	**43%**	96	2	14	13	15	38
f53m1_p	occupational pension schemes	120	5	56	52	43	38	**137**	**44%**	95	2	14	13	15	38
f53m1_q	pension deriving from a life insurance policy	120	5	56	52	43	38	**137**	**44%**	95	2	14	13	15	38
f53m1_r	pension from private pension policies	120	5	56	52	43	38	**136**	**43%**	96	2	14	13	15	38
f53m1_s	other pensions	120	5	56	52	43	38	**121**	**39%**	101	2	18	17	17	38
f53m1_t	no, none of these - no independent income	120	5	56	52	43	38	**41**	**13%**	120	0	46	46	30	31
f53m2_k	pension from the public retirement insurance system, partner	213	28	94	81	55	61	**176**	**33%**	176	17	41	40	27	55
f53m2_l	additional provision from civil service scheme, partner	213	28	94	81	55	61	**200**	**38%**	168	12	36	33	22	61
f53m2_m	company pension, partner	213	28	94	81	55	61	**196**	**37%**	169	13	36	34	23	61
f53m2_n	civil service pension, partner	213	28	94	81	55	61	**201**	**38%**	168	12	35	33	22	61
f53m2_o	agricultural pension scheme, partner	213	28	94	81	55	61	**203**	**38%**	166	12	35	33	22	61
f53m2_p	occupational pension schemes, partner	213	28	94	81	55	61	**203**	**38%**	166	12	35	33	22	61
f53m2_q	pension deriving from a life insurance policy, partner	213	28	94	81	55	61	**202**	**38%**	166	12	36	33	22	61
f53m2_r	pension from private pension policies, partner	213	28	94	81	55	61	**199**	**37%**	167	14	36	33	22	61
f53m2_s	other pensions, partner	213	28	94	81	55	61	**184**	**35%**	176	14	38	37	22	61
f53m2_t	no, none of these - no independent income, partner	213	0	94	81	55	61	**22**	**4%**	213	0	87	77	50	55
D 2 Old-age provision																
f60s	single retired?	42	4	0	0	0	0	**3**	**7%**	41	2	0	0	0	0
f60o	single retired, since?	-	-	-	216	143	90	**159**	**35%**	-	-	-	133	84	73
f61s	couple retired?	641	81	0	9	0	0	**119**	**16%**	566	40	0	6	0	0
f61o1	couple, both retired: since? - interviewed person	-	-	-	219	110	45	**155**	**41%**	-	-	-	127	52	40
f61o2	couple, both retired: since? - partner	-	-	-	230	111	63	**150**	**37%**	-	-	-	135	62	57
f61o3	couple, only interviewed person retired: since?	-	-	-	219	122	83	**109**	**26%**	-	-	-	155	85	75
f61o4	couple, only partner retired: since?	-	-	-	95	49	39	**42**	**23%**	-	-	-	70	34	37
f64m1_a	pension from the state pension insurance scheme	265	52	81	210	213	177	**85**	**9%**	265	52	71	178	170	177
f64m1_b	additional provision from civil service scheme	289	52	81	210	213	177	**95**	**9%**	289	52	72	173	164	177
f64m1_c	company pension	287	52	81	210	213	177	**94**	**9%**	287	52	73	172	165	177
f64m1_d	civil service pension	288	52	81	210	213	177	**102**	**10%**	288	52	70	171	161	177
f64m1_e	agricultural pension scheme	291	52	81	210	213	177	**102**	**10%**	291	52	70	171	161	177
f64m1_f	occupational pension schemes for self-employed people	293	52	81	210	213	177	**100**	**10%**	293	52	70	172	162	177
f64m1_g	pension deriving from a life insurance policy	288	52	81	210	213	177	**85**	**8%**	288	52	72	177	170	177
f64m1_h	pension from private pension policies	290	52	81	210	213	177	**84**	**8%**	290	52	73	177	170	177
f64m1_i	other pensions	293	52	81	210	213	177	**87**	**8%**	293	52	72	175	170	177
f64m1_j	none of these - no independent income	291	52	81	210	213	177	**83**	**8%**	291	52	70	180	171	177
f64m2_a	pension from the state pension insurance scheme - partner	274	30	143	122	168	125	**78**	**9%**	274	30	131	102	122	125
f64m2_b	additional provision from civil service scheme - partner	312	30	143	122	168	125	**78**	**9%**	312	30	130	104	121	125
f64m2_c	company pension - partner	311	30	143	122	168	125	**76**	**8%**	311	30	130	102	125	125
f64m2_d	civil service pension - partner	310	30	143	122	168	125	**82**	**9%**	310	30	130	101	120	125
f64m2_e	agricultural pension scheme - partner	315	30	143	122	168	125	**83**	**9%**	315	30	130	101	119	125
f64m2_f	occupational pension schemes for self-employed people - partner	315	30	143	122	168	125	**83**	**9%**	315	30	130	101	119	125
f64m2_g	pension deriving from a life insurance policy - partner	310	30	143	122	168	125	**69**	**8%**	310	30	131	109	124	125
f64m2_h	pension from private pension policies - partner	314	30	143	122	168	125	**70**	**8%**	314	30	133	107	123	125
f64m2_i	other pensions - partner	313	30	143	122	168	125	**74**	**8%**	313	30	132	103	124	125
f64m2_j	none of these - no independent income - partner	311	30	143	122	168	125	**77**	**9%**	311	30	131	102	123	125
f72s_10	owner Rister-pension	339	0	289	593	333	29	**818**	**52%**	295	0	138	230	73	29

5 CONCLUSION

This paper documented the implementation of a logical panel imputation of the 2003 to 2008 waves of the German SAVE study. After briefly introducing the SAVE dataset, the concept of logical imputation was clarified using different examples. The principles of the underlying logical panel imputation were discussed. Compared to stochastic imputation procedures, a great advantage of the logical panel imputation is the mild assumption that the respondent consistently reports the truth over all the years. After giving a short overview of all the variables logically imputed using the panel structure and a comment about the chosen method, table 6 showed how many missing values could be filled. For remarkably many variables the number of missing values could be reduced by more than 50%. Thus, the applied logical panel imputation improved the quality of the SAVE data notably.

Using the panel dimension of the SAVE dataset can be seen as a first step towards a complete multiple panel imputation procedure. So far the logical imputation in each cross-section and the subsequently logical imputation over the waves 2003-2008 was the starting position for the "Markov Chain Monte Carlo multiple imputation procedure" in each cross-section. This procedure has been improved over the years and was recently standardized. Now, all the datasets from 2003 to 2008 are based on the same imputation procedure, which allows a consistent treatment of all waves using panel estimation techniques.[3] A challenging and work intensive improvement would be a multiple panel imputation. This would not only allow increasing the accuracy of the estimations but also preserving the correlation structure over the years.

[3] The consistently imputed datasets will be available around July 2009. For more information please have a look at the MEA homepage: http://www.mea.uni-mannheim.de/

6 LITERATURE

Börsch-Supan, A., M. Coppola, L. Essig, A. Eymann, and D. Schunk (2008): *"The German SAVE study. Design and Results."* Mea Studies 06, MEA-Mannheim Research Institute for the Economics of Aging, University of Mannheim.

Buck, N. et al. (2006): *"Quality Profile: British Household Panel Survey. Version 2.0: Waves 1 to 13: 1991-2003."* Institute for Social and Economic Research, University of Essex.

Frick, J. R. and Grabka, M. M. (2007): *"Item non-response and Imputation of Annual Labor Income in Panel Surveys from a Cross-National Perspective."* DIW Discussion Paper 736.

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

House, G. (2005): *"General Housing Imputation (excluding utilities) in the Survey of Labour and Income Dynamics (SLID)."* Income Research Paper Series, Statistics Canada, Catalogue no. 75F0002MIE — No. 010.

Kennickell, A. B. (1991): *"Imputation of the 1989 Survey of Consumer Finances: Stochastic Relaxation and Multiple Imputation."* Proceedings of the Section on Survey Research Methods, American Statistical Association. Atlanta, Georgia.

Rässler, S. and R. Riphahn (2006): *"Survey item nonresponse and its treatment."* Allgemeines Statistisches Archiv, 90, 217 – 232.

Schunk, D. (2006): *"The German SAVE Survey 2001 - 2006. Documentation and Methodology."* Mea-Discussion-Paper 109-2006, MEA-Mannheim Research Institute for the Economics of Aging, University of Mannheim.

Schunk, D. (2008): *"A Markov chain Monte Carlo algorithm for multiple imputation in large surveys."* Advances in Statistical Analysis, 92(1), 101 - 114.

Starick, R. (2005): "*Imputation in Longitudinal Surveys: The Case of HILDA."* Research Paper of the Australian Bureau of Statistics. ABS Catalogue no. 1352.0.55.075.

Discussion Paper Series

Mannheim Research Institute for the Economics of Aging Universität Mannheim

To order copies, please direct your request to the author of the title in question.

Nr.	Autoren	Titel	Jahr
161-08	Karsten Hank	Generationenbeziehungen im alternden Europa	08
162-08	Axel Börsch-Supan, Karsten Hank, Hendrik Jürges, Mathis Schröder	Longitudinal Data Collection in Continental Europe: Experiences from the Survey of Health, Ageing and Retirement in (SHARE)	08
163-08	Martin Salm	Job loss does not cause ill health	08
164-08	Martin Salm, Daniel Schunk	The role of childhood health for the inter-generational transmission of human capital: Evidence from administrative data	08
165-08	Christina Benita Wilke	On the feasibility of notional defined contribution systems: The German case	08
166-08	Alexander Ludwig Michael Reiter	Sharing Demographic Risk – Who is Afraid of the Baby Bust?	08
167-08	Jürgen Maurer André Meier	Smooth it Like the “Joneses?” Estimating Peer-Group Effects in Intertemporal Consumption Choice	08
168-08	Melanie Lührmann Jürgen Masurer	Who wears the trousers? A semiparametric analysis of decision power in couples	08
170-08	Jürgen Maurer	Who has a clue to preventing the flu? Unravelling supply and demand effects on the take-up of influenza vaccinations	08
171-08	Johannes Binswanger Daniel Schunk	What Is an Adequate Standard of Living during Retirement?	08
172-08	Mathis Schröder Axel Börsch-Supan	Retrospective Data Collection in Europe	08
173-09	Michael Ziegelmeyer	Documentation of the logical imputation using the panel structure of the 2003-2008 German SAVE Survey	09