EUROMOD WORKING PAPER SERIES

EM 02/26

Imputing Household Expenditure for Indirect Tax Simulation: Extending SWISSMOD Using Statistical Matching

Robin Anderl, Tanja Kirn, and Patrick Oschwald

May 2026

Imputing Household Expenditure for Indirect Tax Simulation: Extending SWISSMOD Using Statistical Matching[1]

Robin Anderl [a]

Tanja Kirn [b]

Patrick Oschwald [a]

[a] Albert-Ludwigs-Universität, Freiburg, Germany

[b] Universität Liechtenstein, Liechtenstein

Abstract

Indirect taxes such as value-added tax (VAT) and environmental levies play a growing role in fiscal and distributional policy in Switzerland, yet existing microsimulation tools lack the consumption information required to analyse their incidence. This paper closes this gap by extending SWISSMOD - the Swiss tax-benefit microsimulation model - through the imputation of detailed household expenditure from the Swiss Household Budget Survey (HBS) into Swiss SILC data. Building on the two-step econometric approach of Akoğuz et al. (2020), we estimate participation and conditional spending using probit and OLS models, followed by Mahalanobis distance matching to transfer observed expenditure structures at the most disaggregated level. We demonstrate how this methodology can be adapted and validated for a non-EU context characterised by distinctive survey design and consumption classifications. Validation using distributional tests, macro-level benchmarks, and income-ventile comparisons shows that the imputed data reliably reproduces key expenditure patterns, particularly for regularly consumed categories. The resulting enriched dataset enables comprehensive simulation of indirect tax reforms within SWISSMOD, providing a robust empirical foundation for evaluating VAT and environmental tax policies in Switzerland.

JEL codes: C15, C83, C81,H31

Keywords: Consumption data imputation, Statistical matching, Household Budget Survey (HBS), Switzerland, Microsimulation, Indirect taxation.

Corresponding author: Robin Anderl robin.anderl@vwl.uni-freiburg.de

[1] This paper is supported by FRIBIS at the University of Freiburg, and received funding from the Erasmus+ project ecoMOD (project number: 2023-1-LI01-KA220-HED-000157594). An earlier version of this study was presented at the European Meeting of the International Microsimulation Association 2025 at the University of Galway, Ireland. We appreciate the valuable comments and suggestions from the participants at this event. We also thank Bernhard Neumärker for his helpful advice. Special thanks go to Martha Brinkmann for excellent research assistance. SWISSMOD is based on EUROMOD model J1.0+. Originally maintained, developed and managed by the Institute for Social and Economic Research (ISER), since 2021 EUROMOD has been maintained, developed and managed by the Joint Research Centre (JRC) of the European Commission, in collaboration with Eurostat and national teams from the EU countries. We are indebted to the many people who have contributed to the development of EUROMOD. We use Swiss SILC data as well as HBS data, which have been made available by the Swiss Federal Statistical Office.

1. Introduction

Indirect taxes, particularly value-added tax (VAT) and environmental levies, constitute a central pillar of Switzerland's fiscal framework. These taxes generate substantial revenue and shape household consumption behaviour, yet they may also impose disproportionate burdens on lower-income households if not carefully designed. As a result, policymakers increasingly rely on microsimulation models to assess the distributional and welfare implications of indirect tax reforms across income groups, household types, and regions.

SWISSMOD, a static microsimulation model of the Swiss tax-and-transfer system (TTS) built on the EUROMOD platform, currently offers detailed simulations of direct taxes and transfers, capturing Switzerland's pronounced subnational heterogeneity, including cantonal variation in income taxation, wealth taxation, joint taxation rules, and social assistance (Kirn, Oschwald, & Anderl 2025). However, like other EUROMOD-based models relying on Swiss EU-SILC data, SWISSMOD lacks household-level consumption information. This represents a crucial limitation: VAT alone is expected to contribute more than 31% of federal revenue in 2026 (FFA 2025), and without reliable expenditure data, the model cannot simulate indirect taxes or evaluate environmental tax instruments. Addressing this gap is therefore essential for providing a comprehensive evidence base for fiscal policy analysis in Switzerland.

To overcome this limitation, we impute detailed consumption expenditure from the Swiss Household Budget Survey (HBS) into Swiss EU-SILC. Our approach follows the two-step econometric framework of Akoğuz et al. (2020), which models participation and conditional expenditure through probit and OLS regressions and then uses Mahalanobis distance matching to assign observed expenditure values from HBS to structurally similar households in SILC. Although subsequent refinements have been proposed by Dreoni, Serruys, Manso, Tudó-Ramírez, and Amores (2025), we deliberately adopt the original method to ensure methodological comparability with previous EUROMOD research and to establish a transparent, reproducible baseline for future extensions.

This paper makes three main contributions to the microsimulation literature. First, we extend SWISSMOD by imputing detailed household consumption expenditure into Swiss SILC, thereby enabling full distributional analysis of indirect taxes in a non-EU setting. This addresses a major limitation of existing Swiss microsimulation models, which currently capture income and wealth taxation, but cannot model VAT or environmental levies. Second, we adapt the widely used two-step imputation framework of Akoğuz et al. (2020) to the specific institutional and statistical environment of Switzerland, characterised by distinct survey structures, COICOP-based classifications, and strong subnational heterogeneity. In doing so, we document the harmonisation, estimation, and matching steps required for a transparent and reproducible application beyond EU countries. Third, we provide a comprehensive validation of the imputed consumption data using distributional comparisons, pseudo-R^2 metrics, Kolmogorov–Smirnov tests, and macro-level consistency checks. This extensive evaluation demonstrates that the method preserves key expenditure patterns and offers a reliable empirical basis for analysing indirect tax reforms, including VAT and carbon levy simulations.

Our results confirm that the approach is feasible and yields robust imputation outcomes for Switzerland. Expenditure categories with regular and frequent consumption - such as housing, utilities, and food - exhibit particularly strong consistency between observed and imputed values, while deviations are more pronounced for infrequent or durable expenditures. Overall, the findings establish a solid micro-data foundation for conducting indirect tax microsimulations in SWISSMOD and open avenues for future analyses of environmental taxation, VAT reforms, and distributional effects across Swiss households.

The remainder of the paper is structured as follows. Section 2 reviews the relevant literature. Section 3 describes the data environment and the harmonisation process. Section 4 outlines the imputation methodology. Section 5 presents the validation and results. Section 6 concludes.

2. Literature Review

The integration of consumption data into income-based microsimulation models has been a longstanding challenge in the field of tax-benefit analysis. Microsimulation models such as EUROMOD traditionally rely on EU-SILC data, which provides detailed socio-economic information but lacks household-level expenditure data. This limitation restricts the ability to simulate indirect taxes, which are increasingly relevant for fiscal and environmental policy.

Early approaches to bridging this gap focused on statistical matching techniques, which combine information from separate datasets based on shared characteristics. Serafino and Tonkin (2017) provide a comprehensive overview of these methods within the European context, emphasizing the importance of harmonization and variable comparability. Building on this foundation, Akoğuz et al. (2020) introduced a two-step econometric approach for imputing consumption expenditures into EU-SILC. Their methodology combines probit models to estimate the probabilitie of participation in consumption categories with OLS regressions determining the expenditure level, followed by Mahalanobis distance matching to transfer observed values on the most detailed level. This approach has been widely adopted in EUROMOD applications and validated across 18 EU countries. Subsequent research has refined these techniques. For example, Dreoni et al. (2025) propose enhancements to improve predictive accuracy and address issues related to infrequent expenditures and zero values.

Despite these advances, applications outside the EU remain scarce. Differences in survey design, household structures, and expenditure classifications pose challenges for transferring EU-based methodologies to non-EU contexts. Switzerland, with its strong subnational fiscal autonomy and Classification of Individual Consumption According to Purpose (COICOP) based expenditure data, offers a unique case for testing the adaptability of these methods. There is evidence that the aggregation of consumption categories in the SWISS HBS deviates from the standard structure when compared to other countries in EUROMOD, even though the Swiss HBS itself is based on COICOP classifications. It should also be noted that the COICOP classifications and household structures are largely, though not completely, harmonized among EU members

(Eurostat 2025a). Previous work on SWISSMOD (Kirn et al. 2025) has successfully integrated wealth and regional information but has not addressed the imputation of consumption data, leaving a critical gap for indirect tax simulation. Notably, similar statistical matching exercises have been conducted to enrich EU-SILC with net wealth information; for example, Donatiello, Toso, and Baldini (2025) combined Italy's SILC, Household Budget Survey, and Survey on Household Income and Wealth using distance-based matching to estimate the joint distribution of income, consumption, and wealth.

This paper contributes to the literature by adapting and validating the EUROMOD-based imputation methodology for Switzerland, assessing its feasibility and robustness, and providing a foundation for future extensions to other non-EU countries.

3. Data

The 2020 Swiss SILC dataset, compiled by the Federal Statistical Office (FSO), serves as the foundation for the initial version of SWISSMOD. Since EUROMOD relies on EU-SILC data, the Swiss SILC is the logical choice, ensuring consistency and international comparability - especially with regard to indicators like the at-risk-of-poverty threshold (European Commission 2025). The dataset is collected following Eurostat standards in both methodology and variable definitions. It also contains detailed information on socio-demographic factors as well as individual histories, for example with regard to the duration of employment (Kirn et al. 2025). What is missing is information on consumption expenditures of the households surveyed. Therefore, we use the information from the Household Budget Survey as the source for the data imputation. The HBS dataset, acquired for this project, includes four waves of data, each covering a three-year period. In addition to sufficient data in income and household structure, this dataset compiled by the Federal Statistical Office (FSO) contains information on household expenditures on a highly detailed level and is therefore a valuable source for the imputation of consumption data. The structure of the expenditure data is explained in Subsection 3.2.

Since the household samples for HBS and SILC surveys do not overlap, combining the datasets using common household identifiers is impossible. Without a common identifier, an econometrical matching process is required to combine the datasets.

Building on the existing literature, we conduct a statistical matching technique to impute the expenditure shares from the latest survey wave of HBS (2015-2017) into the SILC data collected in 2020 for Switzerland. The decision was made to use the most recent survey wave available. Older surveys are not taken into account, as only the most recent consumption patterns are to be estimated to the SILC data. Since, for example, a carbon tax has been in place in Switzerland since 2008 and has been gradually increased, corresponding adjustments in consumption are to be expected, which we would like to use to their full effect in the simulation (FOE 2024). The use of older survey waves could distort current consumption decisions. The remaining available data waves could be used in future to analyze changes in consumption patterns over time. Furthermore, a single survey wave already provides a sufficiently large sample of households, which is why enlargement is not necessary.

As part of the HBS, consumer spending was recorded for a period of one-month for each household surveyed. Over a three-year cycle, different household samples (around 277) were collected in successive months. Since the income data in the SILC set of 2020 refers to the previous year, it corresponds to the status of 2019. Due to the data originating from different years, an adjustment had to be made for inflation rates. The used uprating factor is calculated using the development of the harmonised index of consumer prices (HICP) with the Baseline of 100 in 2015. Since we do not know the precise date in the corresponding period on which a household was surveyed in the HBS data set, the 2015 baseline of 100 is used. Thereby household income as well as the expenditures in absolute terms in HBS are adjusted by a factor of 1.0141 in order to meet the average level in 2019 of 101,41 for all households included in the data set. We therefore assumed that all households were surveyed at the beginning of the survey period. While it could have been conceivable to treat all households as if they had been surveyed in mid-2016, this approach was not adopted because during this period the HCPI trend was frequently negative with an average index-value smaller than 100 (approximately 99,86) (FSO 2025a). Applying this index would therefore imply a nominal decrease in wages and therefore household income. However, such short-term deflationary episodes are not typically associated with immediate reductions in nominal wages due to wages being rigid. For this reason, the mid-2016 index was not used as an starting point for uprating.

3.1. *Data preparation*

The data preparation starts with identifying matching variables to concatenate. In the first step, we identified potential matching variables. These are variables that are part of both surveys and help to identify households that resemble each other, in other words the specific household characteristics. However, not all overlapping variables should be included as matching variables - only those that are relevant to the target variable and exhibit the possibility of similar distributions across the surveys (Serafino & Tonkin 2017). Given the differences in variable names, units, and the scope and detail of available information of the SILC and HBS dataset, an initial harmonization process is necessary before the imputation - meaning that the explanatory variables should be present in both data sets and have the same structure. This process followed the methodology outlined by Akoğuz et al. (2020). The procedure for the data adjustment is as follows: All relevant variables that occur together in both data sets must be identified. It was decided to use 16 variables. The selection of explanatory variables was based on the comparability of their characteristics across both datasets and thus whether the overall structure of the data sets is comparable. Furthermore, the explanatory variables are chosen so that they are as comparable as possible with the 16 explanatory variables in the approach of Akoğuz et al. (2020) since we want to stay as close as possible to this original method in order to show that this approach has worked well. The explanatory variables are listed in Table 1 below. This shows the labels for the selected variables in HBS and SILC, a brief explanation of how the variable was adjusted, and the distinction between categorical and continuous.

Table 1. Overview of explanatory variables

Explanatory Variable (Original)	SILC	HBS	Variable Transformation	New Variable	Type
Disposable Income	HY020	VerfuegbaresEinkommen08	The variable was extrapolated to the annual value. The HBS value was adjusted for inflation to account for the time difference.	hy020_2	Continuous
Civil Status	PB190	Zivilstand03	Distinction in single / never married, married, widowed, divorced / separated. Applies to whole household; only reference person is used.	pb190_2	Categorical
Gender	PB150	Geschlecht98	Counts number of male members (age >= 14) and assigns it to the reference person.	n_gender_male	Continuous
Region	DB040	Grossregion01	Distinguishes between seven main regions of Switzerland: Lake Geneva Espace Mittelland Northwestern Switzerland Zurich Eastern Switzerland Central Switzerland Ticino	db040_2	Categorical
Employed Status	PL031	Erwerbsstatus05	Counts number of household members by employment status and assigns to reference person. Persons in education are only counted if they are over 14 years of age.	n_employed n_self_employed n_in_education n_pension n_unemployed_other	Continuous
Housing Situation	HX070	Mieterhaushalt05	Applies to whole household; only reference person is used for comparison.	hx070_2	Categorical
Age Groups	RX010	Lebensalter98	Counts household members by age group and assigns to reference person.	n_age_0_14 n_age_15_29 n_age_30_44 n_age_45_59 n_age_60_upwards	Continuous
Citizenship	NATIO_CL2	Nationalitaet01	Counts household members without Swiss citizenship and assigns to reference person.	n_nationality_other	Continuous

Unfortunately, not all variables that could possibly be decisive for consumption decisions could be used for the imputation, as in some cases the information was only available in one of the data sets. Therefore, unlike in the original approach, we have some information missing in our datasets. This concerns the status of the number of disabled household members, the status of a household as a farmer and information about the educational level.

To enhance the predictive accuracy of the regression-based imputation, we extended the original set of explanatory variables proposed by Akoğuz et al. (2020). In particular, we included the civil status, the housing situation (tenant/owner), and the number of self-employed household members as additional independent variables. Since the consumption expenditures are only recorded at the household level and cannot be allocated to individuals, all explanatory variables must be aggregated on the household level as well, therefore individual-level information must also be transformed to the household level - which in this case implies that all the information is assigned to the corresponding reference person. Further, some explanatory variables cannot be allocated at the individual level anyway, as in HBS, for instance, the total sum of incomes is calculated for certain types of income in multi-person households and allocated to all household members. Therefore, it is not possible to differentiate between the adults in terms of their contribution share to the household income if both belong to the same employment status.

In a second step, we therefore harmonized the selected variables by converting information from the individual level to the household level in both datasets. In doing so, we proceed as proposed by Akoğuz et al. (2020). As the definition of 'household' in both datasets is based on EUROSTAT's household concept - the structure and defintion of the household head - no unit harmonization is needed. Nevertheless we must convert individual characteristics into household characteristics. If, for example, a person has a certain individual characteristic like the employment status, the household they belong to is characterized by that characteristic of this certain person. Therefore, the number of people in the household belonging to a specific group was recorded (e.g., the number of household members that are of age between 30 and 44).

3.2. *HBS expenditure data*

The Swiss Household Budget Survey (HBS) provides detailed information on consumer spending by private households. It is carried out by the Federal Statistical Office (FSO) and serves as a central data source for analyzing the cost of living, consumer behavior and the social and economic situation of households. The survey contains a total of 9,955 households with 22,315 individuals. As mentioned earlier, the four waves of the HBS provide expenditure data on the household-level. Each wave covers a three-year reference period, with expenditures being recorded at the specific month of observation. The available waves span the periods 2006–2008, 2009–2011, 2012–2014, and 2015–2017. This would allow for a later analysis of trends and changes in Swiss consumer behaviour.

Regarding the recording of the expenditures, a distinction is made between goods and services with regard to the exact recording period. For most items, only expenditures for the observed period (one month) was meant to be documented. For others, however, a longer period is considered. For instance, goods and services for which more than 300 Swiss francs were paid per unit, are observed over five month period. A unit can be either a single item or a group of items. The purchase of a new vehicle is even recorded over a period of 12 months.

These expenditures are then recorded for the respective month and later reported as a monthly average in the HBS. This value is therefore not attributed to the month as if this expenditure were incurred every month in this absolute amount. Instead, the monthly average value calculated from the total annual expenditure is reported. This information is important later on in order to better understand the role of the infrequent expenditure problem. This information can be found in the FSO survey documents, which are available on request.

Initially, it is sufficient to use only one data set as the basis for the imputation of consumption data. For the purpose of this analysis, so far only the most recent wave (2015–2017) is used, as a closer temporal proximity between the donor (HBS) and recipient (SILC) datasets increases the plausibility of the imputed consumption estimates. Furthermore the latest HBS date set ensures the most recent trends in consumption patterns, which are highly relevant with regard to the final goal of simulating the swiss carbon tax. Household consumption expenditures are organized following the COICOP classification, which distinguishes 12 main categories (e.g. housing, food and non-alcoholic beverages, and transportation). In addition, the dataset contains further superordinate expenditure categories that are not considered part of the consumer spending category, such as compulsory transfers or supplementary insurance contributions.

Each main category is further divided into more detailed subcategories across four hierarchical levels of aggregation (in accordance with the classification of individual consumption by purpose). At the highest level under the main category (level 2), expenditures are aggregated into broad groups, while at the most detailed level (level 5), they correspond to specific goods or services. For example, within the "Food and non-alcoholic beverages" category, level 2 might include all food-related spending, level 3 could distinguish between food and drinks, level 4 between describes types of food (e.g. Bread and Cereal Products), and level 5 could refer to specific items like 'rice'.

4. Imputation Methodology: Two-Step Econometric Approach and Mahalanobis Matching

This section outlines the imputation procedure used to transfer detailed consumption information from HBS to EU-SILC. Following Akoğuz et al. (2020), we first estimate a probit model based on the selected independent variables which gives us an estimate of the probability that a household will have positive expenditure in a given category. Then a standard continuous regression equation models the amount of the dependent variable (expenditure) as a function of household characteristics and actual – not estimated – positive expenditure, accordingly, households with zero consumption are not included in this regression. Using this methodology, consumption expenditure for households in the HBS is first estimated, and then consumption expenditure for SILC households is estimated on the basis of the coefficients of the explanatory variables generated in the process. To ensure that only variables with an "appropriate" fit accuracy are included in the Malahanobis distance function, a pseudo-R^2 value is calculated, which serves as a measure of how well the model (which includes both the probit and OLS estimates) can explain the actual observed output behavior, i.e., how much of the variance in the dependent variable is covered by the model.

The Mahalanobis distance function is then applied to identify households in the source and recipient datasets whose consumption behaviour should match. Between these households, the observed expenditure values of the source household are then assigned as imputed values at the most detailed level as expenditures of the recipient household for the corresponding goods. Since this approach uses actual observed values instead of estimated values, it enables effective imputation of values at the finest available level of aggregation and preserves real consumption behaviour.

Figure 1 illustrates the overall imputation process, starting from data harmonization and progressing through the two-step econometric estimation (probit and OLS), combination of predicted expenditure shares, Mahalanobis distance matching, and finally the transfer of observed HBS expenditures to EU-SILC households.

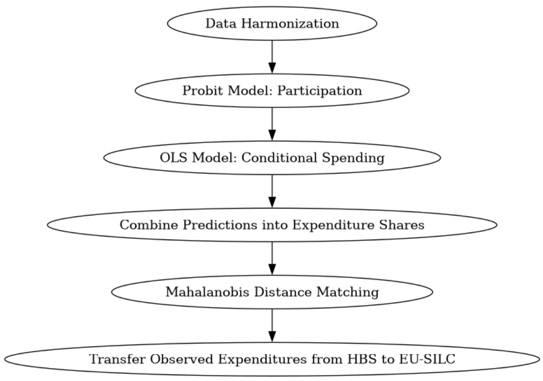

Figure 1. Overview of the imputation process: (1) Data Harmonization, (2) Probit Model for Participation, (3) OLS Model for Conditional Spending, (4) Combine Predictions into Expenditure Shares, (5) Mahalanobis Distance Matching, (6) Transfer Observed Expenditures from HBS to EU-SILC.
Source: Authors' own illustration.

The individual steps of this procedure are described in detail below.

In the initial step, the expenditure share w_{shi} of household h's on a good i in the HBS dataset (indexed by s) is computed. This share is calculated as the ratio of the expenditure e_{shi} on the good i to the disposable income y_{sh}:

$$w_{shi} = \frac{e_{shi}}{y_{sh}}, i \in \mathcal{N} \tag{1}$$

where $\mathcal{N}$ represents the set of indices for goods at the most detailed level.

In the second step, we summarize the income shares of specific expenditures into broad superordinate categories. According to Akoğuz et al. (2020), these categories should be large enough to mitigate the problem of rare expenditures, i.e., expenditures that are reported as zero values only because they were not incurred in the corresponding observation month. Nevertheless, they should remain small enough so that the selected household characteristics can adequately explain the differences in expenditures.

The consumption categories are labelled by $A, B, \ldots$. Accordingly, each index $A, B, \ldots$ represents a non-empty, mutually exclusive subset of $\mathcal{N}$, with $\mathcal{N}$ encompassing all individual goods.

The income share of category X for $X = A, B, \ldots$, denoted as W_{shX}, is the obtained as the sum of the expenditure shares of the goods contained in that category:

$$W_{shX} \equiv \sum_{i \in \mathcal{N}X} w_{shi} \tag{2}$$

The appendix contains a summary table of all selected expenditure categories. The individual categories are aggregated directly to form the corresponding expenditure category used in the subsequent regression analysis. Since the structure of the data sets does not match exactly, the composition of the various aggregation levels had to be adjusted individually. In addition, the categories have been expanded to include expenditures related to secondary homes.

In Akoğuz et al. (2020), some issues with the data are mentioned, which led to individual adjustments to the expenditure values for the consumer goods categories. These problems are particularly evident in discrepancies between data at different levels of aggregation, meaning that in some cases the totals for higher-level categories exceed or fall short of the sum of their subcategories. However, our own review of the available Swiss HBS data did not reveal any such problems. Therefore, in our case, no adjustments had to be made with regard to expenditure. But beside that some of the broad expenditure categories may still contain a significant number of observations with zero expenditures. This is for instance the case for the expenditure category "new cars" which is a sub-category of "vehicles". Here only a small fraction of Households

(4.26 %) has positive values.

For this reason, the next step is to apply a probit model to identify these zero values in the corresponding households as “true zeros,” i.e., if they are not considered to be the result of infrequent expenditures. This is an assumption made by the authors in order not to complicate the procedure. However, since the actual values are transferred at the end of the approach, all zero values are ultimately transferred again. If we were to assume that many of them are due to the problem of infrequent expenditures, they would have to be replaced by positive values, which in turn would have required the creation of a synthetic data set, which is not desirable for various reasons.

Furthermore, the assumption of true zero values is justified for at least some goods due to the survey methods used by the FSO. As described in section 2.3, the observation period is extended for larger purchases, which is why these irregular expenditures are partially covered. Nevertheless, some categorical zero values remain questionable. The category “clothing and shoes” is only recorded on a monthly basis and accounts for 20% of households with zero expenditure. However, given that clothing is a basic necessity, it is not plausible that 20% of households never have consumer expenditure on clothing. This must nevertheless be accepted within this methodology.

The likelihood that a household exhibits positive expenditures to the aggregate category of commodities $X(X = A, B, ...)$ is modeled by using a probit model. This model uses the common variables from the source and recipient data sets as predictors. The formal specification is:

$$Pr(W_{shX} > 0) = 1 - \Phi(-\gamma_X' x_{sh}) = \Phi(\gamma_X' x_{sh}) \tag{3}$$

Here, Φ refers to the cumulative distribution function of the standard normal distribution, x_{sh} is the vector of explanatory variables associated with household h in the source dataset s, and γ_X represents the parameter vector to be estimated.

Subsequently, a standard continuous regression model is employed to estimate the relationship between the income shares of the selected expenditure categories and the explanatory variables, but this is only applied to those households that exhibit actually positive expenditures in the respective categories.

$$W_{shX} = \beta'_X x_{sh} + \epsilon_{hX}, \quad \text{for} \quad W_{shX} > 0 \tag{4}$$

Since the focus of the analysis is restricted to correlational relationships, no correction for sample selection is applied. The model is estimated using the source dataset, with the estimated coefficients represented by hats: $\hat{\gamma}_X$ for the probit models and $\hat{\beta}_X$ for the linear regression models.

In the fourth step, the estimated coefficients are utilized to predict the fitted values for the income shares of expenditures on the broad categories $A, B, \ldots$, for all households in both the source and the recipient datasets, which are indexed by s and r, respectively. The fitted values are denoted by $\hat{W}_{dhX}$ and defined as follows:

$$\hat{W}_{dhX} = \Phi(\hat{\gamma}'_X x_{dh})\hat{\beta}'_X x_{dh} \quad \text{for} \quad d = s, r \tag{5}$$

The first factor on the right-hand side (RHS) represents the estimated probability, as derived from the probit model that household h has positive expenditure on the aggregate category X. The second term represents the estimated income share of the ordinary continuous regression model, which reflects the proportion of income thate household h allocates to category X, conditional on this share being positive.

This two-part approach using the probit model and ordinary least squares regression, followed by a multiplicative link, is significant for the following reasons. Since many observations in the data set correspond to the value zero, OLS regression alone is problematic. If the assumption that the zero values correspond to zero consumption is valid, this leads to the conclusion that there is a two-stage process with two different economic decision stages. (i) The participation decision, i.e., whether a household consumes at all, and (ii) the level of consumption, i.e., how much is spent under the condition of participation. OLS alone can lead to biased coefficient estimates because this approach cannot distinguish between these two stages.

Since OLS assumes a single linear, continuous relationship between the dependent variable and the explanatory variables across all observations, the approach attempts to estimate a single linear relationship that minimizes the average squared deviation across all observations. In the case of numerous expenditures with zero values, this pulls the estimated correlation down. At the same time, positive expenditures are attributed to households that actually consumed nothing, while the expenditure value for households that actually consume is underestimated.

It should be noted that the effect generated by zero values within a sole OLS approach does not necessarily worsen the process of estimating the linear relationship. As long as the same linear assumptions apply consistently to both donor and recipient household data, OLS alone can preserve the true relationship. However, the OLS stage only captures average consumption levels regarding the precise expenditure values conditional on the independent variables and unable to differentiate between genuine non-participation and low positive values. The probit stage therefore offers considerable added value by explicitly modeling the participation decision, i.e., whether a household consumes at all, thus capturing the heterogeneity of the reasons for zeros that cannot be distinguished with OLS alone. By separating participation and intensity decisions, the two-part model preserves the distinction between consuming and non-consuming households, ensures greater variability in predicted participation, and prevents the misallocation of positive consumption to households that would not actually consume. Therefore, this approach provides a more realistic representation of consumption patterns at the household level and mitigates potential systematic downward bias in the estimation.

To solve this problem, a two-stage modeling approach can be used. First, a binary selection model (e.g., Probit) estimates the probability that a household will have consumption expenditure greater than zero, i.e., make a positive consumption decision. Then, for households with positive consumption in the actual data set, the amount spent is modeled using OLS. In the final step, the OLS results are "weighted" using the estimated probabilities of positive consumption. This approach takes into account the actual zero values and at the same time enables an accurate estimation of the nonlinear relationships between household characteristics and consumption levels.

To assess the explanatory performance of the estimated two-step model with respect to observed household expenditure, we calculate a pseudo-R^2:

$$pseudo - R^2(X) = 1 - \frac{\sum_h (W_{shX} - \hat{W}_{shX})^2}{\sum_h (W_{shX} - \bar{W}_{sX})^2} \tag{6}$$

Here $\bar{W}_{sX} = \sum_h W_{shX}/H_s$, with H_s represents the number of households in the source dataset. All summations are taken over the full set of households. Consequently, the pseudo–R^2 of the fitted expenditures serves as a measure of how well the independent variables can explain the variation in the specific expenditure categories and therefore the quality of representation regarding the observed expenditures in HBS (encompassing both the probit and OLS model). Then a distance function was builded based on the adjusted values. This ensures that no estimated variables with low explanatory power are included in the distance function. Therefore, only the adjusted values of categories with a pseudo-R^2 value of 0.1 are retained. This means that only these categories are included in the subsequent matching process. This procedure is intended to prevent households from being compared on the basis of actual behavior that cannot be adequately explained. This results in less variance in consumption patterns being taken into account, as the variables with higher explanatory power lose influence due to the equal weighting of all variables included - this could lead to more inappropriate matches. In addition, lower thresholds were also tested in order to include more categories, but this led to generally poorer results, which is why we adhere to the specifications from Akoğuz et al. (2020) and Dreoni et al. (2025).

In the next formal step, as described, the adjusted expenditure shares above the 0.1 threshold are used as input variables for the Mahalanobis distance function. In order to match each household in the recipient dataset with a household in the source dataset, the household in the source dataset that has the smallest distance to a specific household in the recipient dataset is identified as the donor household (Mahalanobis 2018). The household h in the source dataset is then paired with the closest household g in the recipient dataset. However, this does not mean that exactly one donor household is always assigned to one recipient household. This is obviously not possible due to the different number of households. Instead, the most suitable HBS households are assigned to the SILC households, whereby an HBS household can also act as a multiple donor.

Let $\hat{W}_{dh} \equiv (\hat{W}_{dhA}, \hat{W}_{dhB}, \ldots)$ denote the vector of the chosen estimated expenditure shares used as input for the distance calculation, where $(d = s, r)$ refers to the source (s) and recipient (r) datasets, respectively. The definition of distance between a donor household h and the recipient household g within the Mahalanobis methodology is as follows:

$$\text{dist}(h, g) = d(\hat{W}_{rg}, \hat{W}_{sh}) = \sqrt{(\hat{W}_{rg} - \hat{W}_{sh})' \, \Sigma^{-1} \, (\hat{W}_{rg} - \hat{W}_{sh})}, \tag{7}$$

Σ denotes the variance–covariance matrix of the share vector $\hat{W}$, computed using observations from both the source and recipient datasets.

The observed expenditure values are then transferred to the source budget at the lowest level of goods aggregation. In theory, a matching procedure could also have been applied directly without taking the detour via the consumption expenditure estimate using the regression approach. However, this is helpful in revealing information about the relationship between household characteristics (the explanatory variables) and expenditures (the dependent variables) within the data set, which would be overlooked if only the hot deck matching (HDM) method were used. Direct matching without prior two-stage regression would be possible in principle, but would be based exclusively on observable household characteristics. Such an approach would neglect systematic correlations between these characteristics and specific consumption patterns. Therefore, the first step is to explicitly model the relationship between household characteristics and consumption behavior. In this way, household characteristics are already taken into account in the estimated values, and at the same time, the matching is additionally expanded to include information on how specific combinations of characteristics lead to different consumption decisions. Purely distance-based matching, on the other hand, is not able to give greater weight in the distance function to those household characteristics that are more important for consumption patterns.

The limitations of the approach mentioned in Akoğuz et al. (2020) refer to the exclusion of households with an income of less than zero to an income of 100 euros in the data set. In our case, the number of households with a disposable income of less than or equal to zero accounts for 0.8% in the HBS and 0.1% in the SILC. These households were excluded from the estimation process because the use of the logarithm for the explanatory variable “disposable income” does not allow for negative incomes - beside that, no meaningful income shares can be determined. In our application, however, no households with a low income between zero and 100 had to be excluded, as there are no such households in either dataset.

5. Results of the Imputation Approach

This chapter analyzes the validity and robustness of the final results obtained. First, descriptive statistics are used to provide an overview of the comparability of the two data sets and to indicate any structural differences. This is followed by an overall assessment of the imputation process. In particular, it examines and demonstrates whether the imputed values are economically plausible and whether the consumption patterns observed within the HBS could be retained during imputation and are thus reflected in the SILC data set. This is accompanied by a brief evaluation of whether outliers can be explained by structural differences between the two data sources. In addition, the results are compared with external reference data.

5.1. *Descriptive Overview of HBS and SILC Data*

It should be noted in advance that disposable household income was adjusted at the household level using the OECD equivalence scale. According to this scale, the first adult is assigned a weight of 1, each additional adult a weight of 0.5, and each child a weight of 0.3. As already mentioned, the income data was also logarithmically transformed before the regressions were conducted.

Table 2 below provides an overview of the most important statistical indicators. It is clear that SILC households are characterized by a slightly higher average income. The median income, on the other hand, is very close and differs by only a few hundred Swiss francs. Household incomes in SILC also vary more widely. Furthermore, the data sets are relatively similar in terms of demographic characteristics and age group distribution. However, there is a slightly higher proportion of households with older people in SILC and a higher proportion of households with younger people in HBS, which may be one reason for the differences in average income.

Table 2. Descriptive Statistics: Comparison of HBS and SILC Datasets- General Overview

Variables	**HBS**	**SILC**
Number of Households	9,788	8,141
Percentage of People by Age		
0–14 years	17.79 %	15.39 %
15–29 years	13.53 %	14.95 %
30–44 years	21.17 %	19.15 %
45–59 years	23.58 %	22.47 %
60+ years	23.90 %	28.04 %
Total Persons	22,070	18,191
Average Persons per Household	2.25	2.23
Mean Equivalence-Weighted Income	56,661.16	58,511.09
Median Equivalence-Weighted Income	50,776.64	50,287.93
Standard Deviation	34,371.75	46,896.47
Variance	1.18e+09	2.20e+09
Skewness	7.9827	11.87385
Kurtosis	176.1186	289.479
Tenant Rate (%)	54.9	53.3

Source: HBS dataset (2015-2017) & EU-SILC dataset 2020

Table 3 shows the weighted disposable household incomes sorted by deciles. It shows that, starting with the 7th decile, the SILC incomes are on average higher than those of the HBS households, and that the gap is increasing. The gap is greatest in the last decile, which explains the higher variance and standard deviation in SILC, since at the same time the income amount in SILC is below the HBS value in the lowest decile.

Table 3. Descriptive Statistics: Household Income by Decile

Variable	1.	2.	3.	4.	5.	6.
HBS Mean	20,874	30,567	36,926	42,557	47,942	53,853
SILC Mean	18,930	29,378	35,643	41,333	47,184	53,616
HBS Median	21,775	30,678	37,021	42,628	47,889	53,924
SILC Median	19,931	29,536	35,644	41,374	47,213	53,524

Variable	7.	8.	9.	10.	Total
HBS Mean	60,488	68,750	81,488	123,225	56,661
SILC Mean	61,251	70,538	84,735	142,553	58,511
HBS Median	60,416	68,483	80,879	107,864	50,777
SILC Median	61,130	70,332	84,126	118,184	50,288

Source: HBS dataset (2015-2017) & EU-SILC dataset 2020

As shown in Figure 2 the graphic shows the comparison of the weighted household incomes using a density function, clearly showing that the relative distribution within the two variables is very similar, with only a slight deviation. The basic distribution follows the same pattern.

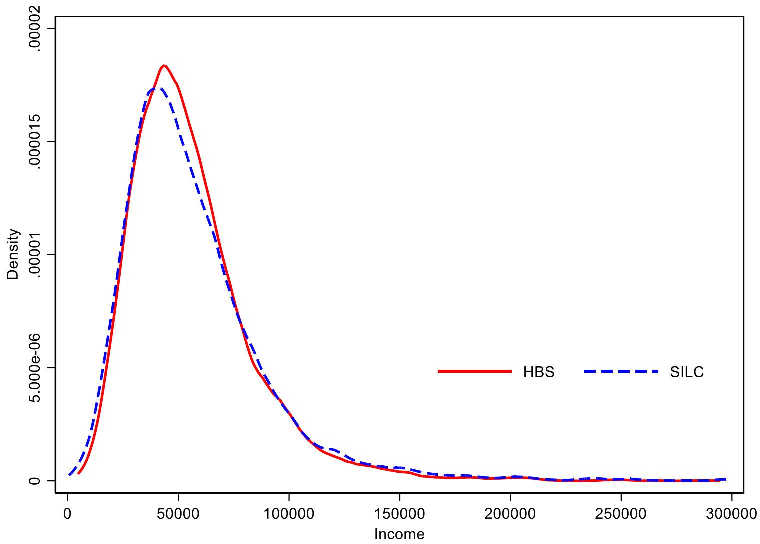

Figure 2. Density Function Disposable Income

Source: HBS 2015-2017; EU-SILC 2020; authors' illustration.

Finally, Table 4 shows once again the distribution of weighted average incomes, this time sorted by major Swiss regions and supplemented by the number of households having their main residence in the corresponding region. There is a general correlation between average incomes, which rise or fall simultaneously depending on the region, with the greatest deviation between the data in Central Switzerland. The number of households surveyed, on the other hand, varies more significantly and differs markedly between Central Switzerland and Ticino.

Table 4. Descriptive Statistics: Household Income by Region

Variable	Geneva region	Swiss Plateau	Northwestern -CH	Zurich region	Eastern -CH	Central -CH	Ticino
HBS Mean	57,428	51,994	58,225	66,029	54,112	59,699	48,698
SILC Mean	56,018	53,072	59,679	67,760	56,779	65,102	47,281
HBS Median	50,340	48,269	53,842	57,693	50,108	52,289	43,917
SILC Median	49,226	47,268	50,523	58,025	48,587	55,618	41,999
N Households	1,558	2,116	1,137	1,650	1,243	1,019	1,065
N Households	1,504	1,893	1,116	1,458	1,084	748	338

Source: HBS dataset (2015-2017) & EU-SILC dataset 2020

As can be seen from the descriptive statistics, the composition of households in both data sources is largely similar. The average equivalence-weighted income, age distribution, and regional distribution of households suggest that the data sets are generally suitable for imputation procedures. For reasons of transparency, these summary statistics are presented here to provide an indication of why the imputation procedure may not have been entirely accurate. However, the main positive results of the analysis are presented below.

5.2. *Data Cleaning and Pre-Processing Steps*

The discrepancies between the number of households specified in Chapter 5.1. and the 9,955 households initially mentioned within the HBS sample are due to further measures taken to adjust the data basis. The number of households was ultimately reduced to 9,788 observations, which were used for the two-stage estimation process. This reduction is primarily due to households with a disposable income of zero or less, which, as explained above, cannot be included in the estimation.

In addition, households with implausibly high expenditure shares for certain consumption categories - defined as values exceeding a threshold of 1.5 - were excluded. This criterion was applied to ensure that extreme or unrealistic expenditure figures did not distort the estimation results. For example, a household in the HBS is characterized by a rent-to-income ratio of 19, which means that its rental expenditure in the month observed is nineteen times its disposable income - clearly an extremely implausible expenditure value. The aim of the threshold is to set it at a level that allows for the inclusion of households whose expenditure may exceed their monthly income due to, for example, borrowing, asset sales, or other intertemporal mechanisms, while at the same time filtering out obvious errors in reporting or documentation. As a result, a total of 80 households were excluded due to negative income and a further 87 due to implausible spending habits. The threshold of 1.5 was chosen because it improved the overall R^2 values. Higher thresholds, on the other hand, noticeably worsened the model estimate. At the same time, the majority of the sample remains intact due to the selected threshold. The exclusion rate of 87 households due to spending anomalies is comparatively low.

It is important to note that particularly high expenditures observed during the survey month are recorded as monthly values - therefore the total sum was converted into monthly values. Consequently, these total amounts should not be interpreted as representative monthly averages. Since those values are now - due to the conversion - monthly values, they should be more likely to not exceed the budget constraint every month. It should be noted that exceptionally high expenditures reported during the survey month are recorded in the HBS as monthly values by converting the observed totals into the monthly accordingly. This implies for our analysis that some unusually large, one-off expenditures were retained rather than excluded, as their conversion to a monthly equivalent should prevent a potential direct violation of the household budget constraint. Substantial expenditures financed through loans (which are not recorded explicitly as income in the datasets) are typically repaid in installments and therefore do not permanently exceed household income, therefore sustainable financing is guaranteed and there appears no exclusion by our approach. For clarity, within the "housing and rental" category, 34 households were excluded based on the 1.5 threshold, while 355 households (3.59% of the original number of households with positive income) fall within the range of 0.5 to 1.5. The vast majority of households thus report expenditure shares below 50%. However, it is also possible that some overlapping is present with other categories in which values above 1.5 also occur.

With regard to the SILC data set, it should be noted that this has been reduced from 8.156 to 8.141 due to negative incomes.

5.3. *Assessing Explanatory Power: Pseudo* $-R^2$ *Values*

The results of the pseudo R^2 value calculation are presented and evaluated here in the same way as conducted by Akoğuz et al. (2020). Table 5 below shows the results for each of the constructed consumption category after all data adjustments and the two-stage regression approach had been carried out

Based on these R^2 values, the threshold value of 0.1 is used to decide which categories have sufficient explanatory power to approximately reflect the actual variation in the specific expenditure category. Therefore, the threshold value serves as a limit for the category whose variance explanation is considered insufficient. This approach is

Table 5. Pseudo-R^2 Values by Consumption Category

Variable	**Pseudo-R^2**
Housing and rental	0.59
Food & non alcoholic beverages	0.39
Utilities	0.33
Communication	0.30
Health & Care	0.28
Insurance	0.18
Personal care	0.09
Culture & leisure	0.08
Private transport	0.07
Tobacco	0.07
Housing goods & services	0.06
Public transport	0.05
Other	0.05
Restaurants	0.04
Alcoholic beverages	0.04
Education	0.04
Clothing and personal items	0.03
Housing durables	0.03
Utilities secondary residence	0.01
Housing and rental secondary residence	0.01
Traveling and holidays	-0.56
Vehicles	-3.70

Source: Computed R^2-values are based on the HBS dataset (2015-2017)

consistent with the approach in Dreoni et al. (2025) and Akoğuz et al. (2020). According to the results, six categories were used for the further process of distance matching: *housing costs, food and non-alcoholic beverages, utilities, health and care, communication, and insurance.* Expenditures in these specific categories alone accounts for 46.9% of total consumer spending. This exclusively comprises expenditure that occurs regularly in fixed cycles and, on average, accounts for a large proportion of disposable household income at just under half. The values were generated taking population weights into account. It should be noted here that parts of the category-specific expenditures were excluded from the calculation because they are not part of pure consumer spending as defined in the HBS-Structure, but are, for example, part of mandatory transfer expenditures. This applies to items A31, A33, A41, A42, and A80, which can be found in Table 8 and Table 9 in the appendix. Otherwise, the comparison would have to be made with total gross income, since ultimately all income is spent within one of the categories and it could be misleading.

A direct comparison with the results regarding the $R^2 - values$ from Akoğuz et al. (2020) or Dreoni et al. (2025) is not possible due to differences in the underlying data sets. Nevertheless, the overall pattern appears to be largely comparable in terms of the number of categories identified as relevant, as well as the type of category. Akoğuz et al. (2020) estimated models for a total of 18 countries. This resulted in four consumption categories achieving $R^2 - values$ values greater than or equal to 0.1 in more than half of the cases. This indicates that these categories were selected at least nine times, whereas two other categories could be selected eight times. On average, there were 6.28 categories per country with $R^2 - values$ above 0.1. In the study of the JRC, which again covered 22 countries, only three categories exceeded the threshold in more than half of the cases. Against this background, the identification of six relevant categories in our analysis suggests that our results are largely consistent with the results of previous studies.

When comparing the categories that exhibit strong or weak predictive performance, a clear pattern emerges. Food and non-alcoholic beverages, utilities, and communications also appear among the consistently relevant categories in both Akoğuz et al. (2020) and the Dreoni et al. (2025). Actual rentals for housing, by contrast, are predominantly selected only in the former. At the same time, there are overlaps between the categories that perform poorly in our analysis and those with low values in the literature. In particular, spending on vehicles and travel/vacation has very low R^2 values. This is most likely due to the extremely low number of positive spending observations in these categories. As a result, the probit model was unable to estimate realistic probabilities of participation in consumption and, in our case, assigned a probability of one to almost all households. When these predicted participation values are combined with the results of the OLS regression and the estimated coefficients are used as the basis for explaining the real data variance, the resulting $R^2 - values$ are correspondingly low.

As mentioned in Section 3.1, additional explanatory variables were added. This is because we had to omit three explanatory variables that were included in Akoğuz et al. (2020), while at the same time wanting to integrate additional variables in order to achieve greater explanatory power within the estimated output categories. A comparison shows that the $R^2 - values$ increased when these additional variables were included. These are listed in Table 6.

Table 6. Comparison of R^2-Values for Selected Consumption Categories

Category	Pseudo-R^2 (with add. var.)	Pseudo-R^2 (without add. var.)	Difference
Housing and Rental	0.5898	0.3716	+ 0.2182
Food & non-alcoholic Beverages	0.3914	0.3880	+ 0.0034
Utilities	0.3272	0.2688	+ 0.0584
Communication	0.3010	0.2980	+ 0.0030
Health & Care	0.2759	0.2642	+ 0.0117
Insurance	0.1769	0.1104	+ 0.0665

Source: Computed R^2-values are based on the HBS dataset (2015-2017)

Higher $R^2 - values$ values were observed for all categories that exceeded the threshold of 0.1, as well as for all but one of the remaining categories. In particular, the value for the "Housing and rental" category has increased. This is probably due to the addition of the explanatory variables 'Homeowner' or "Tenant". Nevertheless, the total number of categories exceeding the threshold remained unchanged.

5.4. *Donor Household Allocation in Mahalanobis Matching*

This subsection discusses the results of the Mahalanobis distance matching. In the course of this, a total of 4,355 households from the HBS dataset were identified as donors for the 8,141 recipient households in the SILC dataset. This corresponds to 44.5% of all HBS households (9,788). On average, each selected HBS household serves as a donor for approximately 1.87 SILC households. The distribution of usage frequencies shows that the majority of donor households (58.32%) were assigned to exactly one recipient household, while a smaller number of households acted as donors multiple times. The maximum donation frequency is 41 for a household. These statistics show that the imputation is based on a large group of donor households, which should ensure sufficient variance in the expenditure structures. With this number of imputed households, it can be assumed that all SILC households receive imputed values based on the closest possible match with an HBS household. At the same time, no SILC household was left without assigned consumption values.

Figure 3 below shows the exact distribution of the frequency of usage of households identified as donors.

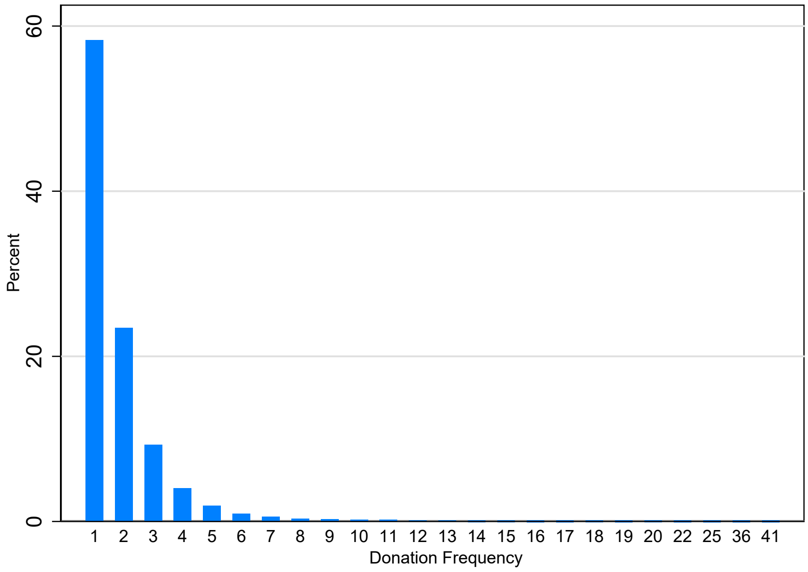

Figure 3. Distribution of donation frequency
Source: HBS 2015-2017; EU-SILC 2020; authors' illustration.

5.5. *Validation of Imputed Consumption Patterns*

This subchapter examines the final results of the estimated consumption patterns in the SILC data set obtained through the data imputation. To this end, graphs are presented like in Akoğuz et al. (2020) to compare the relationship between income shares in specific consumption categories and disposable income. To this purpose, the average income shares per income ventile are compared between the estimated results in SILC and the actual observed values in HBS.[1]

As shown in Section 5.1, the two data sets are already largely consistent in terms of socio-demographic characteristics. In addition, when dividing households into income ventiles, the population weighting variables available in the data sets were applied in advance to improve comparability, as in Akoğuz et al. (2020). This leads initially to a different number of surveyed households per ventile, which is why the average values per ventile in terms of weighted disposable income and income shares for consumption were generated using the population weights again after the division into ventiles has been carried out. In this way, the most accurate mean values possible should be

[1] A ventile is a statistical grouping that divides a distribution into 20 equally sized groups. It is similar to a decile (10 groups) or percentile (100 groups), but more granular than deciles and less granular than percentiles.

calculated, as different weights apply to the households within a ventile - otherwise the mean value would be distorted.

The following section presents the income-consumption relationship for total household expenditure and the results for two of the aggregate consumption categories.

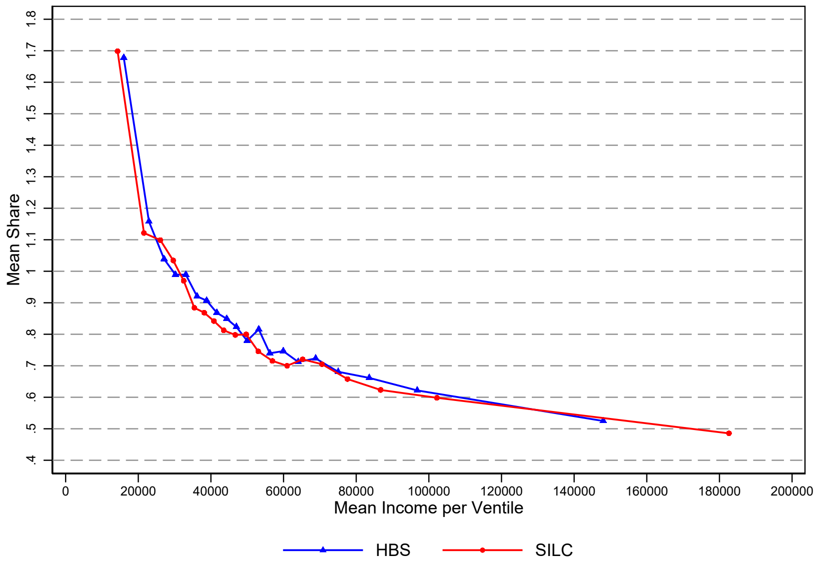

Figure 4. Comparison: average income shares of total expenditures between HBS and SILC per ventile
Source: HBS 2015-2017; EU-SILC 2020; authors' illustration.

Figure 4 displays the average income shares allocated to total household consumption on a monthly basis. In this context, total consumption comprises all subcategories of the HBS category "consumption expenditures." Voluntary insurance premiums or similar items are not included. Although households are, as described above, grouped into ventiles based on their weighted disposable income, it should be noted that average incomes per ventile do not coincide exactly between the HBS and SILC datasets. This is to be expected, as the application of population weights does not fully harmonize the two data sources.

When considering the deviation in overall equivalised mean income across all households, Table 3 shows a difference of CHF 1,850. In the case where population weights are applied, the average deviation in mean income amounts to CHF 2,068. For comparison, separate deciles were additionally constructed for the population-weighted scenario. The average absolute deviation between population-weighted deciles amounts to CHF 2,984, which is lower than in the unweighted case (CHF 3,176), as positive and negative deviations do not offset each other. This indicates a slight harmonization of the datasets and improved comparability, but by no means a complete alignment.

These deviation measures apply to both deciles and ventiles. Consequently, an exact correspondence of total consumption shares between the HBS and SILC datasets cannot be expected.

Applying the analytical framework proposed by Akoğuz et al. (2020) leads to the following conclusions:

First, it should be noted that the pattern of share distributions is largely identical between the imputed and the observed data. The total share of consumer expenditures declines steadily with increasing income, with the exception of a minor and short-lived increase. The decrease in income shares also becomes less pronounced at higher income levels, with both observed and imputed values exhibiting very similar trends across all ventiles. At lower income levels, the decline is steep in both datasets and gradually flattens as income rises.

Next, the vertical distances between the red data points and the blue interpolation line above or below these points must be considered. These vertical distances should always be analyzed and interpreted relative to the scale of the y-axis. The greater the distance between the highest and lowest values, the smaller the individual deviations appear visually. Nevertheless, in the present case, the figure shows relatively small deviations, ranging between 1% and 5%. The overall mean values across the 20 income-share plots amount to 86.13% for HBS and 84.40% for SILC, resulting in an average difference of 1.73 percentage points.

Finally, it is important to assess whether systematic over- or under-imputation has occurred, that is, whether the estimated shares predominantly lie above or below the interpolation line. In the present case, there is a tendency toward slight under-imputation, meaning that the imputed shares lie on average marginally below the blue line. However, this under-imputation is of limited magnitude and does not have a meaningful effect on the overall pattern.

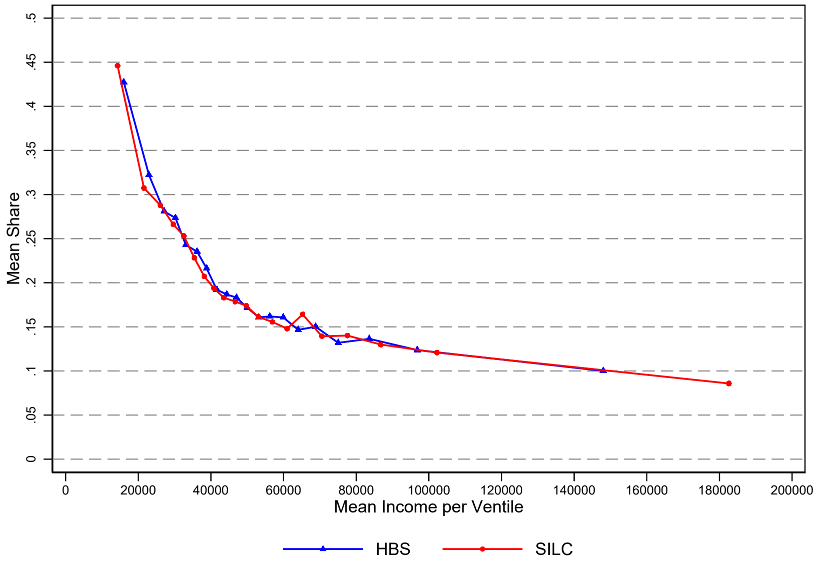

Figure 5. Comparison: average shares of category "Housing and rental" between HBS and SILC per ventile
Source: HBS 2015-2017; EU-SILC 2020; authors' illustration.

Figure 5 illustrates the relationship between income and expenditure shares for the category "housing and rental". As before, the overall pattern is consistent between the imputed and observed values. A negative correlation is evident, and the change in slope is also consistent across the HBS and SILC datasets. However, the vertical distance between the imputed shares and the interpolation line is smaller in this case, resulting in a higher degree of visual alignment, albeit on a tighter scale of the y-axis. This is also reflected in the average expenditure shares across ventiles. The mean HBS share amounts to 20.02%, while the mean SILC share is 19.85%. Overall, there is no indication of systematic over- or under-imputation, at most a slight tendency toward marginally lower percentages.

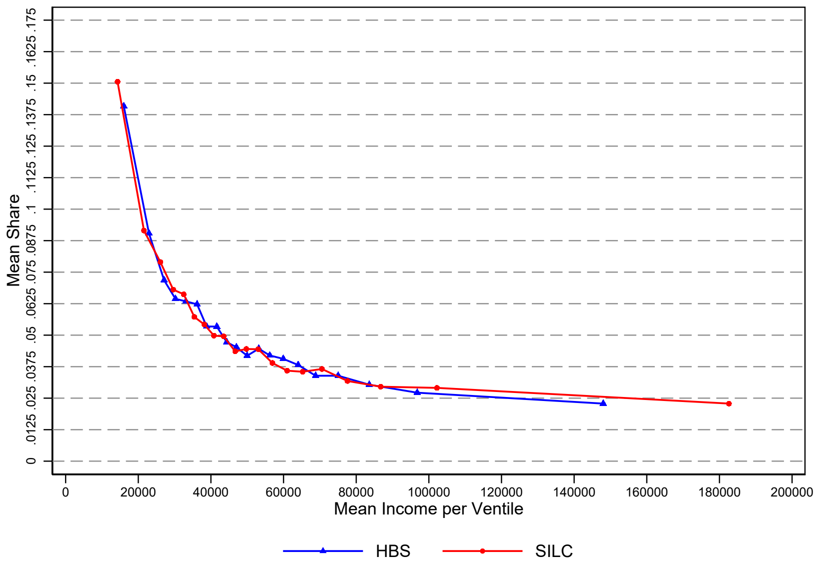

Figure 6. Comparison: average shares of category "Utilities" between HBS and SILC per ventile
Source: HBS 2015-2017; EU-SILC 2020; authors' illustration.

Figure 6 shows the correlation in the category "Utilities". Here, too, the same assessment can be made as before. The general trend, as well as the vertical distances and the pattern with regard to over- or under-imputation, indicate successful imputation. The average income shares are 5.24% for HBS and 5.29% for SILC. The total difference is therefore only 0.05%. Due to the small differences in the visual analysis, this value can also be approximated as the average difference within the individual valves. This result also coincides with the results from Akoğuz et al. (2020), where, among other things, the category "Utilities" performed best in all 18 countries.

Further graphs for the remaining categories can be found in the Appendix 6. In general, the analysis across all aggregated categories shows that the more frequently a category is consumed - i.e., the more regular the consumption and, on the other hand, the higher the share of consumption expenditure in disposable income - the more consistent the imputation results are.

This can be illustrated using the categories "education" and "vehicles." Here, the patterns between imputed and observed values overlap significantly less. In addition, the percentage deviations in average expenditure shares between HBS and SILC are greater when measured relative to the respective category shares. However, this pattern does not apply to all categories: for example, imputation works comparatively well in the “tobacco” category, even though it is characterized by both a low relative expenditure share and a high level of non-consumption. Nevertheless, despite the loss of quality described above, the results are good overall in some categories. It should be noted, however, that with a smaller scaling range, larger errors or deviations occur when only visual analysis is used (Akoğuz et al. 2020). Appendix 6 contains summary tables showing the average values for the valves for all aggregated output categories.

5.6. *Distributional Similarity Assessment: Kolmogorov-Smirnov Test Results*

The Kolmogorov-Smirnov (KS) test is a nonparametric method for comparing two empirical distribution functions (EDFs). It measures the maximum absolute difference between the EDFs of HBS and SILC (D-value) and computes a p-value indicating whether this difference could occur under the null hypothesis that both samples originate from the same distribution (Hedderich & Sachs 2020; Kolmogorov 1933; Smirnov 1939).

Table 7 reports KS test results for all 20 aggregated consumption categories. A low D-value combined with a high p-value suggests strong similarity between distributions. However, given the large sample sizes (approximately 8,000-10,000 observations), p-values tend to be very small even for minor deviations. Therefore, interpretation should primarily rely on D-values (Field 2018).

Overall, D-values are generally low, indicating strong alignment between observed and imputed distributions. Categories with frequent zero expenditures - such as “Vehicles” and secondary residence costs - show particularly high consistency, as both datasets report zeros for most households. However, this pattern underscores a limitation: high similarity in these cases reflects structural zeros rather than predictive accuracy. Consequently, KS test results should be interpreted alongside other validation measures and with attention to category-specific characteristics.

Table 7. Results of the Kolmogorov–Smirnov Test

Category	D-value	p-value
Housing and rental	0.0302	0.001
Food & non-alcoholic beverages	0.0644	0.000
Utilities	0.0224	0.023
Health & Care	0.0631	0.000
Communication	0.0368	0.000
Insurance	0.0436	0.000
Culture & leisure	0.0279	0.002
Private transport	0.0617	0.000
Tobacco	0.0163	0.190
Housing goods & services	0.0338	0.000
Public Transport	0.0085	0.908
Other	0.0152	0.254
Restaurants	0.0137	0.372
Alcoholic beverages	0.0340	0.000
Education	0.0172	0.145
Personal care	0.0371	0.000
Clothing and personal items	0.0248	0.008
Housing durables	0.0459	0.000
Utilities secondary residence	0.0036	1.000
Housing and rental secondary residence	0.0035	1.000
Traveling and holidays	0.0226	0.021
Vehicles	0.0106	0.696

5.7. *Aggregate Consistency Checks and External Benchmarking*

This subsection provides an overall assessment of the imputed consumption data and examines its plausibility through aggregated comparisons with macroeconomic indicators. A direct comparison of consumption-specific expenditures with national accounts (NA) is not feasible because NA values for household consumption are derived using the usage approach, which itself relies on HBS data (FSO 2025b). Consequently, such a comparison would not offer an independent benchmark.

Instead, we use alternative proxies - such as social security contributions and federal tax revenues - that are recorded independently of HBS. These indicators allow us to verify whether aggregated totals in the imputed SILC data align with those observed in HBS and, subsequently, with macro-level figures. This two-step comparison serves two purposes: first, to assess internal consistency between donor and recipient datasets; and second, to provide an initial indication of whether the imputed data reflects a realistic overall picture.

It is important to note that any discrepancies between HBS and official macroeconomic aggregates cannot be fully resolved within the scope of this study, as they may stem from structural differences in data collection or reporting. Nevertheless, presenting these results enhances transparency and helps contextualize the imputation quality beyond micro-level validation.

A first comparison concerns the contributions of Swiss households to health insurance. These insurance policies are taken out individually by households on a private basis, but are nevertheless compulsory. Therefore, there is no employer contribution if the household is in an employee-employer relationship (FOPH 2024). According to social security statistics, revenues in 2019 amounted to 26.9 billion Swiss francs (Schüpbach 2021). In the HBS and SILC data sets, the values in the category "Premiums for basic insurance" were extrapolated using the available population weights. This results in a total of 27.5 billion for HBS households and 26.05 billion for SILC households. The consistency ratio is therefore 1.022 for HBS and 0.968 for SILC.

Other social security contributions include, for example, old age and survivors' insurance (AHV), disability insurance (IV), and income compensation (EO). Unlike health insurance, these are divided equally between employees and employers (FSIO 2025). However, the total amount for these three types of insurance also includes self-employed persons who cannot be excluded. Therefore, households with dependent employees were weighted twice to simulate the employer's contribution. According to social security statistics, contribution revenues amounted to 39.6 billion in 2019 (Schüpbach 2021). Within HBS, the total is approximately 22.6 billion, and in the SILC data set, approximately 26.5 billion. The situation is similar with regard to federal tax revenue. This actually amounted to 11.8 billion Swiss francs in 2019 (FFA 2020). Calculations based on HBS households yield a total of approximately 6.1 billion, while calculations based on SILC households yield approximately 7.7 billion. The ratio for HBS ranges between 50 and 60% in each case. The ratio for SILC performs better, ranging between 65 and 70% in each case. The higher values in SILC are probably due to the higher weighted average income of households. As noted in subsection 4.1, this is almost CHF 2,000 higher.

Comparing savings rates with external data is again difficult, as the official figures are also based on HBS data. First of all, it should be noted that the average values across all households are relatively similar with saving rates of 13.9 % in the

HBS dataset and 15,6 % in the silc data set. The slightly higher savings rate in the SILC data set may partly represent the higher average income. A comparison with the external values reported in the literature indicates that our results are generally comparable (i.e. here Lehner and Hofmann (2023)), although there may be deviations due to variations in the calculation of the savings rate and to our adjustment for inflation rates to bring the data up to the 2019 level. In the present analysis, only expenditures in the "consumer spending" category were considered and calculated on the basis of disposable income, without taking into account expenditure in the areas of insurance, fees, or transfers, as these often include elements of voluntary savings, such as discretionary retirement contributions.

Furthermore, it is striking that the savings rate (negative) is particularly high in the lowest income bracket. It is difficult to find relevant comparisons in the literature, as in most cases overall averages across all income classes are given. In addition, there are structural differences between countries in terms of savings rates, which vary considerably Eurostat (2025b). Nevertheless, the values should be viewed critically at this point, as there are indications in the literature that household income tends to be under reported in the lower income classes (Meyer, Mok, & Sullivan 2015). This makes expenditure appear relatively higher and therefore results in higher (negative) savings rates. However, this is a question of data collection and is beyond our control.

6. Conclusion

As part of this study, consumption data from the Household Budget Survey (HBS) was successfully imputed into the Statistics on Income and Living Conditions (SILC) dataset and analysed. The methodological approach proposed by Akoğuz et al. (2020) proved applicable even for a non-EU country such as Switzerland. Large-scale consumption patterns observed within HBS could largely be retained and transferred to SILC, with only minor deviations likely caused by inherent sample differences that cannot be fully harmonized through weighting.

Prior to the two-stage regression process, data adjustment measures led to the exclusion of 167 households from HBS and 15 households from SILC, ensuring that sufficient variance was retained for the imputation process. Six aggregated consumption categories exceeded the threshold of 0.1 and were used for distance measurement, a result comparable to the results in the literature. Using the Mahalanobis distance, 4,355 HBS households were identified as donors, with each donor assigned on average to 1.87 SILC households. Consequently this approach maintained a high degree of heterogeneity in consumption patterns while allowing the transferred data to closely reflect the structure of the original HBS dataset.

Overall, the results that are shown by the constructed ventile graphs confirm that imputation can reliably preserve key consumption structures dependent of the weighted household income, across datasets. Categories with regular consumption, such as housing, utilities or food exhibited particularly consistent transfers, while less frequent categories showed slightly higher deviations. These findings demonstrate the robustness of the imputation approach and provide a solid basis for further analyses of household consumption behaviour in Switzerland.

Acknowledgements

We would like to thank Elif Cansu Akoğuz for her valuable assitance during the editing process and for helping us better understand the methodlogical approach. Further we would also like to thank Peter Bolliger from the Swiss Federal Statistical Office for his helpful explanations regarding the data set. We are also grateful to Bernhard Neumärker for his constant support and guidance. In addition, we thank Martha Brinkmann for her excellent research support.

Funding

This research was supported by the Erasmus+ Programme of the European Union (ecoMOD Project, project number: 2023- 1-LI01-KA220-HED- 000157594), by the Freiburg Institute for Basic Income Studies (FRIBIS) at Albert-Ludwigs-Universität Freiburg, and by the Forschungsförderungsfond (FFF) of the University of Liechtenstein. The authors gratefully acknowledge the generous support from all these organizations.

Conflict of Interest

No competing interests reported.

Data Availability and Replication Materials

The Swiss HBS and Swiss SILC microdata used in this study are available from the Swiss Federal Statistical Office (FSO) upon request and are subject to confidentiality and data-use agreements. All code for data cleaning, harmonisation, imputation, and validation is available from the corresponding author upon request.

References

Akoğuz, E. C., Capéau, B., Decoster, A., De Sadeleer, L., Güner, D., Manios, K., ... Vanheukelom, T. (2020). *A new indirect tax tool for EUROMOD: Final report* (Tech.

Rep. No. JRC/SVQ/2018/B.2/0021/OC). European Commission, Joint Research Centre (JRC).

Donatiello, G., Toso, S., & Baldini, M. (2025). Matching Wealth Data with EU-SILC: A Statistical Approach for Italy. *International Journal of Microsimulation*, *18*(2), 45–68. Retrieved from https://doi.org/10.34196/ijm.00234

Dreoni, I., Serruys, H., Manso, L., Tudó-Ramírez, J., & Amores, A. F. (2025). *Statistical Imputation and Validation of Consumption Microdata for EUROMOD* (Tech. Rep. Nos. JRC Working Papers on Taxation and Structural Reforms No 2/2025, JRC141513). European Commission Joint Research Centre (JRC). Retrieved from https://publications.jrc.ec.europa.eu/repository/handle/JRC141513

European Commission. (2025). *EUROMOD - Tax-benefit microsimulation model for the European Union.* Retrieved from https://euromod-web.jrc.ec.europa.eu/overview/what-is-euromod (Accessed: 22.10.2025)

Eurostat. (2025a). *Household Budget Surveys - Information on Data.* Retrieved from https://ec.europa.eu/eurostat/web/household-budget-surveys/information-data (Accessed: 22.10.2025)

Eurostat. (2025b). *Sparquote privater Haushalte.* Retrieved from https://ec.europa.eu/eurostat/databrowser/view/tec00131/default/table?lang=de (Online Datacode: tec00131. Accessed: 12.12.2025)

FFA. (2020). *Band 1 – Bericht zum Voranschlag 2020 mit IAFP 2021–2023.* Publication. Bern, Schweiz. Retrieved from https://www.efv.admin.ch/de/voranschlag-mit-integriertem-aufgaben-und-finanzplan#2003%E2%80%932021

FFA. (2025). *Einnahmen.* Retrieved from https://www.efv.admin.ch/de/einnahmen (Accessed: 15.12.2025)

Field, A. (2018). *Discovering Statistics Using IBM SPSS Statistics* (5th ed.). Los Angeles: SAGE.

FOE. (2024). *Schweiz—CO2-Abgabe in der Schweiz von 2008 bis 2023 (in CHF pro Tonne CO2) [Graph].* Statista. Retrieved from https://de.statista.com/statistik/daten/studie/1008172/umfrage/co2-abgabe-in-der-schweiz/ (Accessed: 15.12.2025)

FOPH. (2024). *Die obligatorische Krankenversicherung (Ratgeber).* Bern, Schweiz. Retrieved from https://www.bag.admin.ch/de/krankenversicherung-leistungen-und-tarife#Tarife

FSIO. (2025). *Beiträge an die Sozialversicherungen im Überblick.* Bern, Schweiz. Retrieved from https://www.bsv.admin.ch/de/beitraege-im-ueberblick#Tabellen-%C3%9Cbersicht-aller-Beitr%C3%A4ge

FSO. (2025a). *HVPI (2015=100), Totalindex der Schweiz und der europäischen Länder, Veränderungsraten.* Retrieved from https://www.bfs.admin.ch/bfs/de/home/statistiken/preise/harmonisierte-verbraucherpreise.assetdetail.36093120.html (Accessed: 29.08.2025)

FSO. (2025b). *Revisionsanalyse 2025: Volkswirtschaftliche Gesamtrechnung Details.* Neuchâtel, Schweiz. Retrieved from https://www.bfs.admin.ch/asset/de/be-d-04-VGR-03-sup

Hedderich, J., & Sachs, L. (2020). *Angewandte Statistik: Methodensammlung mit R* (17th revised and expanded edition ed.). Berlin: Springer Spektrum.

Kirn, T., Oschwald, P., & Anderl, R. (2025). SWISSMOD – A New Tax-Benefit Model of Switzerland. *International Journal of Microsimulation*, *18*(2), 1–20. doi:

Kolmogorov, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, *4*.

Lehner, S., & Hofmann, R. (2023). *Die Schweiz ist eine Sparnation.* Retrieved from https://dievolkswirtschaft.ch/de/2023/11/die-schweiz-ist-eine-sparnation/ (Accessed: 10.12.2025)

Mahalanobis, P. (2018). Reprint of: On the generalised distance in statistics. *Sankhya A*, *80*(Suppl 1), 1–7. doi:

Meyer, B. D., Mok, W. K. C., & Sullivan, J. X. (2015). Household Surveys in Crisis. *Journal of Economic Perspectives*, *29*(4), 199–226. Retrieved from https://doi.org/10.1257/jep.29.4.199

Schüpbach, S. (2021). *Gesamtrechnung der Sozialversicherungen 2019* [Publication / Technical report]. Retrieved from https://www.bsv.admin.ch/dam/de/sd-web/JVb9sEzNr1cf/chss-gesamtrechnung-2019.pdf

Serafino, P., & Tonkin, R. (2017). Statistical matching of European Union statistics on income and living conditions (EU-SILC) and the household budget survey [Statistical Working Papers]. , *10*.

Smirnov, N. (1939). On the Estimation of the Discrepancy Between Empirical Curves of Distribution for Two Independent Samples. *Bulletin Moscow University*, *2*, 3-16.

Appendices

Appendix A:

Overview table of selected expenditure categories and their aggregation level

From left to right, Table 8 and Table 9 shows the aggregation level from which the corresponding expenditures per category originate and the title of the expenditure items that make up a category.

Table 8. Overview of consumption categories

Category	Level of aggregation				
	First	Second	Third	Fourth	Fifth
1. Food and non alcoholic beverages		A51 (Food and non-alcoholic beverages)			
2. Housing and rental				A5711 (Net rent or mortgage interest on the primary residence)	
3. Utilities				A5712 (Additional costs of the main residence) + A5713 (Energy of the primary residence)	
4. Communication		A63 (Messaging)			
5. Personal Care			A681 (Personal hygiene)		
6. Insurance	A80 (Life insurance premiums)	A41 (Health insurance: Premiums for supplementary insurance) + A42 (Other insurance premiums)			
7. Alcoholic beverages			A521 (Alcoholic beverages)		
8. Tobacco			A522 (Tobacco products)		
9. Private transport				A6214 (Accessories and spare parts for vehicles) + A6215 (Fuels and lubricants) + A6216 (Vehicle service and repairs) + A6217 (Other services in the passenger vehicle sector)	
10. Education		A67 (School and training fees)			
11. Clothing and personel items		A56 (Clothing and shoes)	A682 (Personal equipment)		
12. Health and Care		A61 (Healthcare spending) + A31 (Social security contributions) + A33 (Health insurance companies: Premiums for basic insurance)			
13. Restaurants			A531 (Restaurants)		

Table 9. Overview of consumption categories - continuation

14. Housing goods and services			A573 (Repairs and maintenance of the apartment) + A585 (Ongoing budget management)		
15. Housing durables			A581 (Furniture, decoration, and flooring, including repairs) + A582 (Household linen and home textiles) + A583 (Household and kitchen appliances) + A584 (Tools for home and garden)		
16. Culture and leisure			A661 (Audiovisual, photographic, and computer equipment and accessories)) + A662 (Additional equipment and items for entertainment purposes) + A663 (Services for sports, recreation, and culture) + A664 (Books, press products, and stationery)		
17. Public Transport			A622 (Transportation services)		
18. Vehicles				A6211 (Cars) + A6212 (Motorcycles, scooters, and mopeds) + A6213 (Bycicles)	
19. Traveling and hollidays			A532 (Accommodation facilities) + A665 (Package tours)		
20. Other			A683 (Social, financial, and other services)		
21. Housing and rental secondary residence				A5721 (Net rent and mortgage interest on secondary residences)	
22. Utilities secondary residence				A5722 (Additional costs for secondary residences) + A5723 (Energy secondary residences)	

Appendix B:

Overview tables of average values per ventile for every selected expenditure category

From left to right, Table 10 and Table 11 shows the average expenditure level of every ventile, that are generated by using the population weights. It starts with the ventile with the lowest average household income.

Table 10. Overview of average expenditure shares per ventile (rounded)

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.	17.	18.	19.	20.
Total Consumption																				
HBS-share	1.678	1.158	1.039	0.989	0.989	0.920	0.907	0.868	0.849	0.823	0.816	0.779	0.747	0.740	0.723	0.712	0.681	0.662	0.622	0.525
SILC-share	1.699	1.122	1.099	1.034	0.970	0.884	0.868	0.842	0.813	0.800	0.798	0.746	0.720	0.715	0.705	0.700	0.658	0.623	0.598	0.486
Housing and Rental																				
HBS-share	0.427	0.322	0.281	0.273	0.243	0.235	0.216	0.192	0.187	0.183	0.172	0.162	0.161	0.161	0.150	0.147	0.136	0.132	0.124	0.100
SILC-share	0.446	0.307	0.288	0.266	0.253	0.228	0.207	0.194	0.183	0.179	0.174	0.164	0.161	0.156	0.148	0.140	0.139	0.130	0.121	0.086
Food & non-alcoholic Beverages																				
HBS-share	0.274	0.192	0.162	0.150	0.140	0.124	0.120	0.118	0.113	0.104	0.100	0.098	0.089	0.085	0.080	0.076	0.069	0.064	0.056	0.042
SILC-share	0.271	0.171	0.158	0.132	0.128	0.115	0.110	0.107	0.101	0.098	0.094	0.084	0.081	0.078	0.077	0.067	0.066	0.057	0.054	0.040
Utilities																				
HBS-share	0.141	0.090	0.072	0.064	0.063	0.062	0.053	0.053	0.047	0.045	0.045	0.042	0.042	0.041	0.038	0.034	0.034	0.030	0.027	0.023
SILC-share	0.151	0.091	0.079	0.068	0.066	0.057	0.054	0.050	0.050	0.045	0.044	0.044	0.039	0.037	0.036	0.035	0.032	0.030	0.029	0.023
Communication																				
HBS-share	0.070	0.049	0.045	0.043	0.041	0.035	0.035	0.033	0.031	0.031	0.029	0.028	0.028	0.028	0.026	0.024	0.023	0.021	0.018	0.014
SILC-share	0.076	0.050	0.042	0.041	0.040	0.034	0.033	0.032	0.031	0.030	0.030	0.029	0.027	0.027	0.024	0.023	0.022	0.021	0.018	0.013
Health & Care																				
HBS-share	0.486	0.347	0.300	0.299	0.297	0.283	0.278	0.277	0.274	0.269	0.256	0.255	0.250	0.246	0.246	0.244	0.237	0.235	0.226	0.212
SILC-share	0.489	0.328	0.303	0.291	0.290	0.273	0.264	0.261	0.260	0.258	0.256	0.248	0.244	0.241	0.234	0.234	0.228	0.227	0.217	0.201
Insurance																				
HBS-share	0.150	0.093	0.090	0.088	0.088	0.087	0.087	0.086	0.086	0.086	0.086	0.085	0.085	0.084	0.083	0.082	0.080	0.080	0.079	0.071
SILC-share	0.156	0.095	0.089	0.088	0.088	0.087	0.087	0.087	0.086	0.085	0.085	0.084	0.084	0.084	0.083	0.081	0.080	0.079	0.078	0.066
Culture & Leisure																				
HBS-share	0.115	0.072	0.072	0.066	0.066	0.065	0.065	0.065	0.064	0.062	0.061	0.061	0.060	0.060	0.056	0.055	0.055	0.054	0.047	0.044
SILC-share	0.116	0.080	0.079	0.071	0.069	0.065	0.064	0.063	0.061	0.061	0.058	0.057	0.056	0.055	0.055	0.055	0.052	0.049	0.047	0.039
Private Transport																				
HBS-share	0.085	0.064	0.061	0.060	0.060	0.058	0.058	0.058	0.058	0.057	0.055	0.055	0.055	0.055	0.053	0.052	0.051	0.049	0.044	0.038
SILC-share	0.087	0.073	0.064	0.060	0.059	0.058	0.057	0.057	0.054	0.053	0.053	0.052	0.052	0.051	0.051	0.051	0.046	0.045	0.043	0.033
Tobacco																				
HBS-share	0.018	0.014	0.013	0.012	0.009	0.009	0.008	0.008	0.007	0.007	0.007	0.007	0.006	0.006	0.005	0.005	0.004	0.003	0.003	0.001
SILC-share	0.021	0.015	0.013	0.012	0.011	0.009	0.008	0.008	0.007	0.007	0.007	0.006	0.006	0.005	0.005	0.005	0.004	0.003	0.003	0.001
Housing Goods & Services																				
HBS-share	0.028	0.023	0.020	0.019	0.019	0.019	0.018	0.018	0.017	0.017	0.017	0.015	0.015	0.014	0.014	0.014	0.014	0.013	0.013	0.013
SILC-share	0.034	0.025	0.021	0.019	0.017	0.017	0.016	0.016	0.014	0.014	0.014	0.014	0.014	0.013	0.013	0.013	0.013	0.012	0.012	0.012

Source: HBS 2015–2017; EU-SILC 2020.

Table 11. Overview of average ependiture shares per ventile (rounded)

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.	17.	18.	19.	20.
Public Transport																				
HBS-share	0.045	0.030	0.030	0.028	0.027	0.027	0.026	0.026	0.026	0.026	0.026	0.025	0.025	0.025	0.024	0.024	0.023	0.023	0.022	0.021
SILC-share	0.038	0.032	0.032	0.030	0.030	0.029	0.027	0.026	0.026	0.026	0.025	0.025	0.024	0.023	0.023	0.023	0.022	0.022	0.020	0.018
Other																				
HBS-share	0.017	0.016	0.015	0.015	0.015	0.015	0.015	0.013	0.013	0.013	0.013	0.013	0.013	0.013	0.012	0.012	0.012	0.011	0.010	0.009
SILC-share	0.019	0.019	0.017	0.016	0.015	0.015	0.014	0.014	0.014	0.013	0.012	0.012	0.012	0.011	0.011	0.011	0.010	0.010	0.009	0.009
Restaurants																				
HBS-share	0.123	0.074	0.072	0.072	0.071	0.071	0.070	0.069	0.069	0.069	0.068	0.067	0.067	0.067	0.067	0.067	0.066	0.066	0.065	0.053
SILC-share	0.118	0.085	0.083	0.079	0.076	0.076	0.075	0.072	0.071	0.070	0.069	0.068	0.067	0.065	0.064	0.063	0.062	0.062	0.060	0.048
Alcoholic Beverages																				
HBS-share	0.025	0.013	0.013	0.012	0.012	0.012	0.012	0.011	0.011	0.010	0.010	0.010	0.010	0.010	0.009	0.009	0.009	0.009	0.008	0.008
SILC-share	0.036	0.016	0.015	0.013	0.011	0.011	0.011	0.010	0.010	0.010	0.010	0.010	0.009	0.009	0.008	0.008	0.008	0.008	0.008	0.007
Education																				
HBS-share	0.013	0.009	0.008	0.008	0.007	0.007	0.007	0.007	0.007	0.006	0.006	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.004
SILC-share	0.017	0.012	0.009	0.008	0.007	0.007	0.006	0.006	0.006	0.006	0.005	0.005	0.005	0.004	0.003	0.003	0.003	0.003	0.003	0.003
Personal Care																				
HBS-share	0.035	0.023	0.021	0.021	0.020	0.020	0.020	0.020	0.019	0.018	0.018	0.017	0.017	0.016	0.015	0.014	0.014	0.014	0.013	0.010
SILC-share	0.038	0.026	0.023	0.021	0.020	0.020	0.019	0.018	0.018	0.018	0.017	0.017	0.017	0.016	0.016	0.015	0.012	0.012	0.012	0.009
Clothing & Personell Items																				
HBS-share	0.059	0.041	0.040	0.040	0.039	0.037	0.036	0.036	0.036	0.036	0.035	0.035	0.035	0.034	0.034	0.034	0.034	0.033	0.032	0.027
SILC-share	0.046	0.044	0.043	0.042	0.041	0.039	0.039	0.039	0.038	0.037	0.037	0.036	0.035	0.035	0.035	0.035	0.033	0.033	0.031	0.025
Housing Durables																				
HBS-share	0.046	0.031	0.030	0.029	0.029	0.028	0.028	0.028	0.026	0.026	0.026	0.026	0.025	0.025	0.025	0.025	0.025	0.024	0.023	0.020
SILC-share	0.043	0.041	0.035	0.031	0.028	0.027	0.027	0.025	0.025	0.025	0.025	0.025	0.025	0.024	0.024	0.024	0.023	0.022	0.020	0.019
Utilities Secondary Residence																				
HBS-share	0.007	0.004	0.003	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
SILC-share	0.003	0.003	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Traveling & Holidays																				
HBS-share	0.040	0.039	0.038	0.037	0.037	0.036	0.036	0.035	0.035	0.033	0.033	0.032	0.032	0.032	0.031	0.030	0.030	0.028	0.023	0.022
SILC-share	0.042	0.042	0.038	0.037	0.037	0.035	0.034	0.034	0.034	0.034	0.033	0.031	0.031	0.031	0.031	0.030	0.028	0.026	0.020	0.020
Vehicles																				
HBS-share	0.035	0.035	0.034	0.032	0.031	0.030	0.028	0.027	0.026	0.026	0.026	0.025	0.025	0.025	0.025	0.022	0.022	0.020	0.020	0.018
SILC-share	0.034	0.034	0.030	0.030	0.029	0.029	0.027	0.027	0.027	0.027	0.026	0.024	0.024	0.023	0.023	0.021	0.020	0.016	0.014	0.012

Source: HBS 2015–2017; EU-SILC 2020.

Appendix C:

Ventile graphs for the remaining selected expenditure categories

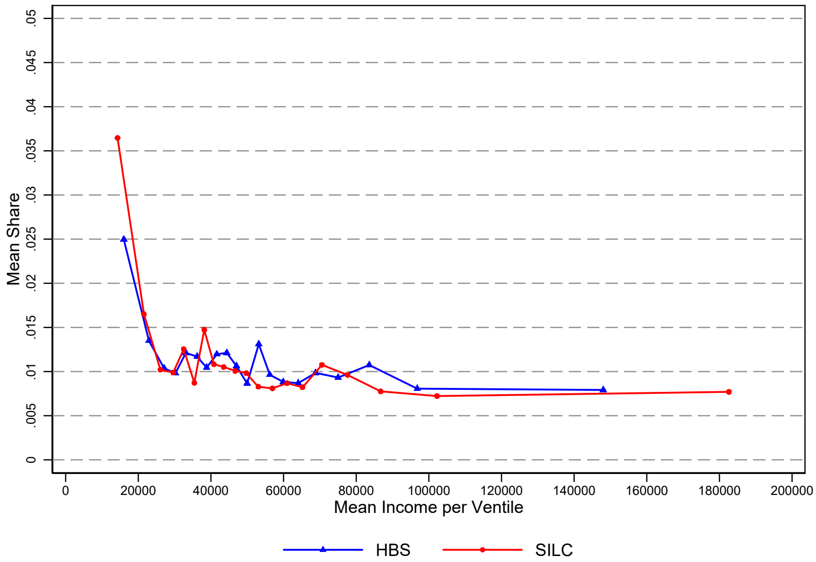

Figure 7. Comparison: average shares of category "Alcoholic beverages" between HBS and SILC per ventile

Source: HBS 2015–2017; EU-SILC 2020; authors' illustration.

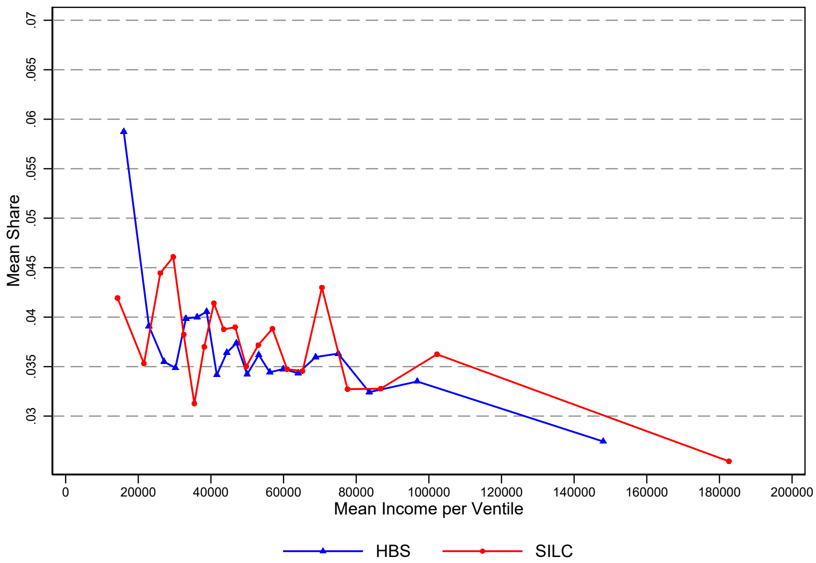

Figure 8. Comparison: average shares of category "Clothing and personal items" between HBS and SILC per ventile

Source: HBS 2015–2017; EU-SILC 2020; authors' illustration.

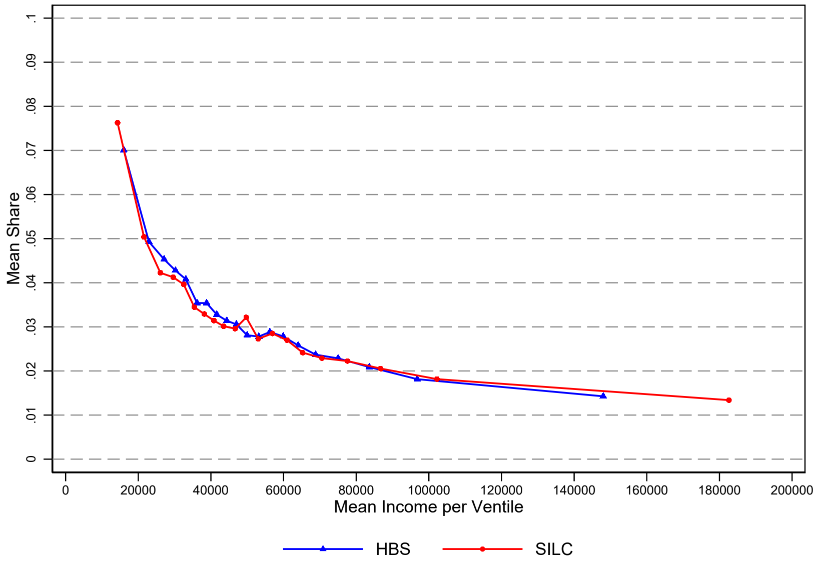

Figure 9. Comparison: average shares of category "Communication" between HBS and SILC per ventile

Source: HBS 2015–2017; EU-SILC 2020; authors' illustration.

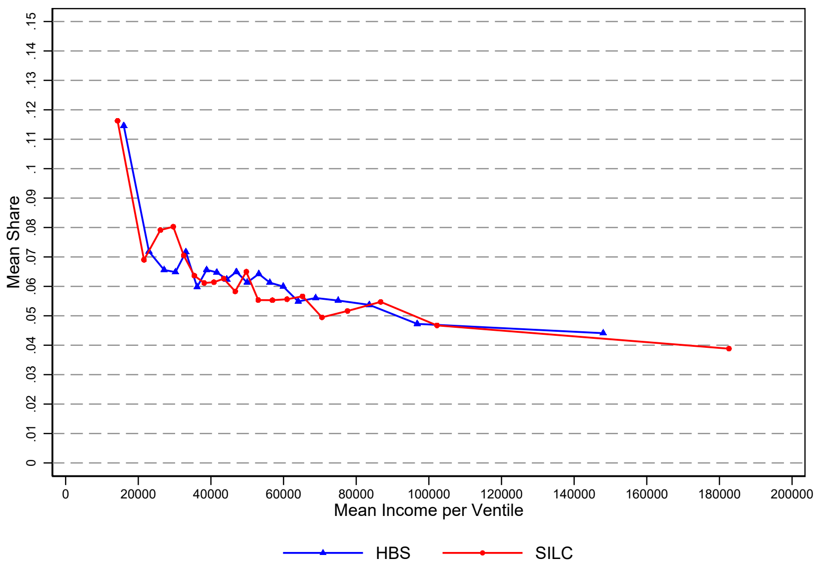

Figure 10. Comparison: average shares of category "Culture and leisure" between HBS and SILC per ventile

Source: HBS 2015–2017; EU-SILC 2020; authors' illustration.

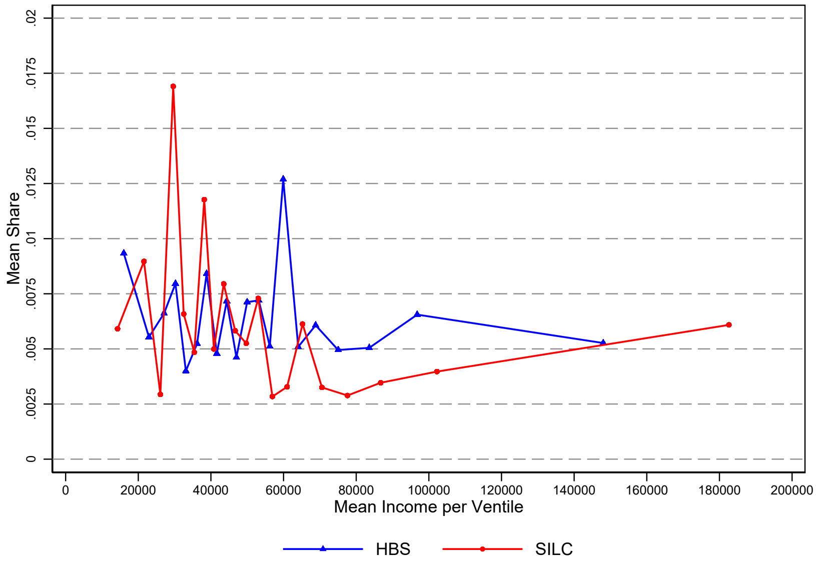

Figure 11. Comparison: average shares of category "Education" between HBS and SILC per ventile

Source: HBS 2015–2017; EU-SILC 2020; authors' illustration.

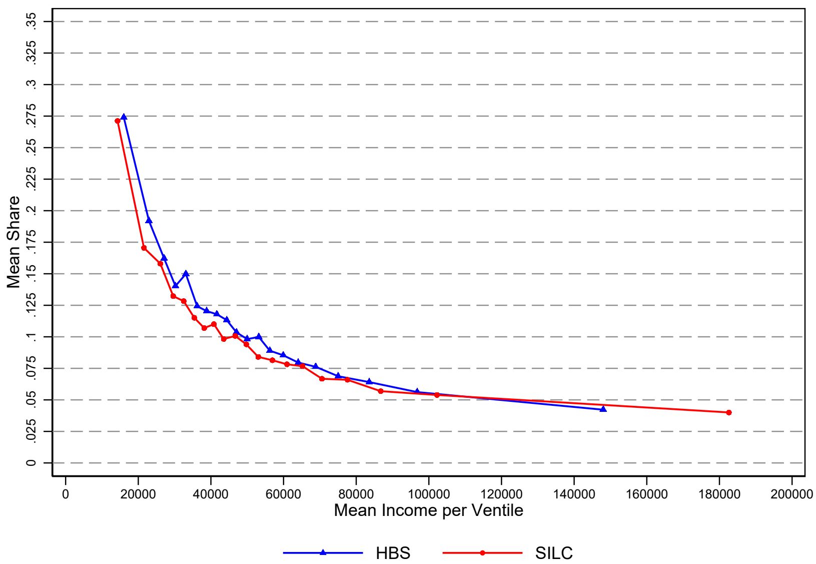

Figure 12. Comparison: average shares of "Food & non-alcoholic Beverages" between HBS and SILC per ventile

Source: HBS 2015–2017; EU-SILC 2020; authors' illustration.

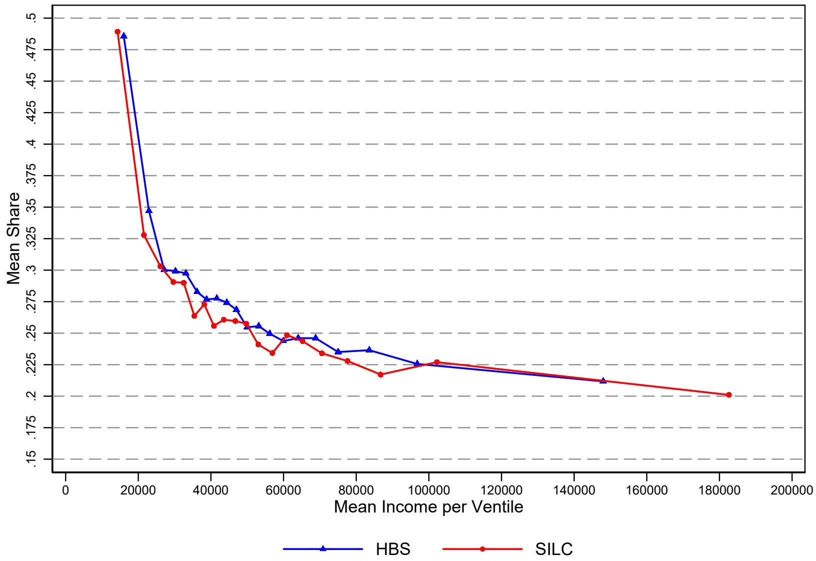

Figure 13. Comparison: average shares of category "Health & Care" between HBS and SILC per ventile
Source: HBS 2015–2017; EU-SILC 2020; authors' illustration.

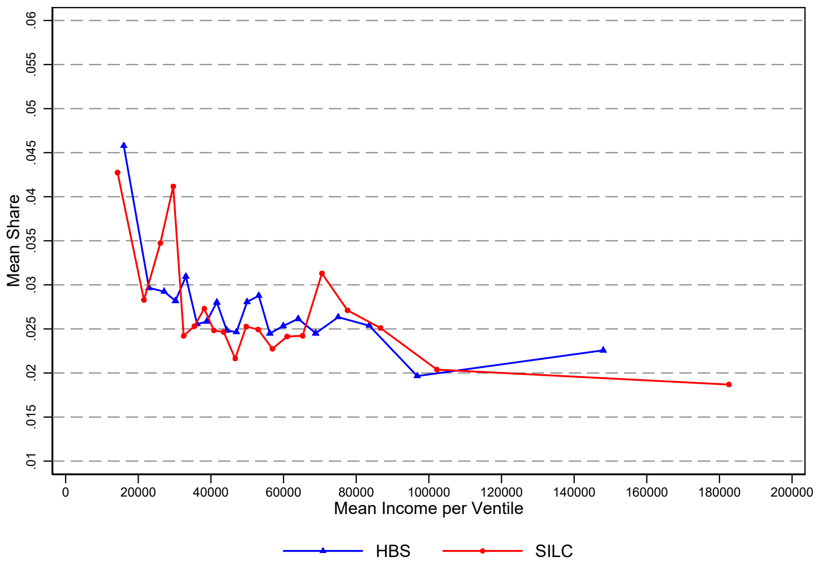

Figure 14. Comparison: average shares of category "Housing durables" between HBS and SILC per ventile
Source: HBS 2015–2017; EU-SILC 2020; authors' illustration.

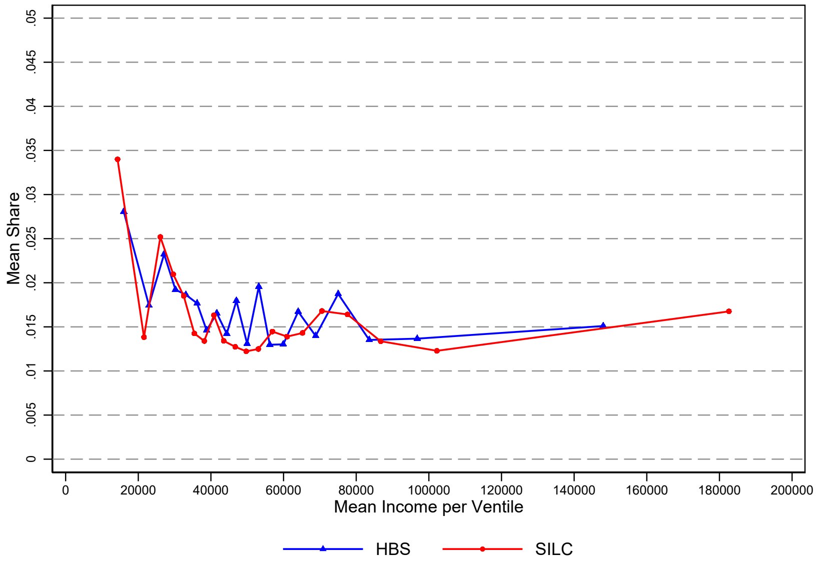

Figure 15. Comparison: average shares of category "Housing goods services" between HBS and SILC per ventile

Source: HBS 2015–2017; EU-SILC 2020; authors' illustration.

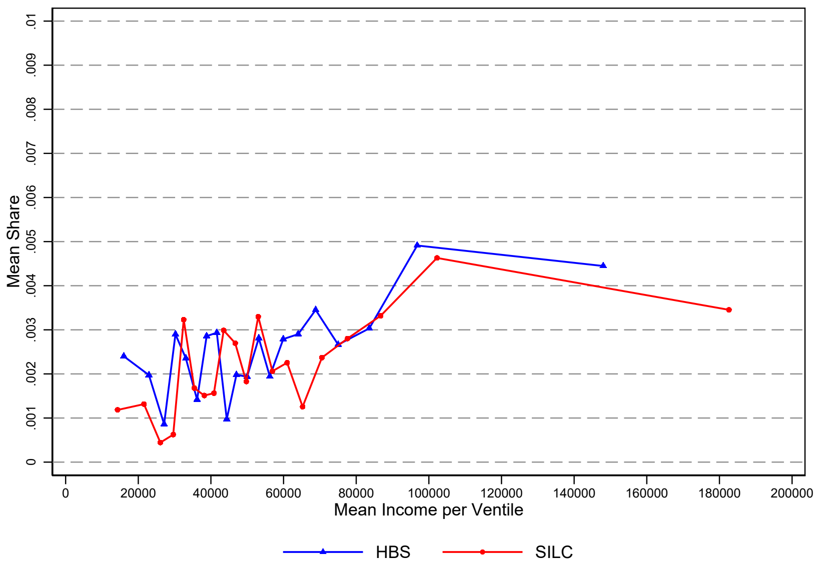

Figure 16. Comparison: average shares of category "Housing and Rental secondary residence" between HBS and SILC per ventile

Source: HBS 2015–2017; EU-SILC 2020; authors' illustration.

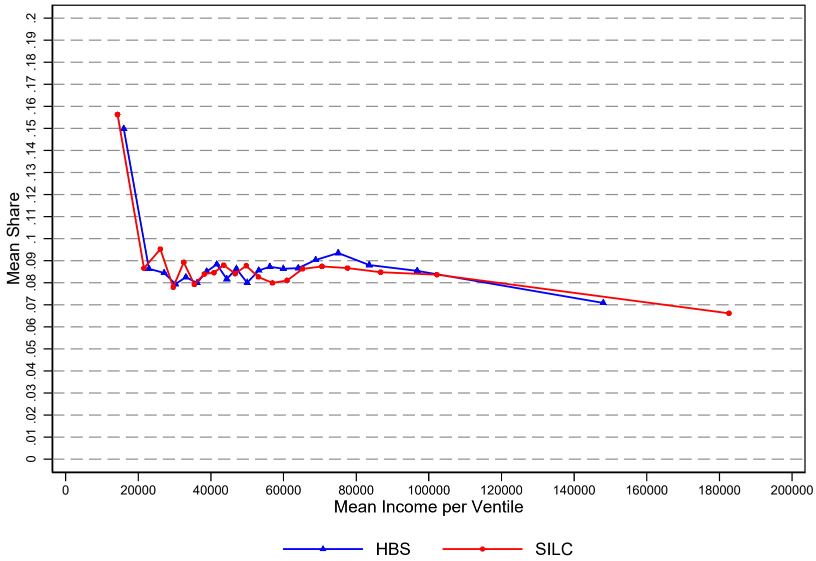

Figure 17. Comparison: average shares of category "Insurance" between HBS and SILC per ventile
Source: HBS 2015–2017; EU-SILC 2020; authors' illustration.

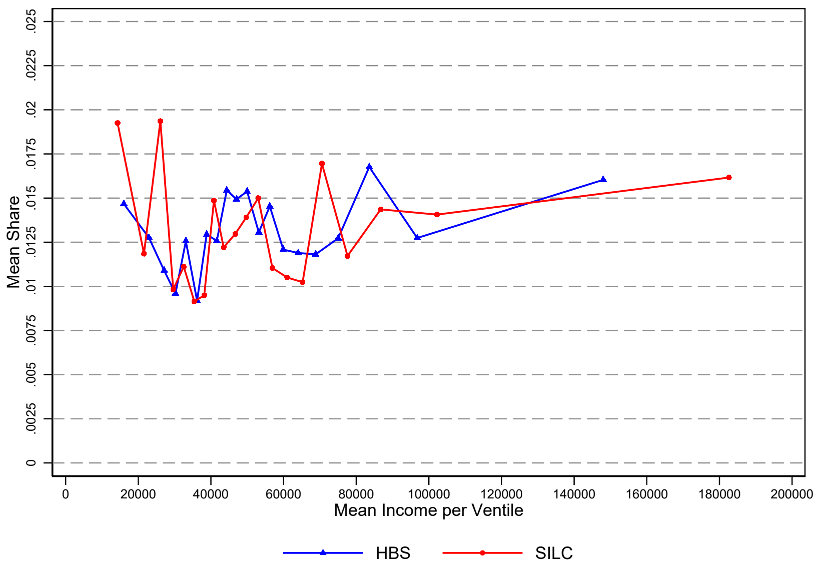

Figure 18. Comparison: average shares of category "Other" between HBS and SILC per ventile
Source: HBS 2015–2017; EU-SILC 2020; authors' illustration.

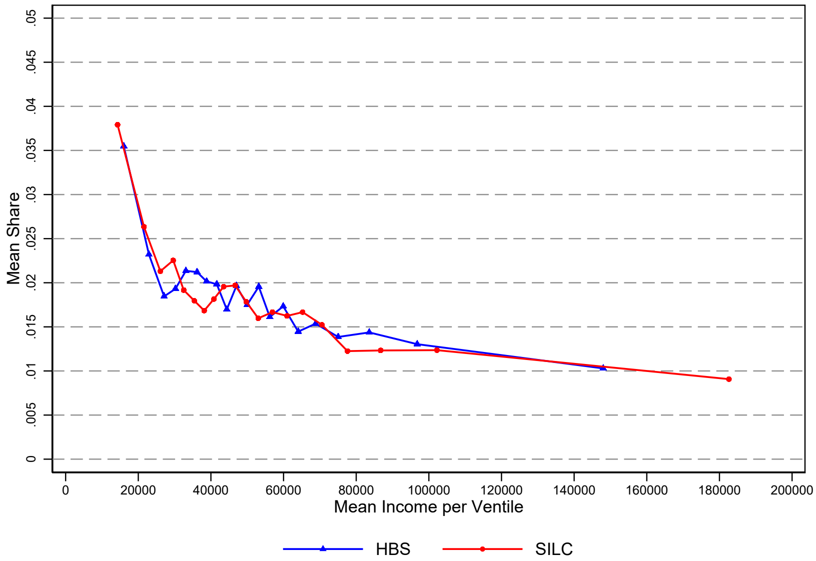

Figure 19. Comparison: average shares of category "Personal care" between HBS and SILC per ventile
Source: HBS 2015–2017; EU-SILC 2020; authors' illustration.

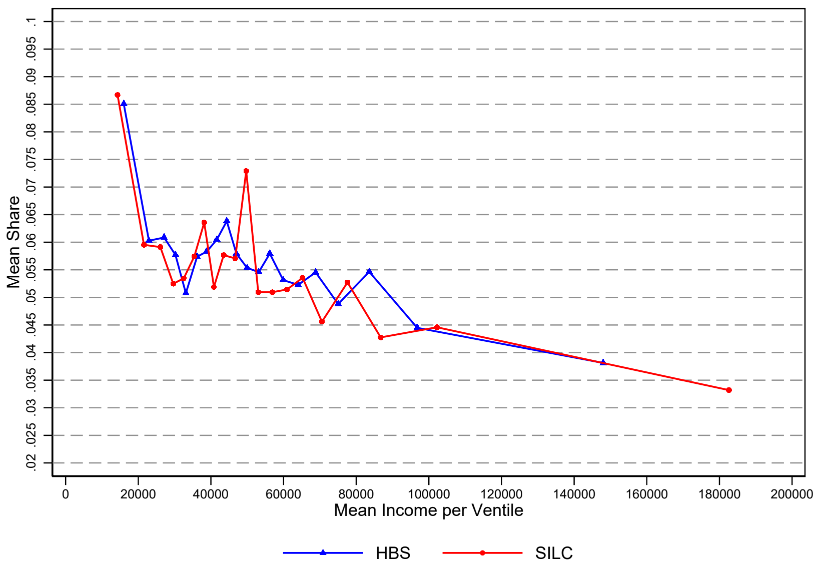

Figure 20. Comparison: average shares of category "Private transport" between HBS and SILC per ventile
Source: HBS 2015–2017; EU-SILC 2020; authors' illustration.

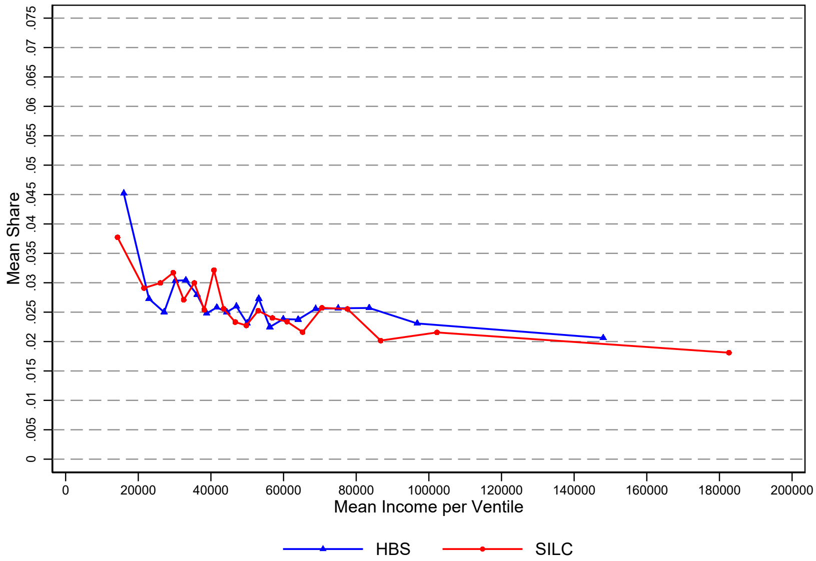

Figure 21. Comparison: average shares of category "Public transport" between HBS and SILC per ventile
Source: HBS 2015–2017; EU-SILC 2020; authors' illustration.

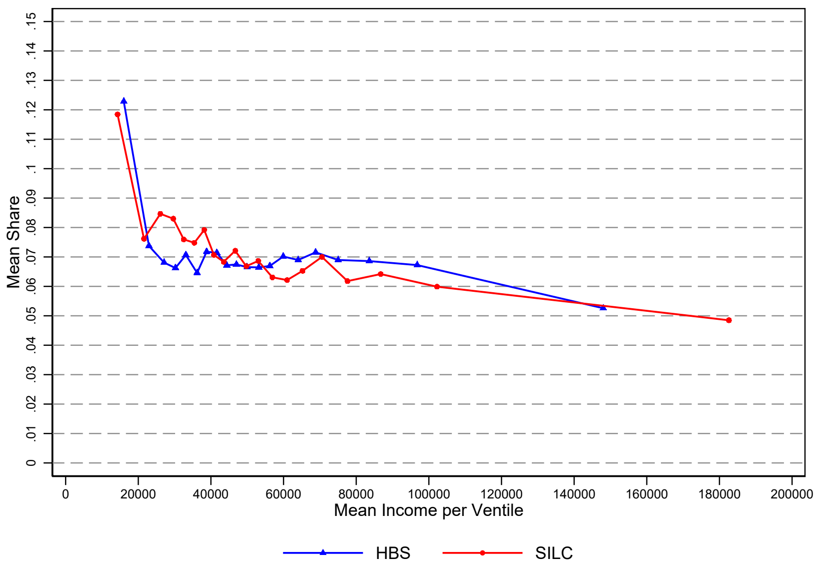

Figure 22. Comparison: average shares of category "Restaurants" between HBS and SILC per ventile
Source: HBS 2015–2017; EU-SILC 2020; authors' illustration.

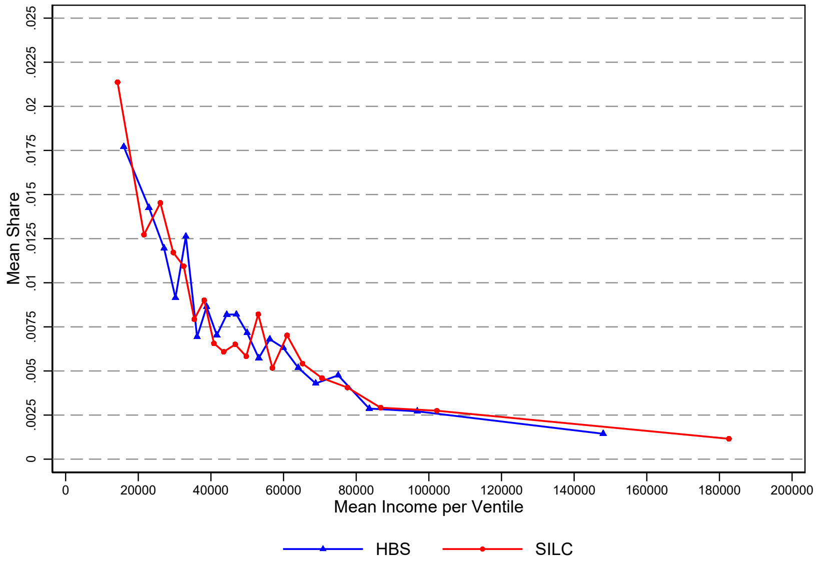

Figure 23. Comparison: average shares of category "Tobacco" between HBS and SILC per ventile

Source: HBS 2015–2017; EU-SILC 2020; authors' illustration.

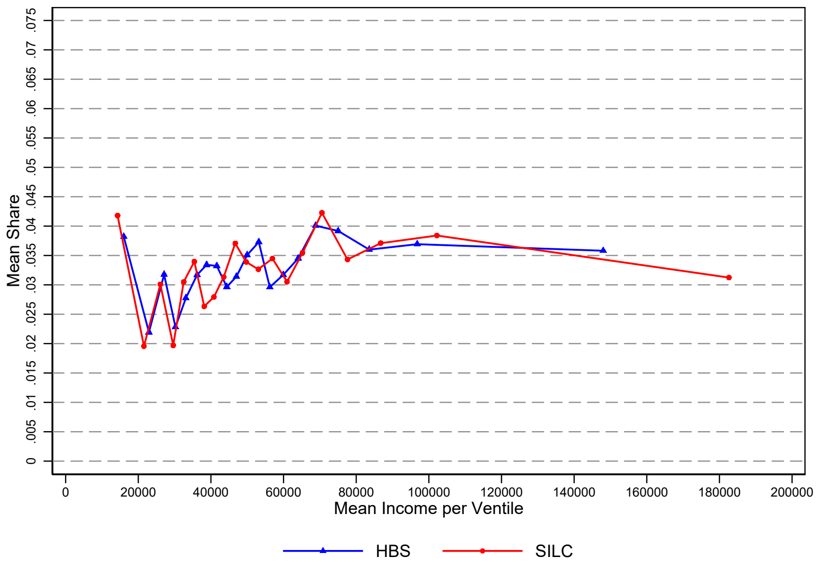

Figure 24. Comparison: average shares of category "Traveling and Holidays" between HBS and SILC per ventile

Source: HBS 2015–2017; EU-SILC 2020; authors' illustration.

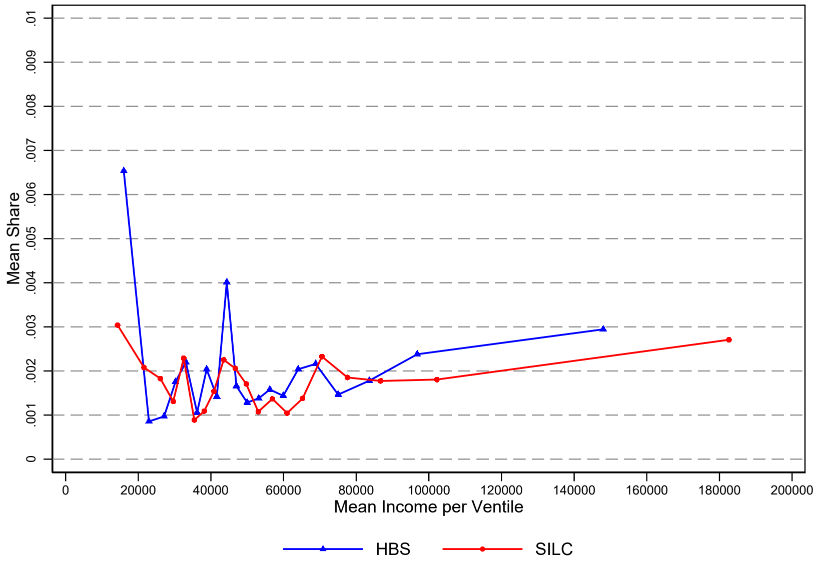

Figure 25. Comparison: average shares of category "Utilities secondary Residence" between HBS and SILC per ventile

Source: HBS 2015–2017; EU-SILC 2020; authors' illustration.

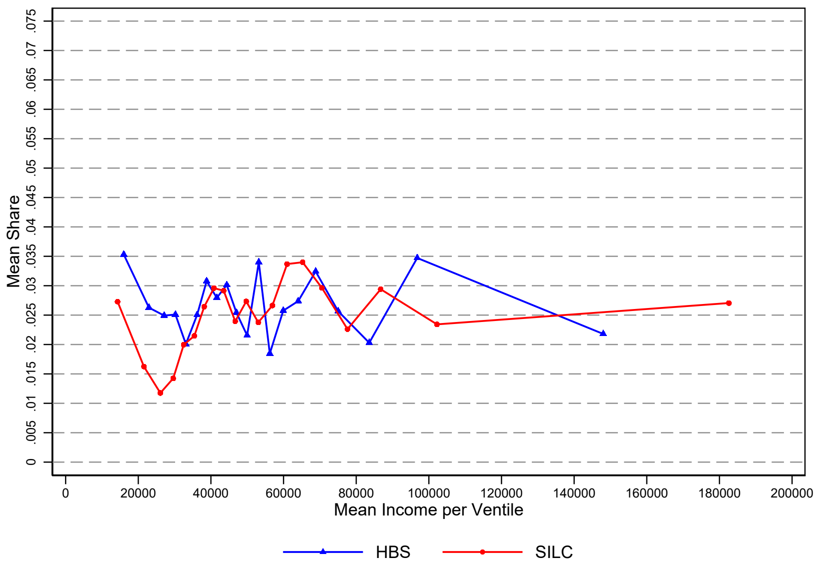

Figure 26. Comparison: average shares of category "Vehicles" between HBS and SILC per ventile

Source: HBS 2015–2017; EU-SILC 2020; authors' illustration.