

Marcus, Jan; Siegers, Rainer; Grabka, Markus M.

**Research Report**

## SOEP 2010 - Preparation of data from the New SOEP consumption module: Editing, imputation, and smoothing

SOEP Survey Papers, No. 145

**Provided in Cooperation with:**

German Institute for Economic Research (DIW Berlin)

*Suggested Citation:* Marcus, Jan; Siegers, Rainer; Grabka, Markus M. (2013) : SOEP 2010 - Preparation of data from the New SOEP consumption module: Editing, imputation, and smoothing, SOEP Survey Papers, No. 145, Deutsches Institut für Wirtschaftsforschung (DIW), Berlin

This Version is available at:

<https://hdl.handle.net/10419/85278>

**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 Documentations

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

145-2013

# SOEP 2010 – Preparation of data from the new SOEP consumption module: Editing, imputation, and smoothing

Jan Marcus, Rainer Siegers, Markus M. Grabka

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 Documentations (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/soepsurveypapers>

**Editors:**

Prof. Dr. Gert G. Wagner, DIW Berlin and Technische Universität Berlin

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

Please cite this paper as follows:

Jan Marcus, Rainer Siegers, Markus M. Grabka. 2013. SOEP 2010 – Preparation of data from the new SOEP consumption module: Editing, imputation, and smoothing. SOEP Survey Papers 145: Series C. Berlin: DIW/SOEP

ISSN: 2193-5580 (online)

Contact: DIW Berlin  
SOEP  
Mohrenstr. 58  
10117 Berlin

Email: [soeppapers@diw.de](mailto:soeppapers@diw.de)

Jan Marcus<sup>1</sup>

Rainer Siegers<sup>2</sup>

Markus M. Grabka<sup>3</sup>

## **SOEP 2010 – Preparation of data from the new SOEP consumption module Editing, imputation, and smoothing**

Abstract:

This documentation describes the data preparation of the new consumption module in the German Socio-Economic Panel Study (SOEP) and introduces the content and structure of the generated dataset “hconsum.” In 2010, the SOEP for the first time included a detailed consumption module in the household questionnaire. This documentation discusses several methodological challenges of the new module and suggests possible remedies. The methodological challenges include inconsistencies between monthly and annual consumption information, missing values, and a high incidence of heaping.

**JEL-Codes:** C81, D30, E21

**Keywords:** SOEP, consumption, heaping, imputation, generalized beta of the second kind

<sup>1</sup> Corresponding author: Jan Marcus, DIW Berlin, Mohrenstrasse 58, D-10117 Berlin, Germany, Tel.: +49-30-89789-308, email: [jmarcus@diw.de](mailto:jmarcus@diw.de).

<sup>2</sup> Rainer Siegers, DIW Berlin, Mohrenstrasse 58, D-10117 Berlin, Germany, Tel.: +49-30-89789-239, email: [rsiegers@diw.de](mailto:rsiegers@diw.de).

<sup>3</sup> Markus M. Grabka, DIW Berlin & TU Berlin, Mohrenstrasse 58, D-10117 Berlin, Germany, Tel.: +49-30-89789-339, email: [mgrabka@diw.de](mailto:mgrabka@diw.de).



## 1. Introduction

Welfare analyses generally rely on information about disposable income. Such information is used, for instance, in the most common measures of inequality at the national level. However, various researchers have suggested that consumption is superior to income in describing well-being and welfare (e.g., Ringen 1988; Crossley and Pendakur 2002). Whereas consumption characterizes the *actual* current living standard, income describes the “*potential* command over resources” (Headey 2008: 24). A person with low income but high consumption expenditures should, following this argument, not be regarded as poor.

The analysis of consumption has a long tradition in both economics and the social sciences. The life-cycle hypothesis described by Modigliani (1966) was based on consumption data. Hall (1978) also made use of consumption data in his permanent income hypothesis. And even the precautionary saving motive (e.g., Leland 1968) is based on consumption data. Finally, the recently published report by the Stiglitz-Sen-Fitoussi Commission (2009) also recommends treating well-being as the result of both income and consumption and thus reinforces the importance of consumption data.

Despite the importance of consumption information, few data sets are available that include detailed information about consumption expenditures due primarily to methodological challenges in the collection of detailed consumption data (see Section 2). In order to improve the data basis for empirical consumption research in Germany, in 2010 the German Socio-Economic Panel Study (SOEP) for the first time included a detailed consumption module. This documentation describes the preparation of data from the new SOEP consumption module and introduces the content and structure of the generated dataset “hconsum.” We were faced with three methodological challenges in generating the final consumption data. Firstly, due to the design of the consumption module, inconsistent answers arose between the monthly and annual amounts spent for consumption. Secondly, we encountered the well-known phenomenon of missing data, here in particular item non-response. And thirdly, consumption data are usually blurred by heaping. For researchers who do not want their consumption variables to include changes from all steps of data preparation, the new data set “hconsum” contains not only the prepared consumption variables but also flag variables providing researchers the opportunity to select individual solutions.

The structure of this documentation is as follows. Section 2 introduces the SOEP 2010 consumption module and other consumption items in the general SOEP questionnaire. Section 3 describes the relevance and incidence of the different methodological challenges, and Section 4 presents how these problems are dealt with. Section 5 compares the consumption data with data from official statistics, and Section 6 presents content and structure of the new generated dataset “hconsum.”

## 2. The SOEP 2010 consumption module

In the past, consumption data was usually collected by national statistical offices such as the Federal Statistical Office in Germany on a cross-sectional basis. A typical example is the Income and Expenditure Survey (EVS) of the Federal Statistical Office in Germany (see Becker et al. 2002). In the EVS, survey participants have to keep detailed household accounts of *all* earnings and expenditures. This book has to be completed in case of the most recent EVS 2008 for a period of three months. In addition, respondents have to keep an even more detailed account of household expenditures on food, beverages, and tobacco products over a one-month period. One of the main challenges for these kinds of consumption surveys is to find a sufficient number of representative participants who are willing to accept the burden of keeping detailed accounts over several months.

Given that detailed accounts of household expenditures seem too burdensome and time-consuming for household (panel) surveys, it is usually only administrative surveys that use this interview mode. To date, other household panel surveys such as the Panel Study of Income Dynamics (PSID), and the British Household Panel Survey (BHPS) have only asked about a small subset of key consumption goods (such as rental costs). However, Browning et al. (2003) proposed an alternative to a detailed shopping diary. They suggested asking about a relevant subset of consumption goods to approximate total consumption. The aim of this subset is to analyze typical consumption in contrast to consumption at a given point in time as is typically the case in administrative surveys. This can be seen as an advantage over a diary, where irregular purchases may bias estimates. The Household, Income and Labour Dynamics in Australia (HILDA) Survey adopted this module and showed that with a short battery of consumption goods, it was able to capture 53% of the total consumption reflected in the Australian expenditure survey (HES) (Headey 2008). In 2010, the German Socio-economic Panel Study (SOEP) also adopted the idea of using a subset of consumption goods to approximate total consumption. In total, 16 consumption categories were asked at the household level (see Figure 1). We refer to these 16 items as the core consumption module. The questionnaire structure of all 16 core consumption items is the same: First, a filter question asks whether or not the household had any expenditures in the previous year. Afterwards, respondents are asked the exact amount — for each item, they can choose between answering on a monthly basis or an annual basis.

Figure 1: Consumption module in the SOEP-2010 household questionnaire

**In the following you see a list of possible expenditures that we have not asked about so far in this survey.**

**Did you or another household member make any of the following expenditures?**

**If yes: how much in total did these expenditures cost your household in 2009?**

*You can state the expenditures for 2009 either as monthly averages or as total yearly expenditures!*

		Expenditures in 2009:			
				per month	per year
<b>1</b>	Food, groceries at home	Yes ..... <input type="checkbox"/> ⇒	<input type="text"/>	or	<input type="text"/> euros
		No ..... <input type="checkbox"/>			
<b>2</b>	Food / drinks outside the home	Yes ..... <input type="checkbox"/> ⇒	<input type="text"/>	or	<input type="text"/> euros
		No ..... <input type="checkbox"/>			
<b>3</b>	Clothing / shoes	Yes ..... <input type="checkbox"/> ⇒	<input type="text"/>	or	<input type="text"/> euros
		No ..... <input type="checkbox"/>			
<b>4</b>	Body care / cosmetics / hairdresser	Yes ..... <input type="checkbox"/> ⇒	<input type="text"/>	or	<input type="text"/> euros
		No ..... <input type="checkbox"/>			
<b>5</b>	Health (e.g., medicines, courses, consultation fee)	Yes ..... <input type="checkbox"/> ⇒	<input type="text"/>	or	<input type="text"/> euros
		No ..... <input type="checkbox"/>			
<b>6</b>	Telecommunication (landline, cellular phone, Internet)	Yes ..... <input type="checkbox"/> ⇒	<input type="text"/>	or	<input type="text"/> euros
		No ..... <input type="checkbox"/>			
<b>7</b>	Education / further training	Yes ..... <input type="checkbox"/> ⇒	<input type="text"/>	or	<input type="text"/> euros
		No ..... <input type="checkbox"/>			
<b>8</b>	Culture (theater, cinema, concerts, museums, exhibitions)	Yes ..... <input type="checkbox"/> ⇒	<input type="text"/>	or	<input type="text"/> euros
		No ..... <input type="checkbox"/>			
<b>9</b>	Leisure activities, hobbies, sports, yard and garden, animals	Yes ..... <input type="checkbox"/> ⇒	<input type="text"/>	or	<input type="text"/> euros
		No ..... <input type="checkbox"/>			
<b>10</b>	Vacation trips including short holidays	Yes ..... <input type="checkbox"/> ⇒	<input type="text"/>	or	<input type="text"/> euros
		No ..... <input type="checkbox"/>			
<b>11</b>	Life insurance, private pension insurance	Yes ..... <input type="checkbox"/> ⇒	<input type="text"/>	or	<input type="text"/> euros
		No ..... <input type="checkbox"/>			
<b>12</b>	Other insurance policies (e.g. car, legal, household goods)	Yes ..... <input type="checkbox"/> ⇒	<input type="text"/>	or	<input type="text"/> euros
		No ..... <input type="checkbox"/>			
<b>13</b>	Motor vehicle repairs (including motorcycle)	Yes ..... <input type="checkbox"/> ⇒	<input type="text"/>	or	<input type="text"/> euros
		No ..... <input type="checkbox"/>			
<b>14</b>	Transport (car, train, bus, etc.)	Yes ..... <input type="checkbox"/> ⇒	<input type="text"/>	or	<input type="text"/> euros
		No ..... <input type="checkbox"/>			
<b>15</b>	Furniture, household appliances not mentioned previously	Yes ..... <input type="checkbox"/> ⇒	<input type="text"/>	or	<input type="text"/> euros
		No ..... <input type="checkbox"/>			
<b>16</b>	Other expenditures	Yes ..... <input type="checkbox"/> ⇒	<input type="text"/>	or	<input type="text"/> euros
		No ..... <input type="checkbox"/>			
		Other expenditures please specify:			



In addition to the 16 core consumption items that were collected for the first time in 2010, there are a couple of selected questions about regular expenditures in the SOEP that were also asked before on a regular basis. These include information about loan repayment, financial support of relatives, expenditures for household and cleaning help, and expenditures for child care. For households in rental housing, we also include costs of heating and hot water, electricity, and gross rent including utilities. For households in owner-occupied housing, we include the costs of heating, electricity, regular expenses (water, garbage removal, sidewalk cleaning, etc.), mortgage, and other housing costs (see Figure A.1 in the appendix for an overview of the questionnaire wordings for these other consumption items).

While the core consumption module asked in 2010 is affected by inconsistencies, non-response, and heaping effects, the additional information about regular expenses is affected only by the problem of non-response. Thus, the latter information is only discussed in Chapter 4.2.

## **3 Methodological challenges**

### **3.1 Inconsistencies**

For each consumption category in the core consumption module of the SOEP, household heads can decide whether to state the average amount spent on an annual or monthly basis. However, particularly in the self-administered PAPI (paper and pencil interviews) without an interviewer, some household heads provide both annual and monthly information. These data do not match in all cases, i.e., for some households that state both amounts, average monthly consumption does not equal one twelfth of average annual consumption. Since only one value for each category will be provided, there is the need to resolve that potential inconsistency.

Table 1 provides an overview of the occurrence of these kinds of inconsistencies for the 2010 consumption module.<sup>1</sup> The table shows that the number of households that provide double information, i.e., monthly and yearly information, is higher for the first consumption categories and lower for subsequent categories, suggesting learning effects of the respondents. The share of households with inconsistent information ranges between 0.35% and 0.8% of all households with

---

<sup>1</sup> Such inconsistencies can only arise for the core 16 consumption categories. For other consumption expenditures, which are asked every year, there is only one response option.

valid positive consumption expenditures in the respective category, and is, hence, clearly less than 1% in all 16 components.

Table 1: Overview over the occurrence of inconsistent monthly and yearly information

Consumption category	Observations with double information	Share of double information (%)	Observations with inconsistent information	Share of inconsistent information (%)
1	552	5.21	83	0.78
2	333	4.05	63	0.77
3	269	2.68	70	0.70
4	342	3.29	63	0.61
5	248	2.58	77	0.80
6	369	3.54	61	0.59
7	57	2.01	10	0.35
8	145	2.26	29	0.45
9	212	2.49	46	0.54
10	86	1.25	31	0.45
11	169	2.83	29	0.49
12	174	1.80	61	0.63
13	72	1.28	21	0.37
14	187	2.73	49	0.72
15	35	1.35	10	0.39
16	50	2.29	10	0.46

Source: SOEP v28

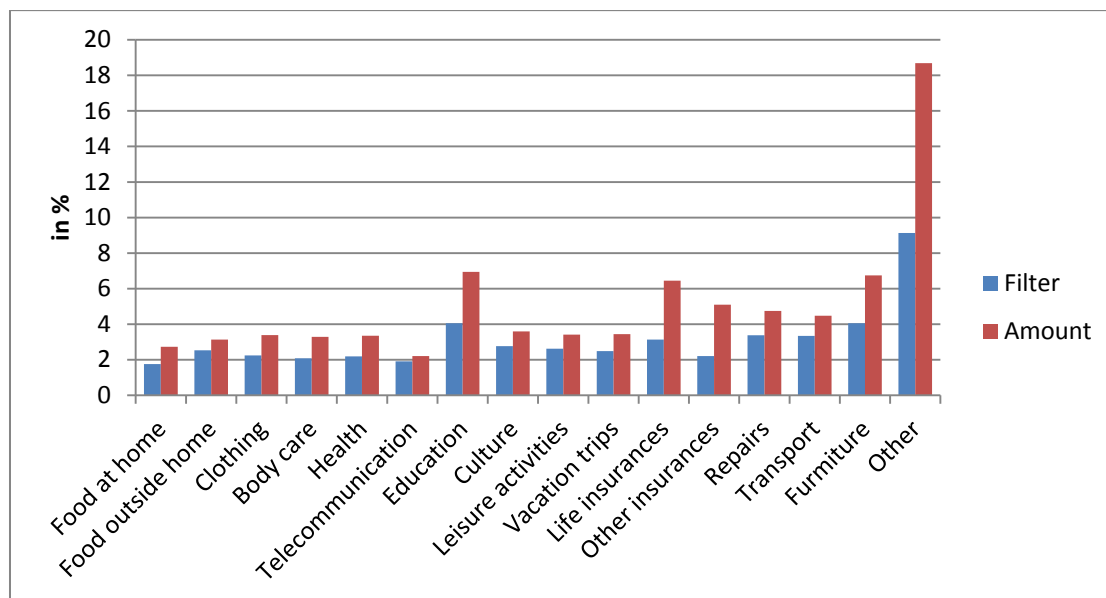
### 3.2 Missing values

Similar to many other variables in non-compulsory surveys, e.g., income and wealth variables, the present data on consumption expenditures also suffers from missing values. These missing values arise when households do not provide information on one or more consumption categories — either because the respondents do not know the information or because they do not want to provide it. In the best case, these values are missing completely at random (MCAR, see Little and Rubin 2002). However, even in this favorable case, dropping these observations would result in less efficient estimators. If the values are not missing completely at random, the estimators will even be biased.

Figure 2 provides an overview over the incidence of item non-response for the different consumption categories. In the core consumption module, around 2-4% of the households do not provide

information on whether or not they have expenditures in the given consumption category (“filter question”).

Figure 2: Share of item-non response on questions about filter and amount – SOEP 2010



Source: SOEPv28, total population (n=10,840)

The last category (other expenditures), is an exception with a much higher incidence of about 9%. Another 3-7% of the households that answer the filter question in the affirmative do not provide the detailed amount of expenditures. Again, the incidence of missing values is much higher for the “other expenditure” category, at nearly 20%. Compared to income and wealth-related questions, the degree of missingness in this consumption module is rather small (see Frick and Grabka 2005).

### 3.3 Heaping

When collecting retrospective data, surveys are confronted with the problem that respondents do not always remember the requested information accurately. It is well known that, for example, in duration analysis, interviewees tend to round certain durations to the nearest year, half year, or month (e.g., Heitjan and Rubin 1990, Kraus und Steiner 1998, El Messlaki 2010). Heaping also occurs in other research areas, such as when one ask former smokers to report when they quit (Lillard et al. 2008), when asking for the number of cigarettes smoked (Wang and Heitjan 2008), or even when asking mothers for the age of their children (Heitjan and Rubin 1990).

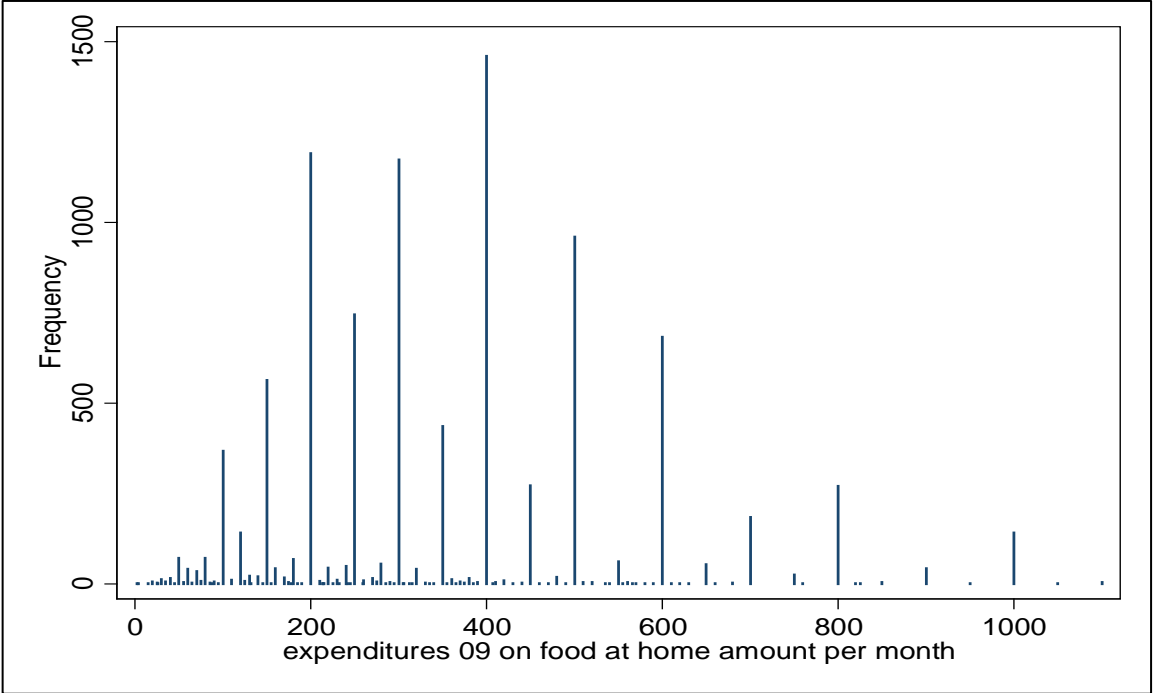
Consumption data is also confronted with the phenomenon of heaping (e.g., Battistin et al. 2003). When asked how much the household spends on specific consumption goods, household heads often round to multiples of 5, 10, or 100.<sup>2</sup> This rounding produces a distribution with distinct heaping points. Heaping and rounding are strongly interrelated, but they are not the same. Not every rounded value marks a heaping point, as heaping only occurs if many observations report the same value. Similarly, not every heaped value is necessarily a rounded value, as heaping might occur “naturally.” For instance, the distribution of the length of unemployment spells might have particular heaping points according to the maximum duration of unemployment benefits. However, such “natural” heaping points are rather unlikely to occur in the consumption data considered here.

The present consumption data in the SOEP are strongly affected by the heaping phenomenon, as can be seen in Figure 3 for food expenditures at home. The figure shows that multiples of 50 and 100 are mentioned particularly often (e.g., 100, 150, 200, 250, and 300). Almost 1,500 households or about 14% of all households report monthly expenditures on food at home of exactly 400 euros. About 88% of all households are located on heaping points, with at least 100 observations. Similar findings emerge for the other core consumption categories of the SOEP data.

---

<sup>2</sup> This problem is particularly apparent in the core consumption module of the SOEP. Also, for other consumption categories, the heaped distribution is much more likely to be the true distribution, as the heaping might occur in the real world (e.g., for loans, alimony, child care, etc.). Therefore, in this section, we only focus on the core consumption module.

Figure 3: Frequency distribution for food expenditures at home – SOEP 2010



Source: SOEPv28.

The strong occurrence of heaping may cause problems because univariate statistics (like mean, median and percentiles) are sensitive to including or excluding limit values. For instance, the share of income spent on food at home by households with less than 60% of the median income (the widely-used at-risk-of-poverty threshold) depends strongly on whether a certain heaped value is above or below this threshold. But also regression estimates might be biased from this type of non-classical measurement error.

## 4 Dealing with measurement problems

### 4.1 Inconsistencies

When monthly and yearly information do not correspond, we use the information that better corresponds to the income information given by the household head. For this purpose, the percentiles in the consumption distribution according to the stated monthly and yearly values is computed for each of the households. Then, this percentile rank is compared to the percentiles of the household net income. In case of inconsistent monthly and yearly consumption information, the consumption value is assigned whose percentile was closer to the income percentile. Given that less than 1% of all observed cases are affected by inconsistent monthly and yearly consumption values, the potential bias due to this procedure is rather negligible.

### 4.2 Missing values

The problem of missing values in the SOEP consumption data is handled by multiple imputation (Rubin 1976). “Imputation” means that we replace missings with an estimate of their value. “Multiple” means that we do not assign a single imputed value but several. This assigning of several (in this case: 5) values reflects the uncertainty in the imputation process. We apply a particular multiple imputation technique, multivariate imputation by chained equations (MICE), as suggested by Van Buuren et al. (1999) and implemented by Royston (2004) into Stata. MICE assume that the missing data are missing at random (MAR).<sup>3</sup> Unlike other multiple imputation techniques, MICE does not assume that all variables with missing data follow a single joint distribution, for example, a large joint normal distribution. Instead, MICE models each variable with missing data separately, conditional on a subset of variables in the data set. This allows modeling each variable according to its “nature”, that is, continuous variables by linear regressions, ordinal variables by ordered logit regressions, etc.

MICE consists basically of two iterating steps.<sup>4</sup> Before the first step, all variables with missing data are imputed preliminarily by an automatic routine provided by Stata’s program “ice.ado.” In the first

---

<sup>3</sup> MAR means that the probability that a value is missing depends only on variables used in the imputation procedure and not on unobserved variables.

<sup>4</sup> In order to implement MICE, we make use of “ice.ado” in Stata 11.2.

step, the observed values of the “first”<sup>5</sup> variable with missing values are regressed on other variables (including the preliminary imputed values). In the second step, missing values of the first variable are replaced with predictions according to the first-step regression.<sup>6</sup> Then, the first and second steps are repeated for one variable after the other. The first iteration is completed when all variables with missing data are imputed based on these regressions for the first time. Subsequent iterations repeat the two steps for all variables with missing values, until the coefficients in the regression models converge. The Brooks-Gelman (1998) diagnostic indicates that after 500 iterations the imputation procedure seems to have converged sufficiently.

The multiple imputation of the SOEP consumption module encompass the 16 categories of the core consumption module\*,<sup>7</sup> payback for loans\*, alimony for relatives\*, expenditures for household and cleaning help\*, expenditures for child care\*. In addition, for households in rental housing, further consumption categories are considered: heating and hot water, electricity, and gross rent including utilities. For households in owner-occupied housing, we impute the costs of heating, electricity, regular expenses (water, garbage removal, sidewalk cleaning, etc.), mortgage payment for the dwelling\*, and other housing costs.

The explanatory variables in the imputation procedure can be roughly divided into household demographics, variables relating to the household economic situation, information about the dwelling, and survey-related information (see Overview 1).<sup>8</sup> Missing values for an explanatory variable are imputed using the information from the other explanatory variables.<sup>9</sup> Additionally, the consumption categories in the set of explanatory variables are used for each another, that is, for each consumption category, the other consumption categories are used as explanatory variables.

---

<sup>5</sup> In our case, “first” refers to the variable with the lowest share of missing values, “second” to the variable with the second-lowest incidence of missing values and so on.

<sup>6</sup> All steps of the imputation procedure are carried out independently for the five values (“implicates”), so the variation among the implicates represents the uncertainty of the imputation process itself. Therefore, the coefficients of the underlying regression models as well as the final predictions are interpreted as random estimators with varying outcomes. Hence, the actually used coefficients are drawn from normal distributions defined by the coefficient estimates and their standard deviations. And, instead of directly using the predictions from the regression models, a term randomly drawn from the observed residuals is added to these predictions. As we perform predictive mean matching, the imputed values contain only values that also exist in the observed data. In order to also reflect the uncertainty in this step, we impute a missing value by matching the value of one *randomly* chosen neighbor from the three nearest non-missing neighbors with closest prediction (Royston/White 2011).

<sup>7</sup> Variables marked with an asterisk indicate that we imputed both a filter variable (i.e., binary variable for whether the household had any positive expenditure in that category) and the exact amount (if the observed or imputed filter was “yes”).

<sup>8</sup> Table A.1 in the appendix provides an overview over the explanatory variables used for each consumption component.

<sup>9</sup> We do not describe the imputation of these variables as the imputation takes place in the same procedure as with the consumption categories, and the focus of this documentation is on the consumption module. We also do not distribute imputed values for these other explanatory variables because we basically use them as auxiliary variables.

Overview 1: Information used in the multiple imputation process

**Demographics**

Age structure of household members, number of children and adults, educational level, sex, migration background, region, health status

**Economic situation of the household**

Employment status, monthly household net income, dis-/saving, windfall income (inheritances, bequests, lottery), filter information and amount of other consumption categories, private transfers received from outside the household

**Information about dwelling**

Tenure status, move, modernization, private household, household type

**Survey-related information**

Interview mode, partial unit non-response

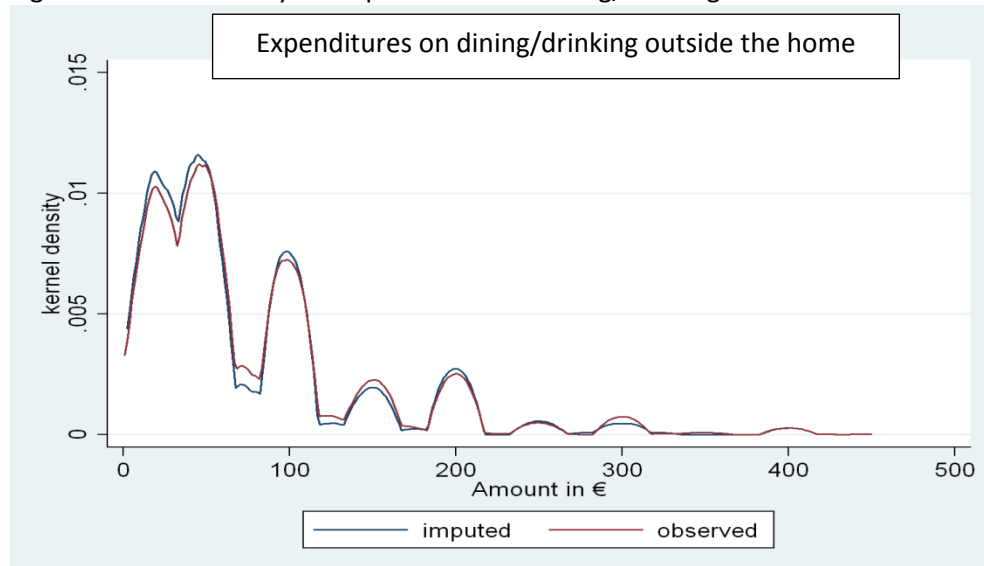
There are not many households with expenditures for alimony, cleaning help, housing, loans, and child care, which can yield to convergence problems in the imputation process. Therefore, for these expenditures, the filter (and not the amount) is considered as an explanatory variable for the imputation process of the other consumption categories. Furthermore, when imputing the filter of the 16 core consumption items, we only include the filter (and not the amount) of the other 15 core consumption items as explanatory variables.

In the end, we assign five different values (“implicates”) for each missing value in order to reflect the uncertainty in the imputation process. Each of the five implicates contains only values that also exist in the observed data to mimic the structure of the observed data. Hence, also the imputed values contain heaped values (see section 2.3).

Figure 4 compares kernel densities of the imputed and the observed values for monthly expenditures on dining and drinking outside the home. The graph clearly depicts that the imputations follow a similar distribution as the observed values, that is, although the problem of item non-response is relaxed, this procedure “generates” a distribution with numerous heaping points. Similar pictures are obtained for the other consumption categories.



Figure 4: Kernel density for expenditures on dining/drinking outside the home – SOEP 2010



Source: SOEPv28.

### 4.3 Heaping

The extent of heaping in the SOEP consumption module is quite large. When analyzing retrospective survey questions in different contexts, other researchers observe non-ignorable heaping patterns as well. Heitjan and Rubin (1990), for example, found heaping in the age of young children as stated by their parents. Wang and Heitjan (2008) provided evidence of heaping in the number of cigarettes, while Lillard et al. (2008) observed heaping patterns in the year of smoking cessation. Also El Messlaki (2010) observed that the distribution of unemployment durations shows specific heaping patterns. All these studies present ways to cope with the heaping phenomenon. However, all suggested solutions are rather special cases designed for the specific variables analyzed. Often external information about specific institutional settings is used to cope with the problem of heaped data. External information is unfortunately not available in the present case. In addition, while all these studies deal with discretely measured variables, the consumption data at hand are inherently continuous.

This section first briefly discusses several alternative solutions to the heaping problem as well as drawbacks of these alternatives. It then describes the procedure used to cope with the heaping phenomenon in more detail.

##### **4.3.1 Alternative solutions to handling heaped data**

One possible solution is the “laissez faire” approach. This means leaving the stated consumption amounts untouched and hoping that the aggregation of all consumption categories into a single total consumption variable smoothes out the heaping points. Yet this solution has two disadvantages. First, researchers still face the same heaping patterns when analyzing single consumption categories. Second, since the heaping patterns are rather strong in the individual consumption components, heaping patterns would persist even in an aggregate measure.

An alternative solution is to fit our consumption data to consumption data from a different source, which does not rely on *retrospective* information. In Germany, the Income and Expenditure Survey (Einkommens- und Verbrauchsstichprobe, EVS) of the Federal Statistical Office might seem most appropriate for this purpose as it collects detailed consumption information by means of a diary. The EVS is a nationwide quota sample of about 55,000 households. A detailed household account log of all earnings and expenditures has to be completed by the participants. This log has to be completed over three months. In addition, a much more detailed household accounts log has to be filled in for food, beverages and tobacco products over one month.

There are several problems with considering the EVS as a benchmark to handle the problem of heaped data. Firstly, the two samples represent different populations. The EVS is a quota sample and, hence, cannot be regarded a random sample, while the SOEP is a random sample of private households in Germany. Secondly, households with a net household income of more than 18,000 euros are excluded from the EVS, while the SOEP does not apply any comparable restriction. Thirdly, the consumption categories used in the EVS do not overlap with those in the SOEP. For instance, the SOEP collects information about transportation in general, while the EVS makes use of a broader concept and includes not only direct costs of transportation but also expenditures for repair, rent for garages, and spare and wear parts. And finally the interview methods differ between the two surveys. The EVS uses a diary to collect information about a certain point in time, while the SOEP uses a retrospective questionnaire to collect information about selected consumption categories for the preceding year. One can assume that these differences between the two surveys might also result in consumption distributions that differ in various ways beyond the existence of heaping points (see also Section 5).

A third alternative is to consider longitudinal information to cope with heaping. For instance, Pudney (2008) makes use of a random effects model to consider different response modes and heaping

behaviors. However, the SOEP implemented the detailed consumption module for the first time in 2010. Thus, no longitudinal information has become available so far that would allow us to follow this research line.

A fourth possibility is described by Battistin et al. (2003). To handle rounding and heaping, they make use of basic household characteristics, “plus a reasonable set of interview quality indicators (such as interview length and interviewer’s assessment of how well the respondent understood the questions), which we assume not to determine consumption level” (Battistin et al. 2003: 370). The information about the interviewer’s assessment on how well the respondent understood the questions is an important variable to explain the heaping process in that paper. Unfortunately this information—or even a comparable variable—is not available in the SOEP, thus their suggested strategy cannot be adapted to the SOEP consumption module. In addition, the applied procedure of Battistin et al. (2003) fails to mirror the lower tail of the distribution in particular for nondurable expenditures.

A fifth alternative defines specific heaping points and allocates observations at these heaping points to the surrounding area. We reject this alternative because it involves several difficult normative decisions: What is a surrounding area? Does the surrounding area differ between heaping points at multiples of 10 vs. multiples of 100 (e.g., 180 vs. 200 euros)? Heitjan and Rubin (1990), for example, attempt to solve the problem of heaping by coarsening data over broad intervals centered around the heaping unit. In addition, which distributional assumptions should be made when assigning observations at heaping points to the surrounding area? And finally, which criteria define heaping points? In regard to this latter question, it is also important to note that not every rounded value constitutes a heaping point. Specifically, it is unclear how many observations have to be at a single point to define this point as a heaping point.

#### **4.3.2 Correcting heaping in the SOEP consumption module**

Instead of relying on a solution to the heaping problem that involves several arbitrary decisions by the researcher, here an approach is selected that is mainly data-driven. The heaping problem is mitigated by approximating the empirical consumption distribution by a theoretical (mathematical) distribution, and then adjusting the consumption data according to the fitted distribution. The first step of this procedure is to determine the theoretical distribution that best fits the consumption data. The remainder of this section deals with this first step. We compare five different theoretical distributions that are known to describe skewed distributions: the Gamma, the Generalized Beta of

the second kind (which includes as special cases the Dagum, the Fisk log-logistic, and the Singh-Maddala distribution), the Gumbel, the Lognormal, and finally the Weibull distribution. For each consumption category, we estimate the parameters of the five distributions by maximum likelihood estimation.<sup>10</sup> Then, we compare how well the theoretical distributions approximate the consumption data by means of graphical and numerical criteria. In a third step, we predict new consumption values according to the parameters of the theoretical distribution that best fits the empirical consumption data. In order to express the uncertainty in the assignment process, we do not just assign a single value but—similarly to the imputation procedure—five different smoothed values.

#### **Graphical criteria**

In the first graphical inspection, we plot the quantiles of the empirical consumption distributions against the quantiles of the fitted theoretical distributions. Figure 5 shows these Q-Q (quantile-quantile) plots for the first consumption category (monthly expenditures on food) for each of the five theoretical distributions.<sup>11</sup> The vertical axes depict values of the empirical distribution; horizontal axes show values of the fitted distribution. Additionally, the graphs include grid lines at the 5, 10, 25, 50, 75, 90, and 95 quantiles.<sup>12</sup> Diagonal lines represent the angle bisectors, i.e., on these lines, each point has the same value on the vertical and the horizontal axis. Points above the diagonal indicate that the values of the empirical distribution exceed the fitted values of the theoretical distributions; analogously, points below the diagonal indicate that the fitted values are larger than the observed values. The large horizontal bars of points signify large heaping points.

In general, it would be ideal to have no areas in which the fitted values systematically deviate from the diagonal. Deviations indicate that the theoretical distribution does not approximate the empirical data well in that area. Although all of the theoretical distributions approximate the consumption data quite well, there are specific areas in which some of the theoretical distributions do not conform to the observed distributions. For instance, for the Weibull and Gumbel distributions on the left tail of the distributions for food expenditures, the points are systematically above the diagonal, indicating that the values of the empirical distribution exceed the fitted values of the theoretical distributions. Similarly, for food expenditures, the lognormal distribution does not approximate the food expenditure data on the upper part properly. The gamma and generalized beta of the second kind

---

<sup>10</sup> User-written programs are provided in Stata for each of the theoretical distributions (see Jenkins 2004).

<sup>11</sup> Due to space limitations, we only present graphs for the first consumption category. Graphs for the other categories are provided upon request. In general, however, the conclusions regarding the fit of the empirical data are the same no matter which consumption category we consider.

<sup>12</sup> In order to mitigate the influence of extreme outliers in the graphical inspections, we only consider the first 99 percentiles, i.e., we disregard the highest percentile.

(GB2) distribution perform particularly well for this consumption category. Furthermore, for both distributions, the diagonals roughly cut the heaping points (the horizontal bars) down the middle, indicating that values at a heaping point are distributed equally to both sides of the point.

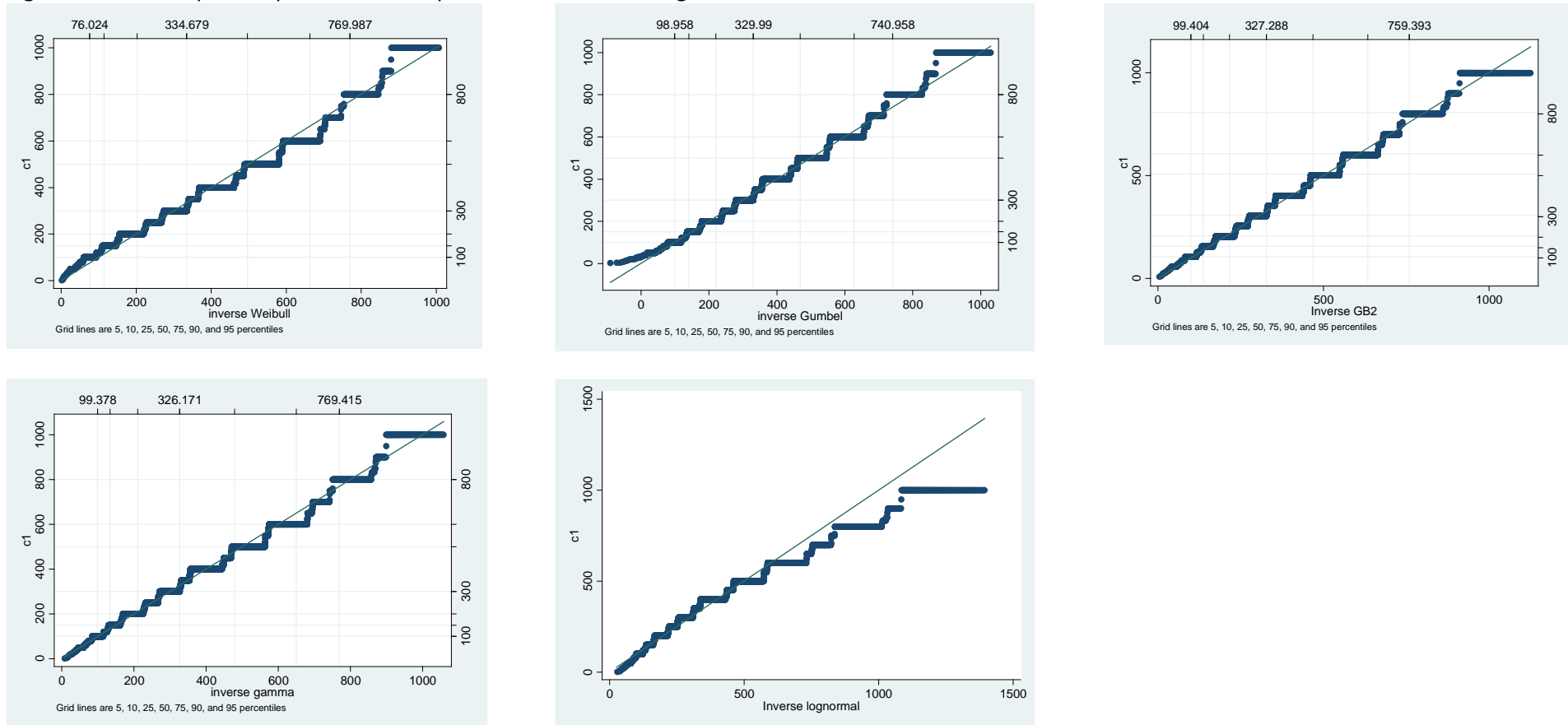
Figure 6, the second graphical inspection, compares the relative and absolute differences between fitted and observed values over the consumption distribution. As with the previous figure, it would be ideal to see no areas in which the differences are systematically below or above the zero line. In these graphs, each dot presents the mean difference (between fitted and observed values) for one of 200 quantiles, i.e., for this consumption component, each dot represents almost 50 households.<sup>13</sup>

---

<sup>13</sup> Similarly to the previous graphical inspection, we do not wish extreme outliers to exert too strong an influence. Hence, we exclude the lowest and the highest ten of the 200 quantiles.

#### 4. Dealing with measurement problems

Figure 5: Quantile-quantile plots for food expenditures at home using different theoretical distributions – SOEP 2010

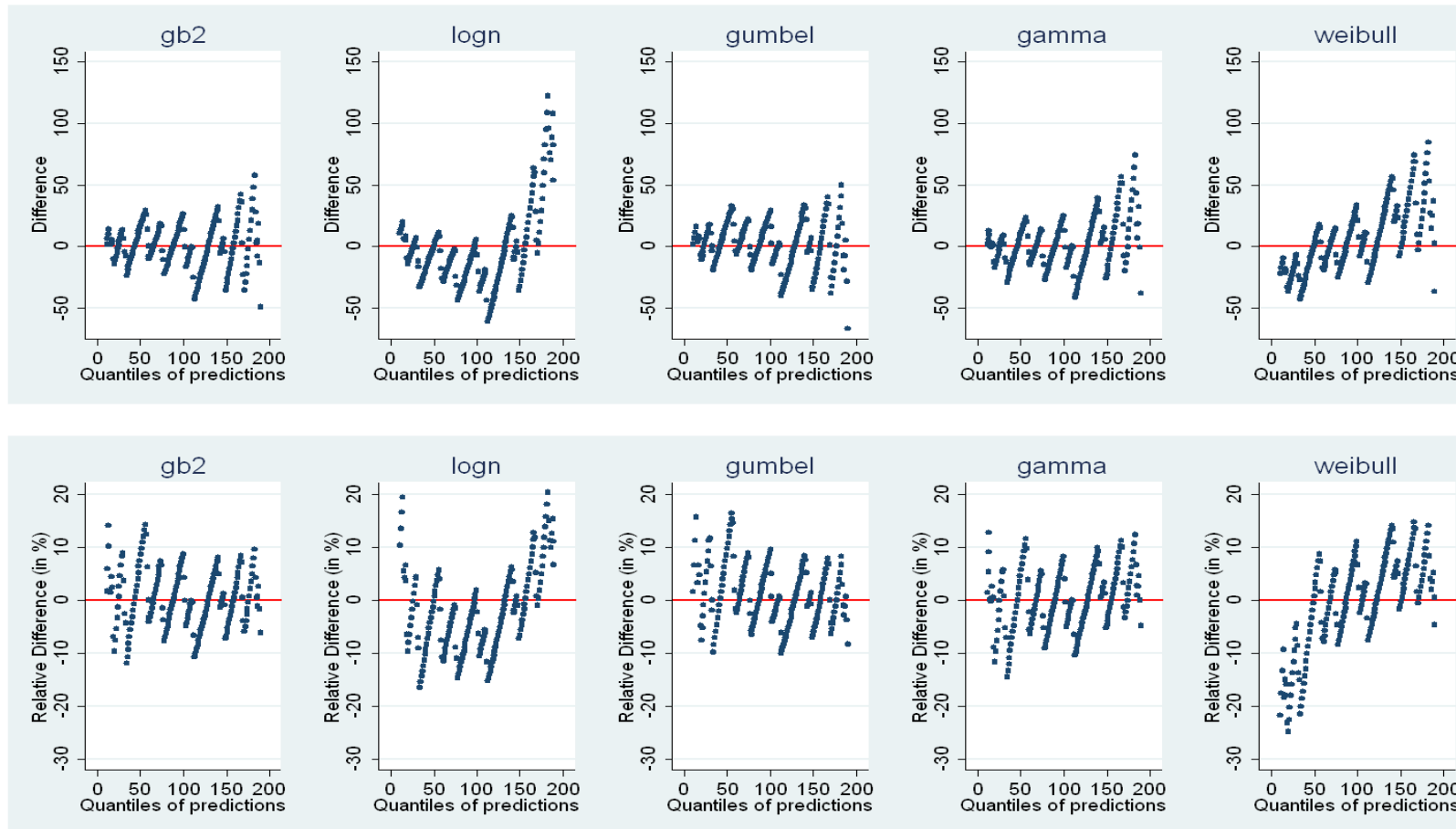


Source: SOEPv28.

#### 4. Dealing with measurement problems

---

Figure 6: Absolute and relative differences between observed and fitted values for food expenditures at home – SOEP 2010



Source: SOEPv28.

In the upper part of Figure 6, absolute differences between fitted and observed values are presented. The focus of the graphs is on the right side of the distribution as the differences naturally increase with the observed values. For the log-normal and the Weibull distribution we find a rather strong deviation in the upper part of the distribution. Starting from the 150<sup>th</sup> quantile, nearly all cases lie above the expected zero line. The gamma distribution performs somewhat better although in the top part of the distribution again nearly all values deviate from the zero line. The GB2 and the Gumbel distribution perform better in comparison to the other three distributions. All values are rather evenly scattered around the zero line.

In order to be able to investigate systematic differences on the left tail of the distribution as well, we also look at relative differences, i.e., differences between fitted and observed values divided by the observed values (lower panel of Figure 6). This graphical analysis confirms the findings of the first by and large. Again, Weibull and Gumbel distribution do not fit the data well on the left tail of the distributions, as well as the Lognormal distribution on the right tail. Yet, this graphical inspection also suggests that the GB2 distribution is slightly superior to the gamma distribution as there are some areas in the graph of the gamma distribution in which the points are systematically different from the zero line, e.g., around the 150<sup>th</sup> quantile.

#### **Numerical criteria**

In the following, we turn to numerical criteria that might help to decide which theoretical distribution best approximates the empirical consumption data. All three criteria considered are based on the idea that we want to mitigate the heaping problem, but want to avoid too many data transformations. The basic idea of the numerical criteria is to minimize the difference between fitted and observed values. The first criterion is the mean squared difference (MSD), which is computed as:

$$MSD = \frac{1}{N} \sum_i (\hat{c}_i - c_i)^2,$$

where  $\hat{c}_i$  is the fitted consumption value,  $c_i$  the observed consumption value, and  $N$  the number of households that spend a positive amount on the consumption component. This criterion computes the squared difference between fitted and observed values and, hence, puts a great deal of weight on large differences. This criterion is particularly sensitive to the right tail of the distribution.

The second criterion, the mean squared relative difference (MSRD), is more sensitive to the left tail of the distribution and is computed as follows:



$$MSRD = \frac{1}{N} \sum_i \left( \frac{\hat{c}_i - c_i}{c_i} \right)^2$$

Similarly, the mean absolute difference (MAD) places less weight on extreme differences, which generally occur on the right tail of the distribution:

$$MAD = \frac{1}{N} \sum_i |\hat{c}_i - c_i|$$

For each of the 16 consumption components, Table 2 displays the name of the distribution with the minimum value for each of the three criteria. The table shows that according to these criteria, the GB2 distribution outperforms the other four distributions. For all 16 components, it provides the smallest mean absolute difference and the smallest mean squared relative difference. And in 14 of 16 cases, the GB2 has also the minimal mean squared error.

Table 2: Numerical criteria to compare several theoretical distributions – SOEP 2010

Consumption category	MSD	MSRD	MAD
1	<b>GB2</b>	<b>GB2</b>	<b>GB2</b>
2	Log-normal	<b>GB2</b>	<b>GB2</b>
3	<b>GB2</b>	<b>GB2</b>	<b>GB2</b>
4	<b>GB2</b>	<b>GB2</b>	<b>GB2</b>
5	<b>GB2</b>	<b>GB2</b>	<b>GB2</b>
6	<b>GB2</b>	<b>GB2</b>	<b>GB2</b>
7	<b>GB2</b>	<b>GB2</b>	<b>GB2</b>
8	<b>GB2</b>	<b>GB2</b>	<b>GB2</b>
9	<b>GB2</b>	<b>GB2</b>	<b>GB2</b>
10	<b>GB2</b>	<b>GB2</b>	<b>GB2</b>
11	<b>GB2</b>	<b>GB2</b>	<b>GB2</b>
12	<b>GB2</b>	<b>GB2</b>	<b>GB2</b>
13	<b>GB2</b>	<b>GB2</b>	<b>GB2</b>
14	Log-normal	<b>GB2</b>	<b>GB2</b>
15	<b>GB2</b>	<b>GB2</b>	<b>GB2</b>
16	<b>GB2</b>	<b>GB2</b>	<b>GB2</b>

Note: Consumption categories according to Figure 1. For each consumption category, the cells contain the name of the distribution with the lowest value for the criterion indicated by the column header. MSD stands for mean squared difference, MSRD for mean squared relative difference, and MAD for mean absolute difference. GB2 indicates the generalized beta of the second kind function.

Source: SOEPv28.

Since we are not only interested in criteria that describe the overall fit, we look at the criteria within part of the distribution, i.e., quartiles. Table 3 provides for each consumption category the theoretical distribution with the smallest mean absolute difference for the four quartiles. The above-mentioned findings are largely confirmed, i.e., the GB2 distribution performs best in comparison to

the other distributions. The GB2 has the minimal MAD in 52 of the 64 cells of the table. The second-best distribution, the lognormal distribution, provides the smallest MAD in only 8 cells.<sup>14</sup>

Table 3: Minimal mean absolute differences according to quartiles – SOEP 2010

Consumption category	MAD (Q1)	MAD (Q2)	MAD (Q3)	MAD (Q4)
1	<b>GB2</b>	Gamma	Gumbel	<b>GB2</b>
2	Log-normal	<b>GB2</b>	<b>GB2</b>	<b>GB2</b>
3	Log-normal	<b>GB2</b>	<b>GB2</b>	<b>GB2</b>
4	<b>GB2</b>	Log-normal	<b>GB2</b>	<b>GB2</b>
5	<b>GB2</b>	Log-normal	<b>GB2</b>	<b>GB2</b>
6	<b>GB2</b>	<b>GB2</b>	<b>GB2</b>	<b>GB2</b>
7	<b>GB2</b>	<b>GB2</b>	<b>GB2</b>	<b>GB2</b>
8	<b>GB2</b>	Weibull	<b>GB2</b>	<b>GB2</b>
9	<b>GB2</b>	Log-normal	<b>GB2</b>	<b>GB2</b>
10	Log-normal	<b>GB2</b>	<b>GB2</b>	<b>GB2</b>
11	<b>GB2</b>	<b>GB2</b>	Log-normal	<b>GB2</b>
12	<b>GB2</b>	<b>GB2</b>	<b>GB2</b>	<b>GB2</b>
13	<b>GB2</b>	Log-normal	<b>GB2</b>	<b>GB2</b>
14	Gamma	<b>GB2</b>	<b>GB2</b>	<b>GB2</b>
15	<b>GB2</b>	<b>GB2</b>	<b>GB2</b>	<b>GB2</b>
16	<b>GB2</b>	<b>GB2</b>	<b>GB2</b>	<b>GB2</b>

Note: Consumption categories according to Figure 1. Cells contain the name of the distribution with the lowest value for the mean absolute differences (MAD) in each quartile for each consumption category. GB2 indicates the generalized beta of the second kind function.

Source: SOEPv28.

### The generalized beta of the second kind function (GB2)

Both graphical and numerical inspections suggest that the GB2 distribution provides the best fit to the SOEP consumption data. The main advantage of the GB2 distribution is its flexibility. The GB2 was designed to describe variables with a skewed distribution and, therefore, was shown to provide a good fit to, e.g., income data (McDonald 1984). Its density function is given by:

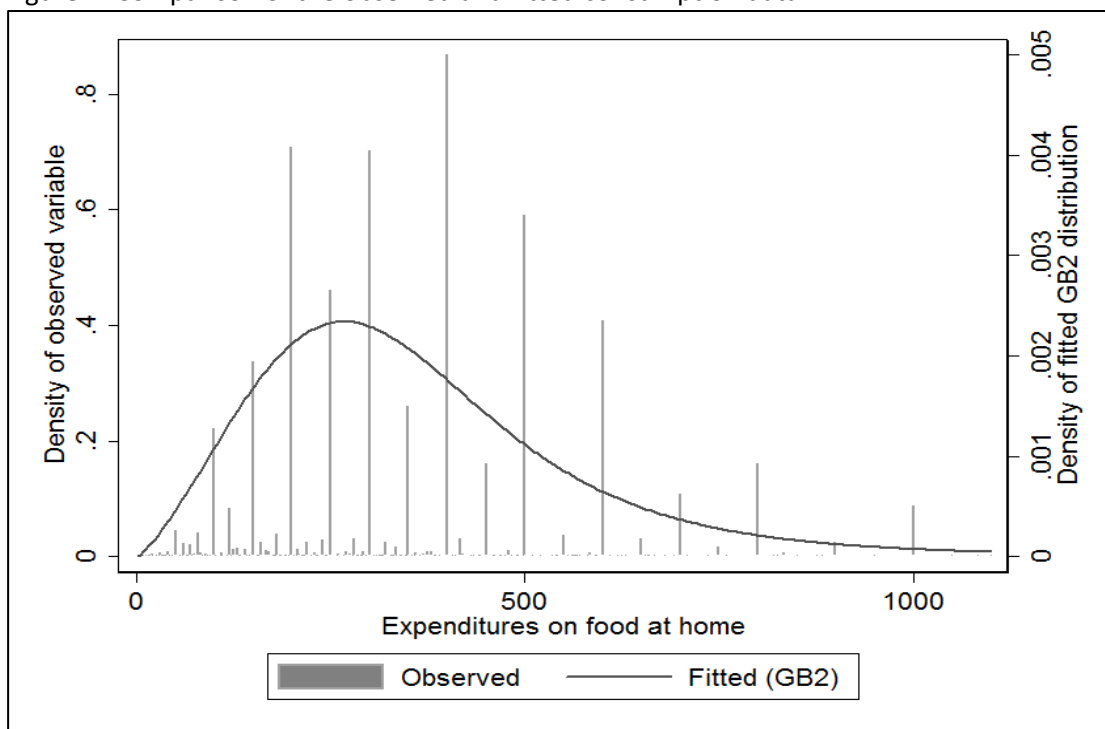
$$f(x) = ax^{ap-1} \times \left\{ b^{ap} \times B(p, q) \times \left[ 1 + \left( \frac{x}{b} \right)^a \right]^{-(p+q)} \right\}^{-1},$$

where  $x$  is a positive random variable (here: the consumption data),  $B(p, q)$  is the beta function, and  $a, b, p, q$ , are positive parameters. The GB2 distribution incorporates as special cases the Singh-Maddala (1976) distribution (for  $p = 1$ ), the Dagum (1977) distribution (for  $q = 1$ ), and the Fisk (1961) distribution (for  $p=1$  and  $q=1$ ), which is also known as log-logistic distribution. These three distributions were also designed to fit right-skewed variables.

<sup>14</sup> Similar pictures emerge when looking at the mean squared difference or the mean relative difference within quartiles.

The GB2-based smoothing procedure works as follows. We first estimate the parameters of the GB2 by maximum likelihood for each consumption category (see Jenkins 2004). Then, we predict new consumption values according to the estimated parameters of the GB2 distribution. Figure 7 compares the distribution of the observed values (spikes) and the fitted GB2 distribution (smoothed line) for the category “food expenditures at home.” This figure also highlights how the approach works: For any given (heaping) point, the observed values are randomly assigned to the neighboring area, so that in the end, the values follow a GB2 distribution.

Figure 7: Comparison of the observed and fitted consumption data



Source: SOEPv28.

### Combining smoothing and multiple imputation

This section describes how we combine the smoothing process with multiple imputation. First of all, note that we first impute the data as described in Section 4.2 and then smooth the consumption data. Smoothing is performed separately for each of the five imputation versions. So far, these five implicates contain the same value for observations with non-missing consumption data. Hence, differences in the parameter estimates of the GB2 distribution between the five implicates are only due to observations with imputed consumption information. In order to better capture the uncertainty in the smoothing process—given that we do not know the true underlying distribution of the consumption data—we estimate the parameters of the GB2 distribution on bootstrapped samples. Therefore, we do not just assign a single value but five different smoothed (and imputed)

#### 4. Dealing with measurement problems

---

values for every household (with positive expenditures in the specific consumption component). Note that while the imputation procedure affects only households with missing data, smoothing affects all households (at least slightly).

Table 4: Observed and generated information for expenditures for food at home – selected households

Household	Observed	Implicate 1	Implicate 2	Implicate 3	Implicate 4	Implicate 5
1	.	381.36	292.42	594.93	629.19	492.56
2	300	294.95	306.82	301.89	296.33	292.69
3	300	300.01	295.22	293.19	292.65	290.79
4	500	463.42	464.79	531.48	508.87	485.10
5	0	0	0	0	0	0
6	200	204.43	181.72	195.62	205.73	179.62
7	400	420.24	405.29	368.72	389.31	417.63
8	500	535.28	509.37	487.80	488.43	515.74
9	250	244.07	243.07	252.57	255.89	243.16
10	500	526.20	481.41	533.20	473.65	506.17
11	.	147.86	201.36	152.65	0	141.28
12	100	82.27	91.93	98.95	93.54	111.80
13	875	875.77	877.25	871.13	861.64	870.29

Source: SOEPv28.

As shown in Table 4, in the end there are five different implicates based on imputed and smoothed information for each household. For instance, the first household refused or was not able to provide the exact amount of food expenditures at home. After the imputation and smoothing process, one gets estimates between 292 and 629 euros. For the second household, the household head stated an amount, but this value was affected by heaping. Due to the smoothing procedure, we get five implicates which range between 292 and 306 euros, which are rather evenly scattered around the heaping point. For household 11, both the filter question and the amount of food expenditures at home are missing. For one implicate, the imputation of the filter question yields a “no”, i.e., in implicate 4 we assign a value of zero to this household. For the other four implicates, the imputation of the filter information yields a “yes”, i.e., the generated information ranges between 141 and 201 euros. The applied smoothing procedure leads to generated values also for cases unaffected by heaping. Household 13 stated expenditures for food at home of 875 euros. This is a rather unique value. However, given that we did not make any assumption to define heaping points, this implies that even for this case, we get five different implicates, which range between 861 and 877 euros. This points to a specific characteristic of the applied method, namely that the ordering of the observed values is retained for the five implicates.

##### **Smoothing at the extreme right tail of the distribution**

The approximation of the empirical distributions using the GB2 distribution does not work well for those with extremely high consumption values. This problem would also arise when making use of alternative theoretical distributions, since only few data points are available to approximate this special section of the distribution. Therefore, we decided not to change the highest values in each consumption category (see Table 5). This implies that any value—even if this is a heaping point—in the top area remains the same after the smoothing procedure.

However, heaping occurs only rarely in the top part of any expenditure category, while rather unique values are in the majority. We apply the following rule to determine the cut-off points: A cut-off point is the last value before two subsequent values of the observed consumption distribution with assigned values of at least two imputation versions both above and below these observed consumption values.

Table 5 provides an overview over the cut-off points for the 16 consumption categories as well as the share of observations with positive, non-missing consumption values on and above these specific cut-off points. These observations are not subject to the smoothing procedure. The table indicates that the share of households with values above the cut-off value ranges between 0.17 % and 2.56 % for the different consumption categories, with median 0.67 %.

Table 5: Overview over the cut-off points and the affected households

Consumption component	Cut-off value	No. of households above cut-off	Share of households above cut-off (%)
1	1800	17	0.17
2	750	25	0.31
3	416	132	1.36
4	350	35	0.35
5	800	20	0.22
6	360	38	0.37
7	525	31	1.17
8	350	13	0.21
9	800	27	0.33
10	1274	115	1.74
11	1285	54	0.97
12	516	98	1.07
13	440	77	1.44
14	870	19	0.29
15	900	62	2.56
16	2000	20	1.13

Source: SOEPv28.

## 5 A comparison with the EVS

The Income and Expenditure Survey (EVS) of the Federal Statistical Office is the only official data source with detailed information about consumption in Germany. Thus the importance of comparing the EVS to the SOEP consumption data seems obvious. However, there are several aspects that need to be considered when comparing the two data sources (see also Section 4.3.1). Firstly the EVS is a quota sample which was conducted in 2008, while the SOEP is a representative sample of the population of persons living in private households. Secondly, a consumption module was used in the SOEP in 2010 and in the EVS in 2008. Thirdly, the EVS excludes households with a net household income of more than 18,000 euros, while the SOEP does not apply any comparable restriction. Fourthly, the interview methods differ between the two surveys. The EVS uses a diary to collect information for a certain point in time, while the SOEP uses a retrospective questionnaire to collect information about selected consumption categories for the preceding year. Finally, the consumption categories collected do not fully overlap. Summing up, due to the many differences between EVS and SOEP, a perfect overlap between the consumption distributions in two data sources should not be expected. However, the two are likely to reflect the relevance of certain consumption categories in similar ways.

Although the observation periods deviate from each other, the number of private households in the target population is only slightly higher in the SOEP (2010), with 40.3 million compared to 39.4 million in the EVS (2008). The Federal Statistical Office announced the mean of total private consumption in the EVS to add up to 2,245 euros per household (Statistisches Bundesamt 2012). However, there are several consumption items that are not considered in this figure, including alimony, insurance contributions, amortization and interest. Additionally, some consumption categories are not explicitly considered in the SOEP. When looking at the consumption categories that are surveyed in a fairly similar manner in both surveys, one finds total private expenditures of about 2,320 euros per month per household in the EVS and of only 1,870 euros in the SOEP, i.e., the EVS shows 24% more consumption. However, also the mean net household income is 25% higher in the EVS (Statistisches Bundesamt 2010). Therefore, if one relates the total consumption to mean and median monthly net household income, the ratio is rather similar across the two surveys. Total private consumption makes up a share of 80% of mean net household income in both EVS and SOEP. The higher consumption value and higher net household income provided by the EVS could be the result of the underlying quota sample.<sup>15</sup>

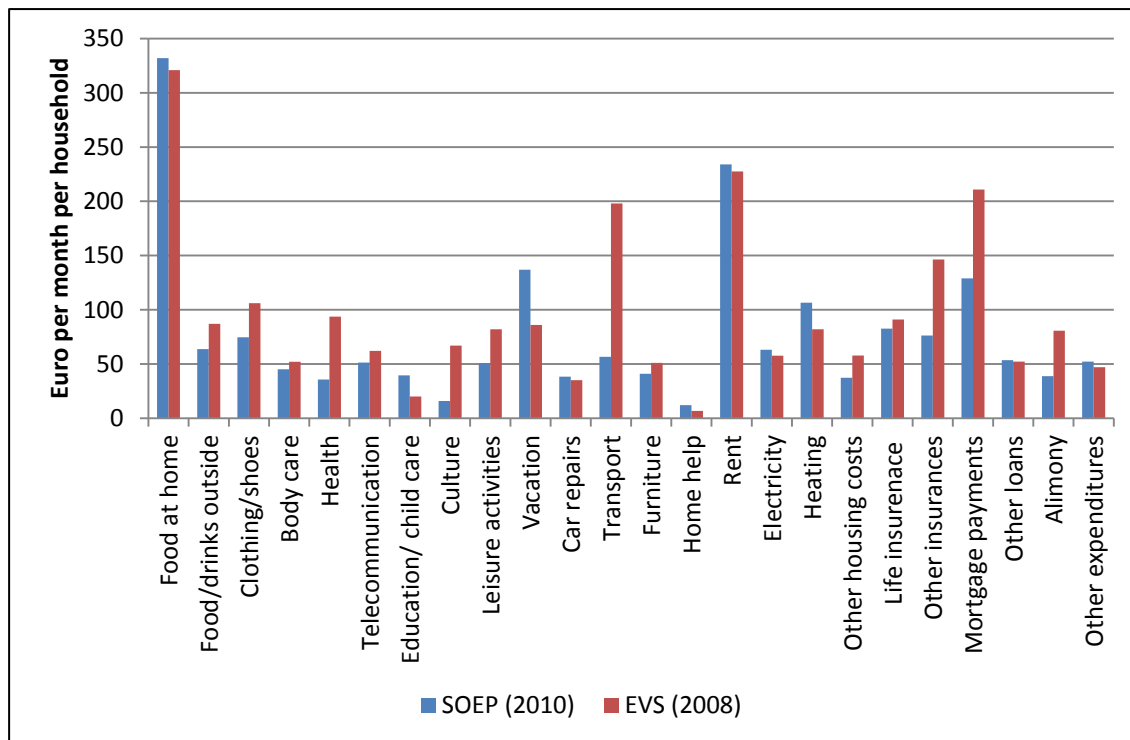
The consumption items that are part of total private consumption in the EVS but not considered in the SOEP are imputed rents<sup>16</sup>, interior/home appliances, costs for vehicles, telephone, leisure tools, and maintenance for renters. They accumulate to an average of 505 euros per month. If one restricts the comparison to those items that are covered in both surveys (Figures 8 and 9), one observes fairly close conformity overall between the two.

---

<sup>15</sup> The EVS shows a couple of prominent deviations compared to the German Microcensus (see, e.g., Becker et al. 2002).

<sup>16</sup> The SOEP consumption module did not collect information about imputed rents. However, the SOEP provides a generated value of net imputed rents. We refrain from a comparison to the EVS because the EVS provides only gross values of imputed rent, which by definition yield higher values than the net amounts in the SOEP.

Figure 8: Consumption portfolio: absolute amounts in SOEP and EVS



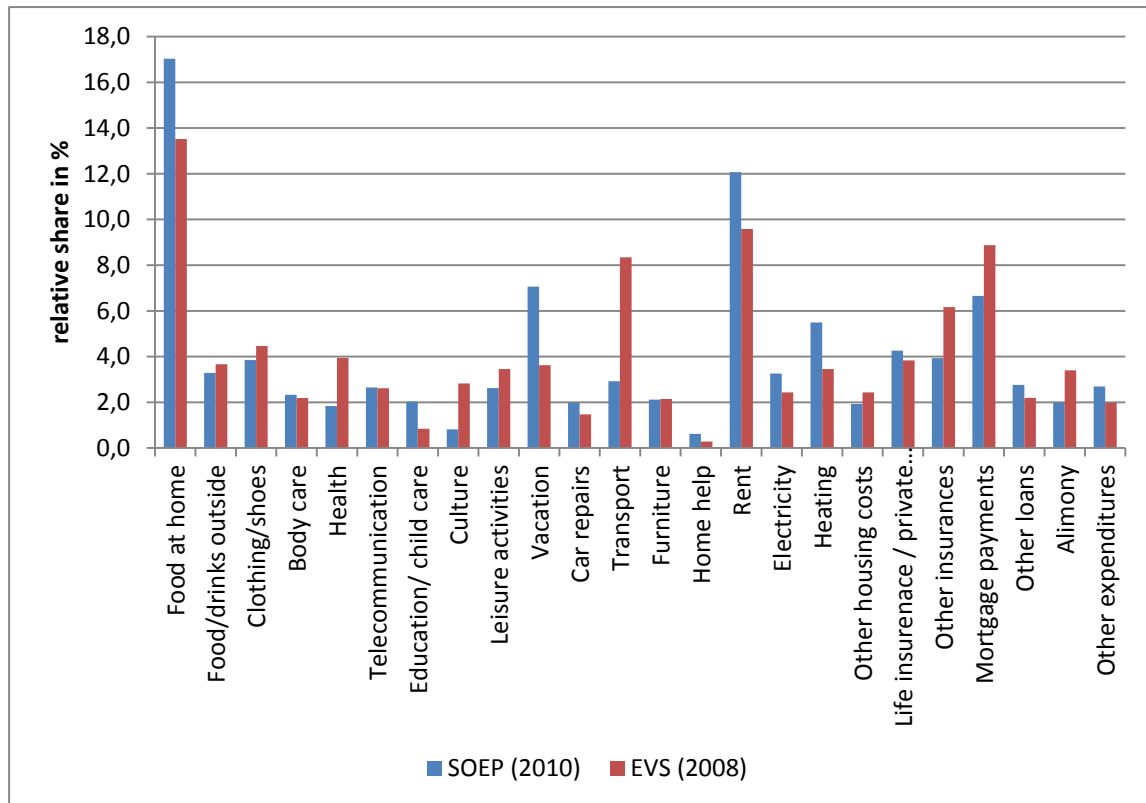
Source: SOEPv28 and EVS 2008.

In both surveys, the three most important consumption components are food at home, rent, and mortgage payments. They add up to 759 euros in the EVS and 693 euros in SOEP, which correspond to a share of 32% of total private consumption in the EVS and 36% in the SOEP. However, a distinct difference occurs for transportation and vacation. While the latter is clearly higher in the SOEP, the former is much higher in the EVS. The costs for transportation are more comprehensive in the EVS than in the SOEP: The SOEP collects information about transportation in general, while the EVS includes not only direct costs of transportation but also expenditures for repairs, rent for garages, and spare and wear parts. Another explanation could be that travel costs are subsumed under the topic of “transportation” in the EVS, while SOEP respondents provide this information in the item “vacation.”

For most consumption items, total monthly amount and share of total consumption are fairly similar between the two surveys. However, for some smaller consumption items, there are also a few differences, e.g., for health, education, and culture. Given that the two surveys differ in various respects, discrepancies in these areas have to be expected.



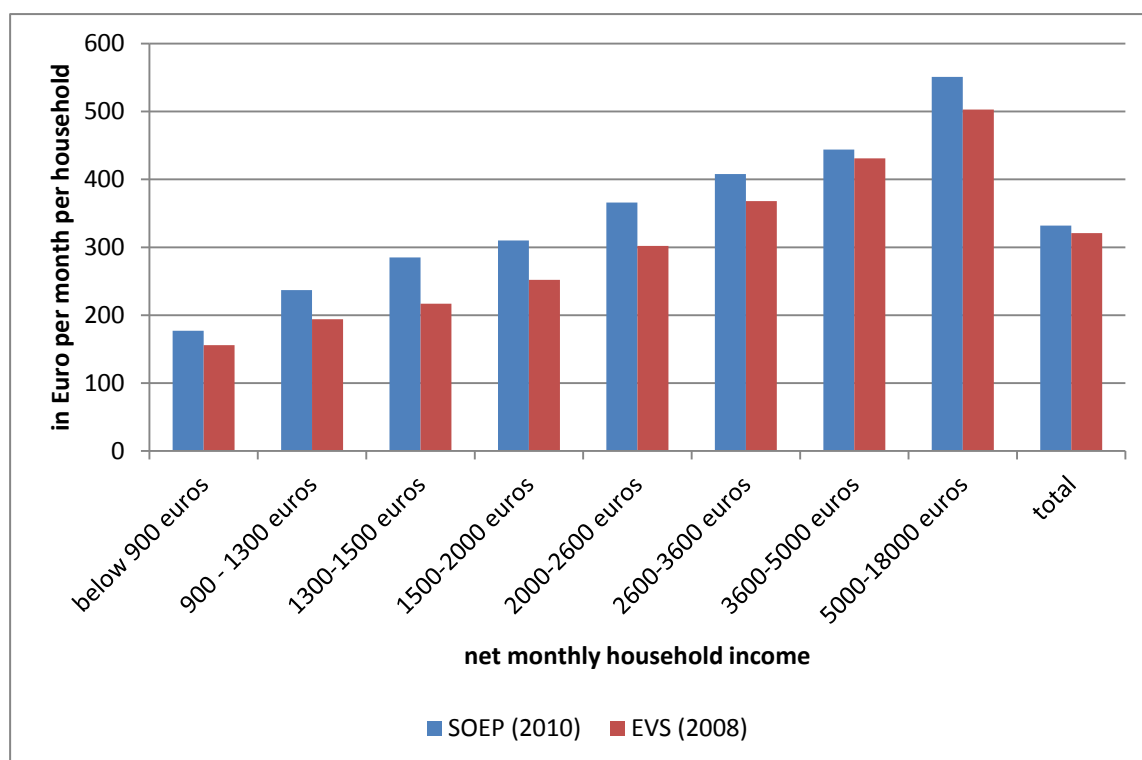
Figure 9: Consumption portfolio: Relative amounts in SOEP and EVS



Source: SOEPv28 and EVS 2008. Note: the considered consumption categories add up to 100% respectively.

Food at home is the most significant consumption category in both surveys. Thus, in an additional comparison, we break down spending in this consumption category by income groups (Figure 10) and by household size (figure 11). In addition, the two surveys only differ by 3% in the absolute amount spent on food at home, which eases the comparison. Income is given here by net monthly household income. Figure 10 shows that in both surveys, the absolute amount spent on food at home increases the larger the household's income. In the top income group, this amounts to more than 500 euros in both the EVS and the SOEP. Over the whole income distribution, spending on food at home is always somewhat higher in the SOEP. The biggest difference of 20-30 % occurs for those households with a net household income between 900 and 2000 euros. A potential explanation might be a recall error in the SOEP data, given that the respondents might not perfectly remember the exact amount spent on food. Another factor could be the increase in food prices between 2008 and 2010.

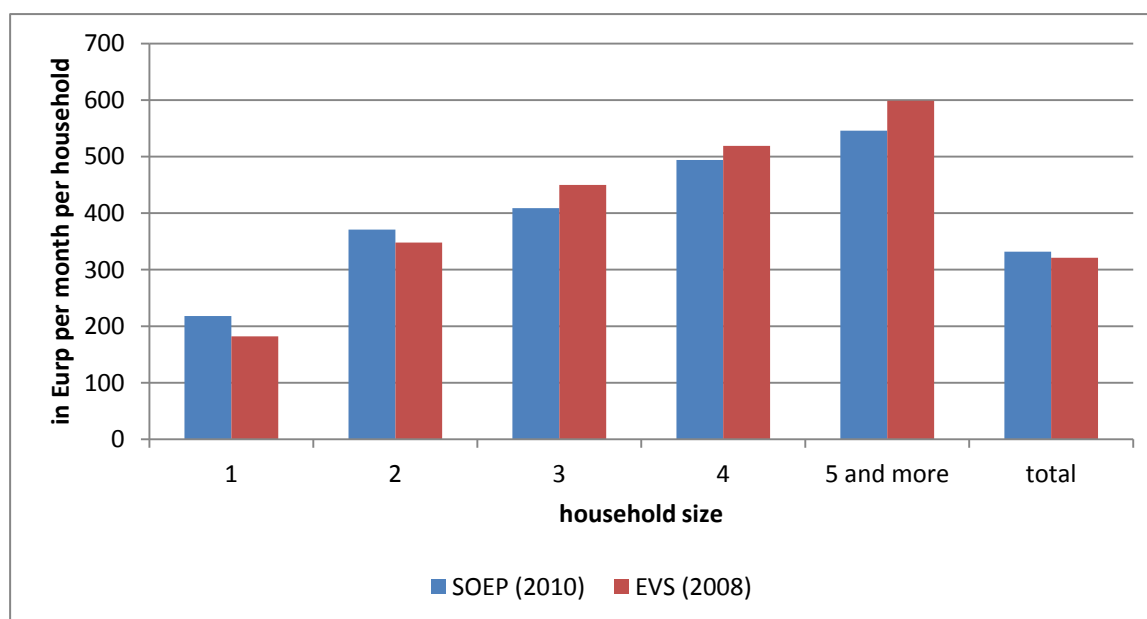
Figure 10: Expenditures on food at home by income groups in SOEP and EVS



Source: SOEPv28 and EVS 2008.

When comparing the costs for food at home in the two surveys by household size, we find again that with increasing number of household members, spending on food also rises. However, the increase is more pronounced in the case of the EVS. The diary method might yield more precise information about actual private consumption. The more members a household has, the more difficult it becomes to estimate the amount spent on food when answering a simple question about the total sum spent on food, as used in the SOEP.

Figure 11: Expenditures on food at home by household size in SOEP and EVS



Source: SOEPv28 and EVS 2008.

## 6 Data distribution

The consumption variables generated according to the procedures described in section 4, are distributed in the new generated SOEP data set “hconsum.” The dataset “hconsum” is organized in the “wide” format, meaning that for every household there is one line in the data set that contains the values of the five implicates (imputation versions) for each of the 28 consumption items (16 core and 12 other consumption items). Additionally, the data set contains a flag variable for every consumption category indicating which changes were made to each consumption value. Table 6 provides an overview of the possible flag values, where editing refers to the process of correcting inconsistent information (see 1.1 and 2.1).

Table 6: Codes of the flag variables

Code	Label
0	no change
1	edited
2	value imputed
3	filter imputed
10	smoothed
11	edited & smoothed
12	value imputed & smoothed
13	filter imputed & smoothed

In case researchers do not want their consumption variables to include changes from all steps of the data preparation, the flag variables provide researchers the opportunity to select individual solutions. For instance, a researcher who wants to use only the smoothed and edited values but not the imputed values can exclude values with flag codes 2, 3, 12, and 13.

Table 7 lists the variables that are included in the data set “hconsum.” Variables “consum1”-“consum16” refer to the sixteen core consumption questions. Variables starting with the letter „f“ indicate the flag variables; the letter “x” at the end of the variable name indicates the imputed version (“a”, “b”, “c”, “d”, “e”).

Table 7: Overview over the variables in the data set “hconsum”

Variable	Description
hhnrakt	Current Wave HH Number
svyyear	Survey year
consum1x	c: food at home, vers. x
consum2x	c: food/drinks outside the home, vers. x
consum3x	c: clothing/shoes, vers. x
consum4x	c: body care, vers. x
consum5x	c: health, vers. x
consum6x	c: telecommunication, vers. x
consum7x	c: education/further training, vers. x
consum8x	c: culture, vers. x
consum9x	c: leisure activities, vers. x
consum10x	c: vacations, vers. x
consum11x	c: life insurance, private pension insurance, vers. x
consum12x	c: other insurances, vers. x
consum13x	c: motor vehicle repairs, vers. x
consum14x	c: transport, vers. x
consum15x	c: furniture, vers. x
consum16x	c: other expenditures, vers. x
rheat_x	c: heating/warm water (renter), vers. x
rrent_x	c: rent (renter), vers. x
relectr_x	c: electricity (renter), vers. x
oheat_x	c: heating (owner), vers. x
oelectr_x	c: electricity (owner), vers. x
outil_x	c: other costs (owner), vers. x
mortgage_x	c: mortgage payments (owner), vers. x
housing_x	c: housing costs (owner), vers. x
loan_x	c: loans for consumer expenditures, vers. x
alimony_x	c: alimony for relatives, vers. x
clhelp_x	c: cleaning/household help, vers. x

## 6. Data distribution

---

chcare_x	c: child care, vers. x
fconsum1	flag: food at home
fconsum2	flag: food/drinks outside the home
fconsum3	flag: clothing/shoes
fconsum4	flag: body care
fconsum5	flag: health
fconsum6	flag: telecommunication
fconsum7	flag: education/further training
fconsum8	flag: culture
fconsum9	flag: leisure activities
fconsum10	flag: vacations
fconsum11	flag: life insurance, private pension insurance
fconsum12	flag: other insurances
fconsum13	flag: motor vehicle repairs
fconsum14	flag: transport
fconsum15	flag: furniture
fconsum16	flag: other expenditures
fmortgage	flag: mortgage payments (owner)
fhousing	flag: housing costs (owner)
floan	flag: loans for consumer expenditures
falimony	flag: alimony for relatives
fclhelp	flag: cleaning/household help, vers.
fchcare	flag: child care
frheat	flag: heating/warm water (renter)
frrent	flag: rent (renter)
frelectr	flag: electricity (renter)
foheat	flag: heating (owner)
foelectr	flag: electricity (owner)
foutil	flag: other costs (owner)

**Appendix**

Table A.1 Explanatory variables used in the imputation procedure

Explanatory variables	For all										For renter			For owner							
	F: 16 core cons.	A: 16 core cons.	F: loans	A: loans	F: alimony	A: alimony	F: cleaning help	A: cleaning help	F: child care	A: child care	Heating	Electricity	Rent + utility	F: mortgage	A: mortgage	Heating	Electricity	Running costs	F: housing	A: housing	
Consumption	F: 16 core consum.	x	-	x	-	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	A: 16 core consum.	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	F: loans	x	x	-	x	x	-	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	A: loans	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	F: alimony	x	x	x	x	-	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	A: alimony	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	F: cleaning help	x	x	x	x	x	x	-	x	x	x	x	x	x	x	x	x	x	x	x	x
	A: cleaning help	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	F: child care	x	x	x	x	x	x	x	x	-	x	x	x	x	x	x	x	x	x	x	x
	A: child care	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	Heating	x	x	x	x	x	x	x	x	x	-	x	x	-	-	-	-	-	-	-	-
	Electricity	x	x	x	x	x	x	x	x	x	x	-	x	-	-	-	-	-	-	-	-
	Rent/utility	x	x	x	x	x	x	x	x	x	x	x	-	-	-	-	-	-	-	-	-
	F: mortgage	x	x	x	x	x	x	x	x	x	-	-	-	-	x	x	x	x	x	x	x
	A: mortgage	x	x	x	x	x	x	x	x	x	-	-	-	-	-	x	x	x	x	x	x
	Heating	x	x	x	x	x	x	x	x	x	-	-	-	x	x	-	x	x	x	x	x
	Electricity	x	x	x	x	x	x	x	x	x	-	-	-	x	x	x	-	x	x	x	x
Running costs	x	x	x	x	x	x	x	x	x	-	-	-	x	x	x	x	-	x	x	x	
F: housing	x	x	x	x	x	x	x	x	x	-	-	-	x	x	x	x	x	x	-	x	
A: housing	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
HH demographics	HH-typ (8)	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
	Age 17-30	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
	Age 31-60	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
	Age 61+	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
	Kids [0, 4]	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
	Kids [5, 10]	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
	Kids [11, 15]	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
	Kids [16, 18]	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
	# adults (4)	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
	Age HHH	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
	Sex HHH	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
	Migback HHH	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
	East	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
HH economic	HH income	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
	Windfall	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
	Income vs. cost (3)	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
	Poor health	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
	Working	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
	Civil service	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
	Selfempl.	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
	Educ. (3)	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
Dwelling	Renter	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
	Rooms	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
	Size	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
	Renovations	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
Survey	Survey instrum. (7)	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
	Institut. HH	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
	HH moved	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
	PUNR	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	

Figure A1: Further consumption items in the SOEP

42. Are you currently paying back **loans and interest on loans** that you took out to make large purchases or other expenditures?

Please *do not include* loan, mortgage or interest payments which you have already stated in previous questions.

Yes .....

No .....

**Skip to question 45!**

43. How high is the monthly rate that you pay on **these** loans?

If you don't know the exact amount, **please estimate!**

Loan repayment  
 (include interest payments) .....  euros per month

53. Do you or another household member **currently** provide regular financial support to family members or relatives?

including former spouse.

Yes .....  .....  euros per month

No .....

62. Do you regularly or occasionally employ household help?

Yes, regularly .....

Yes, occasionally ....

No .....

How much do school, care, and the activities described above cost you? .....  euros per month

70. How much do school, care, and the activities described above cost you?

Average monthly cost in euros .....  .....  .....  .....

No costs .....  .....  .....  .....



**For renter**


23. How much is the monthly rent?

euros      I don't pay rent .....  ➔ Skip to question 38!

24. Are the costs for heating (and usually also warm water) included in the rent?

Yes .....       No .....

25. How high are the heating costs per month?

 If you don't know the exact amount, **please estimate!**

euros      Don't know ....

25a How high are your average monthly electrical costs?


euros      Don't know ....

26. Are other costs included in the rent, for example for water, garbage removal, etc.?


Yes, included in full .....  ➔ How much are they?  euros per month    Don't know ....   
 Yes, included in part ...   
 No .....

**For owners**

29. Do you still have financial obligations, for example loans or a mortgage, for this house or flat in which you live?

Yes .....        No .....  ➔ Skip to question 31!

30. How high are the monthly loan or mortgage payments including interest for this loan or mortgage?

 If you don't know the exact amount, **please estimate!** Please do so also in the next questions.

Loan or mortgage payments and interest .....  euros per month

32. What were the costs for heating last calendar year? .....  euros per year

32a How high were your electrical costs in the last calendar year? .....  euros per year

33. And how high were the costs for water, garbage removal, street cleaning, etc. last year? .....  euros per year

34. Do you pay fees for the management or maintenance of the building?

Yes .....  ➔  euros per month  
 No .....

---

## References

- Battistin, E., R. Miniaci and G. Weber (2003): What do we learn from recall consumption data? *Journal of Human Resources*, 38(2): 354-385.
- Becker, I., J.R. Frick, M.M. Grabka, R. Hauser, P. Krause and G.G. Wagner (2002): A comparison of the main household income surveys for Germany: EVS and SOEP. In: Hauser, R. and Becker, I. (Eds.): Reporting on income distribution and poverty. Perspectives from a German and a European Point of View. Springer, p. 55-90.
- Blundell, R., L. Pistaferri and I. Preston (2005): Imputing consumption in the PSID using food demand estimates from the CEX. The Institute for Fiscal Studies, WP 04/27.
- Brooks, S.P. and A. Gelman (1998): Alternative methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7: 434-455.
- Browning, M., Th.F. Crossley and G. Weber (2003): Asking consumption questions in general purpose surveys. *The Economic Journal*, 113(491): F540–F567.
- Crossley, T. and Pendakur, K. (2002): Consumption inequality. In: D. Green and J.R. Kesselman (eds.): Dimensions of inequality in Canada. Vancouver: UBC Press.
- Dagum, C. (1977): A new model of personal income distribution: Specification and estimation. *Economie Appliquée*, 30: 413-437.
- El Messlaki, F. (2010): Making use of multiple imputation to analyze heaped data. Master's thesis, University of Utrecht. (<http://igitur-archive.library.uu.nl/student-theses/2010-0705-200143/UUindex.html>).
- Fisk, P.R. (1961): The graduation of income distributions. *Econometrica*, 29(2): 171-185.
- Frick, J.R. and M.M. Grabka (2005): Item-non-response on income questions in panel surveys: Incidence, imputation and the impact on the income distribution. *Allgemeines Statistisches Archiv (ASTA)* 89(1): 49-61.
- Hall, R.E. (1978): Stochastic implications of the life cycle-permanent income hypothesis: Theory and evidence. *Journal of Political Economy*, 86: 971-987.
- Headey, B. (2008): Poverty is low consumption and low wealth, not just low income. *Social Indicators Research*, 89(1): 23-39.
- Heitjan, D.F. and Rubin, D.B. (1990): Inference from coarse data via multiple imputation with application to age heaping. *Journal of the American Statistical Association*, 85: 304-314.
- Jenkins, S.P. (2004). Fitting functional forms to distributions, using ml. Presentation at Second German Stata Users Group Meeting, Berlin. <http://www.stata.com/meeting/2german/Jenkins.pdf>
- Kraus, F., & Steiner, V. (1998). Modelling heaping effects in unemployment duration models: With an application to retrospective event data in the German Socio-Economic Panel. *Jahrbücher für Nationalökonomie und Statistik*, 217(5): 550-573.
- Leland, H.E. (1968): Saving and Uncertainty: The precautionary demand for saving. *Quarterly Journal of Economics*, 82: 465-473.
- Lillard, D.R., H. Bar and H. Wang (2008): Heap of Trouble? Accounting for mismatch bias in retrospectively reported data (with application to smoking cessation and (non) employment). Paper presented at the 8<sup>th</sup> International German Socio-Economic Panel User Conference, July 9-11 2008, Berlin.

- Little, R.J.A. and D.B. Rubin (2002): *Statistical analysis with missing data*. New York: John Wiley & Sons.
- McDonald, J.B. (1984). Some generalized functions for the size distribution of income. *Econometrica*, 52: 647-663.
- Modigliani, Franco (1966): The life cycle hypothesis of saving, the demand for wealth and the supply of capital. *Social Research*, 33(2): 160-217.
- Pudney, S. (2008): Heaping and leaping: Survey response behavior and the dynamics of self-reported consumption expenditure. Institute for Social and Economic Research, working paper No. 2008-09.
- Ringen, S. (1988): Direct and indirect measures of poverty. *Journal of Social Policy*, 17: 351-365.
- Royston, P. (2004): Multiple imputation of missing values. *The Stata Journal* 4(3): 227-241.
- Royston, P., I.R. White. (2011) Multiple imputation by chained equations (MICE): Implementation in Stata. *Journal of Statistical Software* 45(4): 1-20.
- Rubin, D.B. (1987): *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons.
- Singh, S.K. and G.S. Maddala (1976): A function for size distribution of incomes. *Econometrica*, 44(5): 963–970.
- Statistisches Bundesamt (2010): *Wirtschaftsrechnungen. Einkommens- und Verbrauchsstichprobe Aufwendungen privater Haushalte für den Privaten Konsum 2008*. Fachserie 15 Heft 5.
- Statistisches Bundesamt (2012): *Wirtschaftsrechnungen. Einkommens- und Verbrauchsstichprobe Einkommensverteilung in Deutschland 2008*. Fachserie 15 Heft 6.
- Stiglitz, J.E., A. Sen, J.-P. Fitoussi (2009): Report by the Commission on the Measurement of Economic Performance and Social Progress. (<http://www.stiglitz-sen-fitoussi.fr/en/index.htm>)
- Wang, H. and D.F. Heitjan (2008): Modeling heaping in self-reported cigarette counts. *Statistics in Medicine*, 27(19): 3789-3804
- 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.