Dreoni, Ilda; Serruys, Hannes; Manso, Luis; Tudó-Ramírez, José; Amores, Antonio F.

Working Paper

Statistical imputation and validation of consumption microdata for EUROMOD

JRC Working Papers on Taxation and Structural Reforms, No. 2/2025

Provided in Cooperation with:
Joint Research Centre (JRC), European Commission

Suggested Citation: Dreoni, Ilda; Serruys, Hannes; Manso, Luis; Tudó-Ramírez, José; Amores, Antonio F. (2025) : Statistical imputation and validation of consumption microdata for EUROMOD, JRC Working Papers on Taxation and Structural Reforms, No. 2/2025, European Commission, Joint Research Centre (JRC), Seville

This Version is available at:
https://hdl.handle.net/10419/322094

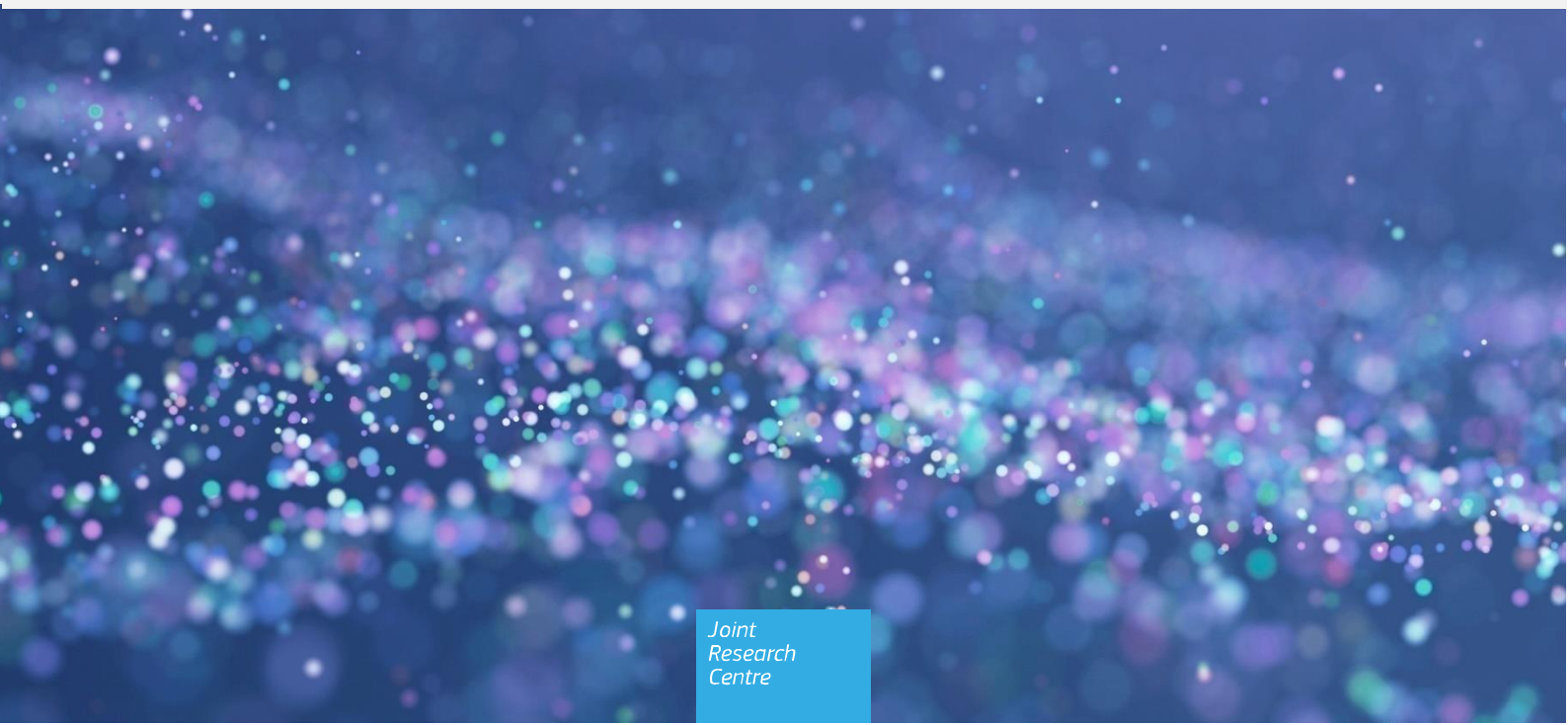
European Commission

Statistical imputation and validation of consumption microdata for EUROMOD

JRC Working Papers on Taxation and Structural Reforms No 2/2025

Dreoni, I., Serruys, H., Manso, L., Tudo Ramirez, J., Amores, A.F.

2025



Joint Research Centre

Contact information
Edificio EXPO, c/ Inca Garcilaso 3, E-41092 Sevilla
JRC-EUROMOD@ec.europa.eu

EU Science Hub
https://joint-research-centre.ec.europa.eu

Seville: European Commission, 2025

How to cite this report: Dreoni, I., Serruys, H., Manso, L., Tudó, J., Amores, A. F., *Statistical imputation and validation of consumption microdata for EUROMOD*, European Commission, Seville, 2025, JRC141513.

Contents

Abstract

Consumption taxes are a crucial revenue source for EU Member States, yet they also potentially have non-negligible impact on income distribution. The EU's tax-benefit microsimulation model, EUROMOD, has recently been extended to simulate consumption taxes (CT) across all 27 EU countries allowing researchers and practitioners to examine carefully their design and assess trade-offs. The CT simulation uses consumption patterns derived from Household Budget Survey (HBS) microdata, which are imputed into EUROMOD's input data using the European Union Statistics on Income and Living Conditions (EU-SILC) microdata which contains detailed socio-demographic and socio-economic information. The imputation process employs a statistical matching procedure that joins HBS (the donor survey) with EU-SILC (the recipient survey) using a predictive mean matching method. Expenditure data are integrated into the recipient survey using a multi-stage procedure that involves the use of estimated probit and linear regression models combined with a distance-hot deck approach for the final observation mapping. This methodology offers enhanced results compared to traditional approaches such as only regression-based or distance-based. The imputation performance in distributional terms and the macro validation of the resulting datasets are thoroughly examined. We assess the impact of potential distortions from the statistical matching process by conducting a set of exploratory and comparative analyses, and also by using an administratively matched dataset for Czechia from 2019 to 2021. Our findings in this specific case indicate that, on average, the majority of imputed expenses are exactly the same when comparing the original HBS data with the matched SILC data that includes fitted expenditures.

Acknowledgements

We are grateful to the previous work by the consortium among Praxis and the Universities of Leuven and Essex to develop the consumption taxes in EUROMOD, as well as all previous developments. Specifically, we are grateful to Paola De Agostini, Sofia Maier, Mattia Ricci, Sara Riscado, Annamaria Maftei, Elif Cansu Akoğuz, Bart Capéau, André Decoster, Liebrecht De Sadeleer, Duygu Güner, Kostas Manios, Alari Paulus, Toon Vanheukelom, Holly Sutherland, Francesco Figari, Chrysa Leventi, Carlo Fiorio, Chiara Gigliarano, Olga Cantó, Jack Kneeshaw, Richard Ochmann, Kevin Spiritus and Nicolas Bouckaert.

We appreciate also the provision of the raw input data from the Household Budget Survey – wave 2015 by Eurostat and Statistics Austria. Specifically, we appreciate the help from Aleksandra Bujnowska, Jakub Hkral and Erika Taidre (Eurostat), and Jerome Olsen and Holger Breiholz (Destatis) for their clarifications on the data. Manuel Tomás (Basque Centre for Climate Change, BC3) provided ideas to address some challenges.

The insightful comments by Salvador Barrios, Paola De Agostini, Sofia Maier, Mattia Ricci and an anonymous JRC internal reviewer to drafts of this work were also very useful. Further, we would like to thank Fidel Picos, as coordinator of the team, and all the developers of the EUROMOD team during 2024 at JRC for their support with the analysis of the validation results: Alberto Mazzon, Andrea Papini, Chrysa Leventi, Hugo Cruces, Katya Bornukova, Klaus Grunberger, Michael Christl, Mattia Ricci, Paola De Agostini, Silvia Navarro, Sofía Maier and Viginta Ivaskaite-Tamosiune.

Authors

Ilda Dreoni

Hannes Serruys

Luis Manso

José Tudó-Ramírez

Antonio F. Amores

Executive summary

This paper provides an overview of the methodology and results of integrating household expenditure data from the Household Budget Survey (HBS) with the European Union Statistics on Income and Living Conditions (EU-SILC) dataset. This integration is crucial for extending the EUROMOD tax-benefit microsimulation model to include taxes on consumption.

Policy context

EUROMOD, the EU27 tax-benefit microsimulation model, has been extended to include taxes on consumption, requiring the integration of household expenditure data from the HBS with the EU-SILC dataset. This extension aims to provide a more comprehensive picture of household behaviour and policy impacts, enabling better decision-making and analysis.

Key conclusions

The statistical matching procedure used to integrate HBS and EU-SILC data has been enhanced. The methodology uses a predictive mean matching method including a two-step regression followed by a non-parametric distance matching.

The validation results show that the matching procedure is able to preserve the values for most household-level expenses, with an average preservation rate of 80-90%. This suggests that the integrated dataset can be used for reliable analysis and simulation of consumption taxes, providing valuable insights for policymakers and researchers.

Main findings

The distributional validation results indicate that the matching procedure performs well in preserving the consumption patterns across different income levels, with a low mean absolute deviation index. The macro validation results show that the matching procedure does not introduce significant distortions compared to the original HBS data, ensuring the accuracy of the integrated dataset.

The ad-hoc validation using an administratively matched dataset for Czechia confirms that the statistical matching procedure is reliable, with a high percentage of correctly matched households and expenses. This provides additional confidence in the accuracy and reliability of the integrated dataset, enabling its use for a range of analytical purposes.

Quick guide

To use the integrated dataset, users can access the EUROMOD software and documentation, which provides guidance on how to run simulations and analyse the results. The dataset is available for most EU countries, and users can select the desired country, year, and expenditure category to run the simulations, enabling flexible and tailored analysis to meet their specific needs and requirements.

1 Introduction

EUROMOD, the tax-benefit microsimulation model of the European Union, has recently been extended with modelling taxes on consumption, (i.e., harmonised excises on energy, alcohol and tobacco, and VAT). Before the introduction of the CT, the direct taxes and social benefits got computed in EUROMOD using a version of to the European Union Statistics on Income and Living Conditions (EU-SILC) datasets adapted to be used in EUROMOD ([1]). EU/SILC contains detailed socio-demographic and socio-economic information, and for which the survey is harmonized across the ESS (European Statistical System) countries.

To compute consumption taxes, the EUROMOD input dataset needs to include information on expenditures of households, which are not included in EU-SILC. Hence, a statistical matching procedure had been developed to augment the input data, derived of EU SILC, with the Household Budget Survey (HBS), which contains detailed expenditure information on households.

This document serves two purposes. It explains the methodology, initially developed by Akoğuz et al. (2020) that has been used in producing the input data and the modifications that have been made to it. Secondly, it discusses the imputation performance in distributional terms, i.e. called distributional validation henceforth, and the macro validation of the produced datasets. The methodology uses a predictive mean matching (PMM) method to impute pattern of consumption, i.e., consumption shares, estimated from the HBS survey to the EUROMOD input dataset.

[1] Depending on the dataset further augmented with National SILC variables.

2 Methodology

To produce the microdata input used for the simulation of consumption taxation, we need to impute expenditure variables from the Household Budget Survey (HBS) to EU-SILC datasets. This section briefly describes the overall methodological approach adopted in this process (Akoğuz et al., 2020) and the modifications implemented to the previous methodology by the JRC team.

The production of the CT microdata is based on a statistical matching procedure that matches HBS (donor survey) to EU-SILC (recipient survey) using a method based on the predictive mean matching (PMM) method. Expenditure data are imputed into the recipient survey using a multi-stage procedure that involves the use of estimated probit and linear regression models combined with a distance-hot deck approach for the final observation mapping.

The probit and linear regression models seek to estimate the relationship between socioeconomic variables and expenditures and capture some form of behavioural relation between household socioeconomic characteristics and their consumption behaviour. The fitted expenditures are then used as input for the second stage of the procedure, the Mahalanobis distance matching. This second stage aims at imputing household expenditures for each household in the EU-SILC dataset while maintaining the variance that characterizes the original expenditures dataset.

For this purpose, we built on the methodology used to produce the input microdata for earlier versions of the CT tool (Akoğuz et al. 2020; Capéau et al. 2022) to create the new augmented CT microdata for 26 countries ([2]) aiming at improving the distributional validation of the matched dataset. The main improvements done by the JRC to the original methodology are as follows:

1. Enrichment of the probit and linear regression models with additional covariates to improve statistical model performance and to better capture the relationship between household socioeconomic variables and their consumption behaviour.
2. Introduction of multiple cut-off values of pseudo-R^2 to define the threshold of inclusion into the distance function.
3. Development of a new distributional validation measure that allows a synthetic measure of the differences between imputed and original expenses across income ventiles, i.e., the distortion created by the statistical process, to be used for the selection of the best model.
4. Production of extensive distributional validation and macro validation results to be used for the validation of the imputed datasets ([3]) including a validation exercise that compares statistically matched dataset with administratively merged dataset, i.e., where recipient and source dataset can be linked directly by using household IDs, for Czechia.

In the following sub-sections, we summarize the imputation procedure (Section 2.1), we describe the data harmonization procedure (Section 2.2.) and discuss in detail the changes applied to both stages of the matching procedure (Section 2.3).

2.1 A step-by-step description of the imputation procedure

This section provides a systematic account of the imputation methodology adopted in the statistical matching procedure used to match consumption shares from the HBS to the EU-SILC datasets.

1. A household *h*'s expenditure on a commodity i in the source dataset (the HBS, indexed by superscript H), denoted by e_{hi}^H, is converted into a share, w_{hi}^H, of disposable income, y_h^H, i.e.:

   [F. 1] $$w_{hi}^H = \frac{e_{hi}^H}{y_h^H},\ i \in N$$

   where *N* is the set of indices of commodities at the most detailed level in the HBS.

[2] Italy is currently excluded as HBS does not have net income variable.

[3] The validation output will be made available together with the dataset used for the CT modelling in EUROMOD.

2. The above expenditures as shares of income are aggregated under twenty broad categories of commodities (cf. Table A.1.1 in Appendix A.1). We index these categories by $C = A, B, \ldots N_C$ is subset of N, denoting the set commodities belonging to category C. Thus, the income shares of expenditure category c, W_{hC}^{H}, is defined as:

$$[\text{F. 2}] \qquad W_{hC}^{H} \equiv \sum_{i \in N_C} w_{hi}^{H}$$

3. Consumption shares of income for aggregated categories, W_{hC}^{h} are regressed against a relevant set of covariates common to both the donor (HBS) and the recipient (SILC) datasets. Although there is no structural interpretation to the regression model, the selection of covariates is inspired to the specification of Engel curves. Note that aggregated categories $C = A, B, \ldots$ may still contain a substantial number of zero observations. At this level of aggregation, these are considered true zeros ([4]). To account for zero expenditures, a two-step regression is performed, as described in points a) and b) below.

   (a) The probability that a household exhibits positive expenditures on commodity aggregate X is modelled by a probit model, using the common variables in the source and recipient dataset as explanatory variables. Formally:

   $$[\text{F. 3}] \qquad Pr(W_{hC}^{H} > 0) = 1 - \varphi(-\gamma_C' x_h^H) = \varphi(-\gamma_C' x_h^H)$$

   where $\varphi(\cdot)$ denotes the standard normal distribution function, x_h^H is the vector of explanatory variables for household *h* in the source dataset *s*, and the vector $\gamma_C^{'}$ contains the parameters to be estimated.

   (b) Next, an ordinary continuous regression model is formulated to assess the relation of positive income shares for broad expenditure categories with the common variables. Formally:

   $$[\text{F. 4}] \qquad W_{hC}^{H} = \beta_C^{'} X_h^H + \varepsilon_h, \; W_{hC}^{H} > 0$$

4. Using the estimated models, income shares spent on the broad categories $C = A, B, \ldots$ are fitted for all households in both the source dataset HBS (H) and the recipient datasets SILC (S), i.e.:

$$[\text{F. 5}] \qquad \widehat{W}_{dhC} = \varphi\left(-\widehat{\gamma_C^{'}} x_{dh}\right)\widehat{\beta_C^{'}} X_{dh}, \; d = H, S.$$

5. Denote a vector of fitted shares retained as input for the distance by $\widehat{W}_{dh} \equiv (\widehat{W}_{dhA}, \widehat{W}_{dhB}, \ldots)$, where $d = H, S$. Using the Mahalanobis distance metric, the distance between a household *h* in the source data, and a household *g* in the recipient data is defined as:

$$[\text{F. 6}] \qquad dist(h, g) = dist\left(\widehat{W}^{H}{}_{h}, \widehat{W}^{S}{}_{g}\right) = \sqrt{(\widehat{W}^{H}{}_{h} - \widehat{W}^{S}{}_{g})' \Sigma^{-1} (\widehat{W}^{H}{}_{h} - \widehat{W}^{S}{}_{g})}$$

   where Σ stands for the variance-covariance matrix of the vector $\widehat{W}$, using data from both source and recipient.

6. A match for household *g* in the recipient dataset is defined as the household *h* in the source dataset that has the smallest distance to household *g*, where the distance is measured in terms of equation [F. 6].

7. For each match (h, g), income shares of expenditures at the most detailed level of goods disaggregation, $i \in N$ for the recipient household *g*, are obtained from the corresponding values of the source household *h*:

$$w_{gi}^{S} = w_{hi}^{H}$$

[4] Which is to say not a consequence of the infrequent expenditure problem.

2.2 Harmonization – COICOP codes and 20 broad expenditures

The harmonization of data is a crucial step to ensure an adequate matching procedure. It occurs in the preparatory phase, prior to the imputation and matching procedure, however it can have a significant impact on the results of the statistical matching procedure, including the validity and reliability of the matched dataset. Expenditure data in HBS is organized using the Classification of Individual Consumption According to Purpose (COICOP) developed by the United Nations Statistics Division (UNSD) to classify and analyse individual consumption expenditures incurred by households. The COICOP is part of a set of classifications of expenditures according to purpose that are part of the System of National Accounts. As such, the classification system has been evolving overtime and across different waves of the HBS. Therefore, the harmonization procedure needs to account for potential differences in COICOP versions across HBS surveys. In our case, we are matching different HBS versions to different years of EU-SILC. For this purpose, we are using two versions of HBS, 2010 and 2015 that use COICOP versions 2003 and 2013 (ECOICOP1) respectively, both developed by Eurostat by adding a level (fifth digit) to the COICOP 1999 from UNSD. The COICOP classification used in HBS 2010 and HBS 2015 have 4 levels and the codes are composed as follows:

coicopcode = WXYZ

where W, X, Y, Z are either digits or blank. For the COICOP classification 2013, which was used for HBS 2015, we have for example:

COICOP-code	Description
CP01 (CPW)	Food and non-alcoholic beverages
CP011 (CPWX)	Food
CP0111 (CPWXY)	Bread and cereals
CP01111 (CPWXYZ)	Rice

We conveniently define the concept of parent and child categories, where a parent category is the category which value is equal to a category with the last digit trimmed away. In our example above CP01 is the parent category of CP011. Conversely, CP0111 is a child category of CP011. In order to ensure a consistent matching procedure, we have harmonized our datasets across multiple steps:

Step 1. Redefinition of demographic variables and harmonization of net income across the two datasets

Relevant demographic variables of EU-SILC get recoded to correspond to the variable definition in HBS, which are defined at the household level.

a) Net income in EU-SILC is converted to euro if needed and adjusted in correspondence to the HBS year by multiplying it with the ratio of the mean in HBS year divided by the mean in EU-SILC year.

Step 2. Ensure consistency of the expenditure structure

Missing values of parent categories are replaced by the sum of their sub-categories – e.g. CP01 = CP011 + CP012 if CP01 is missing and CP011 and CP012 are defined.

b) When data is missing for one child category, there are two cases when we impute its values:

i) Its parent category is not missing and there is just one child category.

	before	after
CP0211 (Spirits and liquors)	15	15
CP02111 (Spirit and liquors)	missing	*15*

ii) Its parent category is not missing and the other child categories are not missing

	before	after
CP0314	15	15
CP03141	missing	10
CP03142	5	5

c) When all of the child categories are defined but the sum of the categories is greater than the parent category, the value for the parent category is replaced by the respective sum.

	before	after
CP0314	15	*19*
CP03141	14	14
CP03142	5	5

To allow for the comparison between matched datasets with different HBS COICOP classifications (2010 or 2015), we produced two version of the harmonized HBS for the year 2015. The first version uses the COICOP 2013 classification that was originally distributed with the 2015 HBS dataset. The second version is re-mapped to COICOP 2003 that was originally distributed with the 2010 HBS dataset. Essentially, the COICOP 2013 classification is a refinement of the COICOP classification of 2003. However, in some cases there was a more fundamental restructuring of the COICOP classification. For example in the case of Austria, a national version of HBS has been used for the production of the matched dataset, and this required a mapping of national COICOP classification to the standard 2003 or 2013 COICOP classification.

2.3 Imputation procedure

The imputation procedure adopted here was adapted from Akoğuz et al. (2020). They use a two-part model approach, a widely accepted modelling approach to deal with data containing observations with zero expenditures (i.e., Deb & Norton, 2018). In the first stage (step 3 and 4 in section 2.1), a pre-specified set of variables homogeneous across all EU countries were chosen based on their availability within each country dataset. In the second stage (step 5 and 6 in section 2.1), a distance metric was calculated using a vector of fitted expenditure shares calculated for both the HBS (donor) and the EU-SILC (recipient). The distance step uses a Mahalanobis distance metric, an effective multivariate distance metric that measures the distance between fitted and observed expenditure shares taking into account also their covariance matrix. The choice of the vector of fitted share expenditures to be included in the distance matrix is based on a synthetic measure, referred to as pseudo-R^2, which measures the amount of variability explained by the combination of probit and linear regression model for each expenditure category. The choice of pseudo-R^2 threshold needs to take into account the trade-offs between reliability, i.e. using only the expenditure categories characterized by a high explanatory power of covariates as inputs for the distance matrix, and coverage, i.e., retaining a higher number of aggregates in the distance function to better preserve the correlation structure of expenditures across aggregated categories. In their work, Akoğuz et al. (2020) chose an arbitrary fixed threshold of 0.1 for the pseudo-R^2 measure after performing some tests on Belgium data using also 0.3 and 0.45 cut-off thresholds.

To produce the new augmented micro dataset for EUROMOD including CT, we expanded and enhanced this methodology in two meaningful ways. First, by adding model interactions deemed as theoretically relevant to improve the explanatory power of the regression-based stage (cf. Table A.1.2 in Appendix A.1). Second, by testing multiple pseudo-R^2 cut-off thresholds (0.05, 0.1, 0.25 & 0.4) to find the best balance between reliability and coverage. Using this enhanced procedure, we produced eight different datasets for most countries. We then select the best-matched dataset based on its comparative empirical performance

according to our synthetic measure of distributional validation, provided that the macro validation of resulting expenditures does not vary too much across the matched dataset options. The distributional validation index is described in more detail in section 3.1.1.

2.3.1 Regression-based stage

The choice of covariates (cf. Table A.1.2) to be included in the regression analysis is based on both the initial list of variables included in Akoğuz et al. (2020) and the overview of summary statistics calculated for HBS and EU-SILC datasets (e.g., a categorical variable that exhibits very little variance has not been included).

For each country we further included few selected interactions which have been pre-selected based both on conceptual assessment and empirical testing of potential improvements of distributional validation for few countries (Austria, Belgium, Cyprus, Germany, France, Czechia). We tested both the inclusion of additional interaction terms within both probit and linear regression model as well as just into linear regression model. We found that the inclusion of additional co-variates in linear regression model performed almost the same as the inclusion of additional variables in both models, consequently we choose to include them just in the linear regression model.

2.3.2 Mahalanobis distance metric and the pseudo-R^2 threshold

To provide an overview of the trade-offs involved with the choice of pseudo-R^2, i.e., between reliability and coverage, Table 1 and 2 present a list of fitted expenditure items that enter the Mahalanobis distance algorithm for each country and each expenditure items for the matched dataset that uses SILC 2022. Table 1 provides a list of each broad expenditure category with the corresponding number of countries for which the category enters the distance function as well as the average pseudo-R^2 value exceeding the 0.1 threshold.

Table 1. Broad expenditure categories considered in the Mahalanobis distance and relative average pseudo-R^2 across countries for SILC year 2022

Category	Number of times selected	Average pseudo-R^2
Food and non-alcoholic beverages	20	0.42
Utilities	19	0.44
Communications	18	0.41
Housing and rental	9	0.29
Health and care	5	0.15
Tobacco	5	0.09
Insurance	4	0.36
Public transportation	4	0.22
Culture and recreation	4	0.12
Education	3	0.49
Alcoholic beverages	3	0.27
Other	3	0.26
Private transportation	2	0.08
Personal care	1	0.41
House goods and services	1	0.29
Travel and holidays	1	0.20
House durables	1	0.15
Restaurants	1	0.07
Clothing and personal items	1	0.06
Vehicles	0	0.00

The regression models for three specific categories, i.e., food and non-alcoholic beverages, utilities and communications seem to perform significantly better compared to other categories for almost all countries. These expenditures categories do not suffer from the infrequent expenditure problem nor the zero-expenditure problem being consumption goods of daily use. Surprisingly, categories such as housing and rental and tobacco, which are expected to suffer from this problem, still retain quite a high predictive power and enter the distance function respectively nine and five times. However, they exhibit a relatively low pseudo-R^2 especially in the case of tobacco. Other categories that suffer from this problem, such as vehicles, travelling and holidays and house durables, very rarely enter the distance function and if so, they exhibit a low pseudo-R^2.

Table 2. Expenditure categories considered in the Mahalanobis distance and relative pseudo-R^2 for each country and SILC year 2022

	AT	BE	CY	CZ	DK	EE	EL	ES	FR	HR	HU	IE	LT	LU	LV	MT	NL	PT	RO	SE	SI	SK
Food and non-alcoholic beverages	0.35	0.33	0.52	0.4	0.44	0.33	0.52	0.32		0.45	0.36	0.37	0.43	0.41	0.29	0.49	0.43	0.27	0.62	0.46	0.56	0.39
Utilities	0.23	0.5	0.45	0.52	0.42	0.34	0.76	0.42	0.46	0.44	0.43		0.21	0.17	0.35	0.69	0.62	0.54	0.09		0.56	0.44
Communications		0.76	0.31	0.61	0.81	0.11	0.66	0.24	0.3	0.19			0.22	0.19	0.32	0.25	0.56	0.33	0.31	0.83	0.42	0.1
Housing and rental	0.29	0.29	0.18	0.11	0.47							0.2	0.07	0.33			0.54	0.2	0.09	0.59		0.1
Health and care											0.2			0.11				0.13	0.07	0.33		0.05
Tobacco										0.07			0.13		0.07	0.09		0.1	0.07			
Insurance		0.15			0.16										0.06		0.8			0.25	0.35	0.33
Public transportation			0.25	0.08			0.48	0.05														
Culture and recreation				0.25	0.11		0.23			0.08						0.06			0.08	0.18	0.26	
Education	0.35										0.35	0.19		0.76								
Alcoholic beverages							0.37			0.05		0.53							0.22			
Other					0.46						0.19					0.14						
Private transportation										0.1			0.06								0.25	
Personal care							0.38					0.41						0.21			0.17	
House goods and services							0.24													0.29	0.12	
Travel and holidays															0.2							
House durables									0.11						0.15					0.2		
Restaurants													0.07									
Clothing and personal items																			0.06			
Vehicles																						

3 Distributional validation and macro validation

Statistical matching is a complex procedure that seeks to provide joint information on variables and indicators that were collected through more than one source. In its simplest format, statistical matching can be regarded as an imputation problem where target variables from the donor sample are imputed in a recipient dataset. Consider the following example of two sample surveys A (recipient) and B (donor) that share a group of common variables X. In contrast, variables Y and Z are only present on samples A and B respectively.

Variables / Dataset	Recipient specific Y	Covariates X	Donor Specific Z
Recipient - A	$y_1 \dots y_n$	$x_1 \dots x_n$	
Donor - B		$x_1 \dots x_n$	$z_1 \dots z_n$
Matched - C	$y_1 \dots y_n$	$x_1 \dots x_n$	$z_1 \dots z_n$

Technically, statistical matching explores the relationship between Y and Z through the set of common variables X in order to obtain a statistically consistent matching between observations in the donor and recipient datasets. This link is then used to enrich the recipient dataset (Y, X) with the donor's information (X, Z). The resulting product is a synthetic dataset (X, Y, Z) that provides joint information on variables and indicators that were not observed together in the first place.

The process becomes increasingly more difficult when we factor in the need to assess the quality of the matching procedure. There are several publications that address the importance of validation (e.g., Kaplan & Turner, 2013); Leulescu & Agafitei, 2012); Gao et al.,(2017); D'Alberto & Raggi,(2023), however, most of them refer to the framework set forth by Rassler (2004). In this framework, the author establishes four levels of validity that matched datasets should strive to achieve in order to verify the quality and justify its use:

Level 1: Preservation of individual values

The matching procedure is capable of preserving true but unobserved values of the donated variables in the recipient dataset such that:

$$z_1 \dots z_n = \tilde{z}_1 \dots \tilde{z}_n$$

Where z_n represents the true but unobserved value of the set of variables z for a given observation and $\tilde{z}_n$ represents the imputed value of the set of variables z for the same observation. Simply put, this level of validation assesses the ability of the matching procedure in imputing accurate values that would have been observed for that observation if the information was collected. This validation method requires the existence of a dataset where we can compare imputed vs observed values for the same household

In Section 4, we show the results of the comparison between a statistically matched dataset and its equivalent administratively merged in order to assess the quality of the matching procedure. Further, we show that our matching procedure is able to preserve true but unobserved values for most household-level expenses (except COICOP 10 - education). Specifically, the current matching procedure is able to preserve the true on average in about 80-90% of the cases.

Level 2: Preservation of joint distributions

The matching procedure is capable of preserving the true and unobserved joint distribution of all variables in the synthetic dataset, such that:

$$\tilde{f}_{X,Y,Z} = f_{X,Y,Z}$$

Where f denotes the observed joint distribution and $\tilde{f}$ denotes the distribution obtained in the synthetic dataset. The most important objective of statistical matching is to generate a synthetic dataset that can be used to make valid statistical inference. In this regard, the validation of the synthetic data should be more concerned with preserving the joint distribution $f_{X,Y,Z}$ rather than preserving individual values. In Section 4, we show the difference in the distribution of selected expenditures across income ventiles (cf. Figure 8, 9, 10) based on a comparison between administratively merged and statistically matched dataset. We provide the percentages gaps across ventiles between the two datasets in order to evaluate the ability of the matching methodology to preserve the joint distribution.

Level 3: Preservation of correlation structures

The matching procedure is capable of preserving the correlation structure, such that:

$$\widetilde{cov}(X,Y,Z) = cov(X,Y,Z)$$

Where cov denotes the covariance matrix observed and $\widetilde{cov}$ the same matrix calculated in the synthetic dataset. These results at aggregated level are presented in section 3.1.3 ([5]).

Level 4: Preservation of marginal distributions

The resulting matched dataset should, at least, preserve the marginal and joint distributions of the variables in the donor sample. In this respect, should ensure $\tilde{f}_Y = f_Y$ and $\tilde{f}_{Y,Z} = f_{Y,Z}$ when imputing Y in a (X,Z) sample. This is often considered the minimum validation requirement for a statistical matching procedure. An example of how the preservation of marginal distribution is assessed is found in sub-Section 3.1.1 and sub-Section 3.1.2.

Finally, we use external information from National Accounts to macro validate the dataset produced (see section 3.3).

In the following, we discuss the main analysis carried out to evaluate the statistical imputation performance and validate our matched dataset. The performance and validation data used in these sections are available for each dataset distributed together with EUROMOD.

3.1 Distributional validation

Our enhanced matching methodology has been used to produce eight different matched datasets for most countries. Out of these eight datasets we choose the best matched dataset based on their empirical performance across the income distribution according to our synthetic measure of distributional validation provided that the differences in the macro validation results of expenditures were negligible. The distributional validation index is described in more detail in the following section.

3.1.1. Distributional validation - distribution of imputed vs observed expenditure shares

The performance assessment of the statistical matching is based on the comparison of imputed and observed distribution of income shares of expenditures across income ventiles as developed by (Akoğuz et al., 2020; Capeau et al., 2022). This comparison is carried out for each of the 20 broad expenditure categories used to impute expenditures (see Table A.1.1).

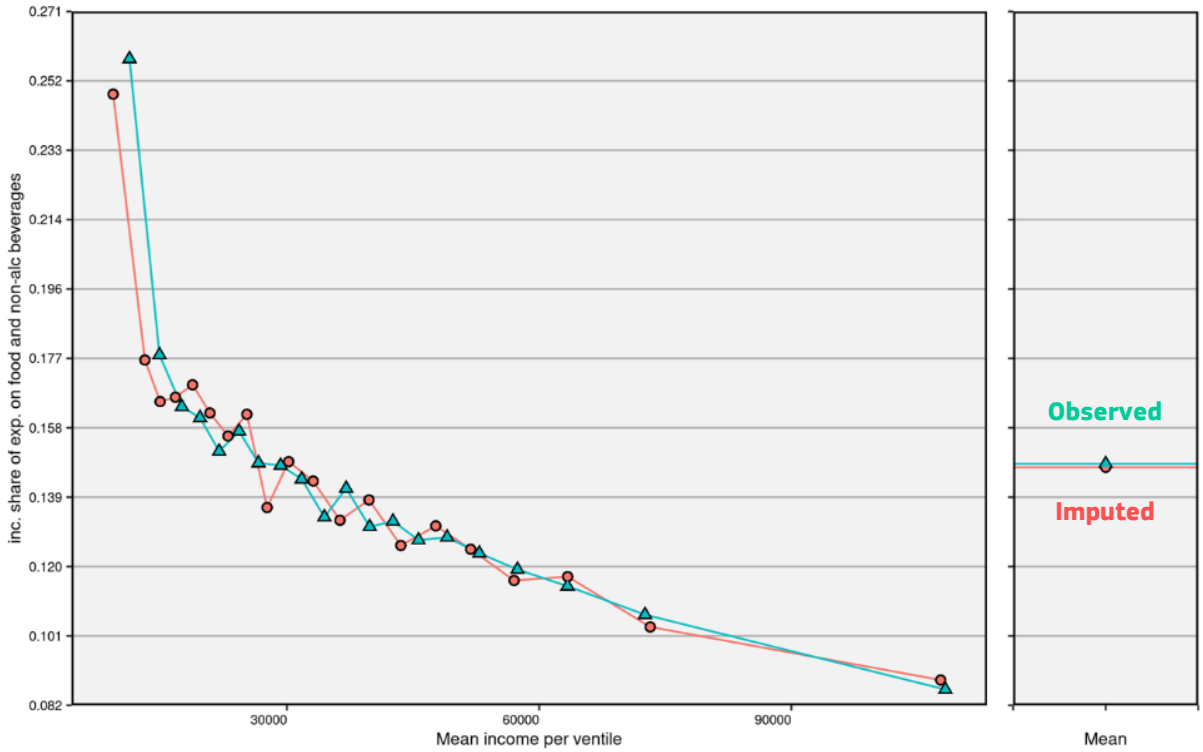
The distribution of expenditure shares over income across income ventiles can be assessed graphically using a ventile diagram (see Figure 1), where income ventiles are plotted on the x-axis and the expenditure shares (of income) on the y-axis for both the original HBS data, and the matched SILC data. The closest the two curves are the best the matching performs because it is able to retain the consumption patterns across different income levels.

Absolute value comparisons, such as Mean Absolute Deviation (MAD), directly measure the absolute deviation between observed and imputed expenditure shares across income ventiles without normalizing by the observed values. This is crucial when dealing with expenditure shares, which can be very small and lead to

[5] There are detailed information on correlation structure of matched dataset for each dataset produced for the CT modelling in EUROMOD which are available together with the dataset. These evaluation data may be available on request as well.

distortions in percentage-based metrics like MAPE or WAPE. Conversely, percentage-based metrics require division by observed values, which can lead to extremely high percentage errors or undefined values when the shares are very small or zero in some categories. For this purpose we have opted to use the MAD to assess the distribution of imputed vs observed expenditure shares.

Figure 1. Ventile Diagram of income shares for food and non-alcoholic beverages (Belgium) observed vs. imputed



For a given expenditure category, the matching procedure will be assessed on a micro-level using a MAD (mean absolute deviation) calculated on the distribution of income shares across ventiles (Miller & Blair, 2009). The MAD is calculated based on this graph and for a specific broad category of consumption is defined as:

$$MAD_X = \sum_{k=1}^{20} \frac{|\omega_{kX}^S - \omega_{kX}^H|}{20}$$

with ω_{kX}^S and ω_{kX}^H, the average share of expenditure for a broad category X within income decile *k*, and *S* denoting the matched SILC dataset and *H* the donor dataset HBS. Note the absolute value operator in the numerator to avoid positive and negative values cancelling each other. We further aggregate the MAD index, which is category-specific, to provide a unique metric to assess each of the 8 model for each country. The aggregation is computed as a weighted average, using the household weights, of the MAD index across all 20 broad categories,

$$\Omega = MAD_X \cdot s_X$$

with s_X the average share of expenditures for the broad category X of the donor dataset. Find an example in Table 3. Note that the best dataset is chosen by ranking the final datasets in terms of the aggregated MAD index Ω.

It can be interpreted as a synthetic measure of how well each imputation process performs in terms of minimising the differences between imputed and observed value disaggregated for each expense category – the lower the value (i.e., the closer to zero) the better the imputation performance for the relative expense category.

Table 3. Example of calculation of MAD index for Belgium SILC year 2022

Category	MAD index	Category	MAD index
cat_food_bev	0.011362	cat_public_trans	0.002901
cat_housing_rental	0.016558	cat_travel_holiday	0.009638
cat_house_goods_serv	0.004848	cat_education	0.002713
cat_utilities	0.009940	cat_vehicles	0.022594
cat_communications	0.003453	cat_house_durables	0.009036
cat_culture_leisure	0.010103	cat_clothing_pers_item	0.003826
cat_pers_care	0.002563	cat_health_care	0.014228
cat_insurance	0.004181	cat_restaurants	0.007907
cat_alcohol	0.002672	cat_other	0.011612
cat_tobacco	0.001294	cat_private_trans	0.009212
Weighted MAD index		**0.009924**	

3.1.2. Distributional validation - distribution of imputed vs observed absolute expenditures

The imputation performance of our matched dataset is also assessed against the distribution of absolute expenditure values. This additional check is needed to evaluate the performance of our matching procedure once the absolute expenditures are retrieved and multiplied by income from SILC. The imputation procedure imputes income shares calculated according to equation F.1 in section 2.1, the income shares are calculated using income from the HBS survey. However, the absolute expenditures used for CT modelling are produced by multiplying imputed income shares by simulated disposable income (including uprating exercises if needed). As such, we produced also the distribution of absolute expenditures across income ventiles for each broad category for each dataset that we produced so that data users can evaluate the quality of the matching and the subsequent consumption taxation modelling. The income used for the production of ventiles of absolute expenditures is disposable income included in EU-SILC survey (i.e. before being refined by EUROMOD simulation). Find in Figure 2 below an example of a ventile diagram of absolute expenditures for food and non-alcoholic beverages (Belgium).

3.1.3. Correlation matrix

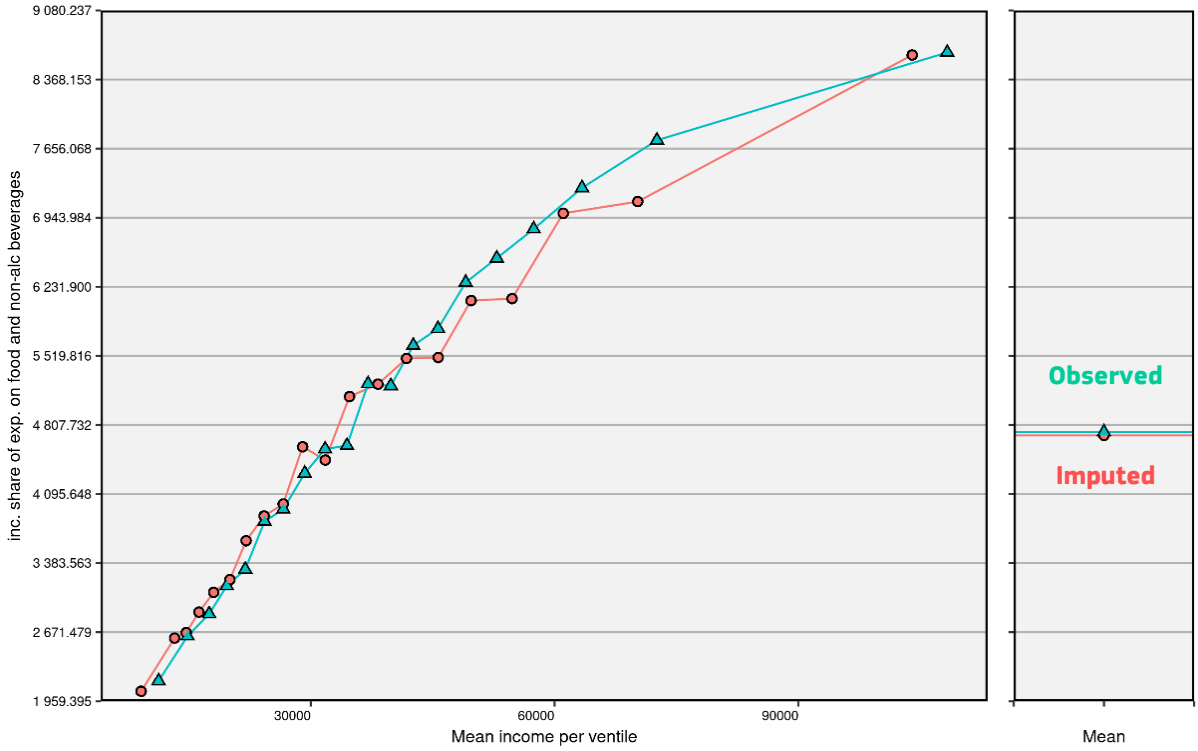
The correlation matrix of the multivariate population distribution – SILC and HBS income, imputed or observed income shares and set of socio-demographic covariates – allow understanding the similarity across SILC and HBS as well as understanding whether both dataset allow us to make inferences about the same population. Specifically, if this holds true then the differences across the two correlation matrices should be equal to 0 (cf. Akoğuz et al., 2020). The correlation matrix is examined by looking at three different aggregates:

1. Mean absolute differences in correlations between HBS and SILC within the common socio-demographic characteristics – this indicator is used to evaluate whether inferences on the socio-demographic household characteristics are similar if based on the HBS or SILC. It provides an indication of whether there are substantial differences between the two surveys, i.e. surveys are administered to different socio-demographic samples, and thus may lead to a bad imputation performance - *Within covariates correlation.*
2. Mean absolute differences in correlations between HBS and SILC within the common socio-demographic characteristics and income shares of 20 broad expenditure categories – this indicator

is used to evaluate whether the relationship between household characteristics and income shares is preserved in the imputed dataset - *Between covariates and expenditures correlations.*

3. Mean absolute differences in correlations between HBS and SILC within income shares of 20 broad expenditure categories – this indicator is used to evaluate how well the correlation within income shares of expenditures on broad categories is preserved in the imputed dataset – *Within expenditures correlations.*

Figure 2. Ventile Diagram of absolute expenditures for food and non-alcoholic beverages (Belgium) observed vs. imputed



In Figure 3, we plot the mean absolute differences of correlations between covariates and expenditures against correlations within covariates. As discussed above, larger differences in correlations within covariates indicates higher differences in the two surveys which will likely lead to a worse imputation independently of the imputation process. Therefore, by looking at Figure 3, we see how countries perform both on the *Within covariates*-dimension and the *Between covariates and expenditures*-dimension. Where lower values on the x and y-axis reflect a better performance. For Denmark, France, Sweden, Ireland and Greece the quality of imputation is deemed to be lower compared to countries that exhibits similar within covariates correlation (e.g., Slovenia, Austria, Luxembourg). These countries exhibit higher differences with respect to between covariates and expenditures correlations indicating a worst performance of the matching methodology in terms of preserving the correlation structure between socio-economics variables and expenditures.

In Figure 4, we shows that even when looking at the relationship between correlations within covariates and within expenditures countries like Denmark, Ireland, Sweden, Greece and France are still those indicated as having lower quality matching results when comparing it to countries with similar correlation structures within covariates (e.g., Slovenia, Austria, Luxembourg). These countries exhibit higher differences with respect to within expenditures correlations indicating a worst performance of the matching methodology in terms of preserving the correlation structure between expenditure categories.

We produced multiple sets of matched datasets for different SILC years using our enhanced procedure. Specifically, we produced five different sets of matched dataset using as donor HBS2015 data and match it with five different SILC dataset for the income years 2015 – 2019 – 2020 – 2021 and 2022, except when any country had availability of SILC for any year. Italy has not been imputed due to the lack of net income data within HBS2015. For Germany, we imputed HBS 2015 just with SILC income year 2015 and 2019. In Table A.2.1 in Appendix A.2, we provide a summary of imputed datasets produced using our methodology.

Figure 3. Mean differences in correlation "within covariates" vs "between covariates and expenditures"

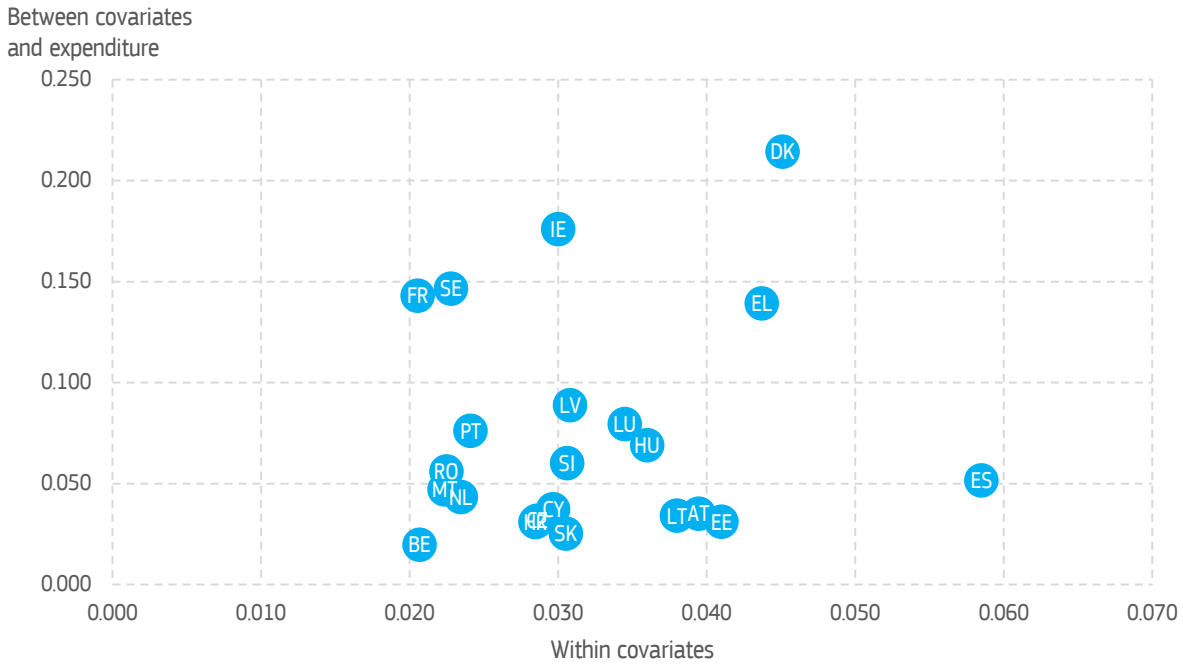
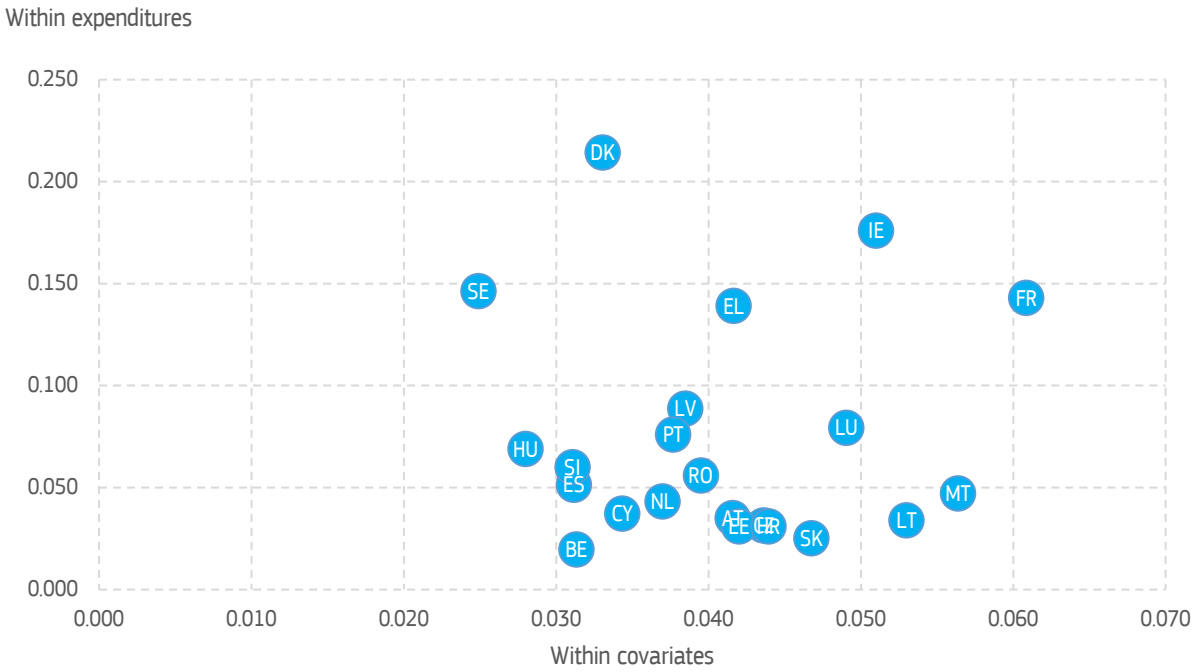


Figure 4. Mean differences in correlation "within covariates" vs "within expenditures"



For each dataset, an extensive set of outputs is produced, so quality of imputation can be assessed in more detail when looking at a specific country/specific expenditure item. Specifically, for each dataset we produce ([6]):

[6] This information are available for each matched dataset that is distributed together with the model.

- Descriptive statistics of variables employed in the imputation process (income, expenditures, expenditure shares as well as set of common co-variates entering the regression stage).
- Results of probit and linear regression for each broad expenditure category.
- Mean expenditure share across income ventiles for each broad category including ventile graphs plotted for quick graphical analysis.
- Mean absolute expenditures (expenditure shares imputed multiplied by SILC income) across income ventiles for each broad category including ventile graphs plotted for quick graphical analysis.
- Differences in the correlation structure to evaluate the preservation of correlation structure.

3.2 Macro validation

The matching process described above produces input microdata for EUROMOD that includes household expenditure by detailed COICOP category. In order to finalize and validate the input data for the consumption taxes (CT) in EUROMOD, we performed comparisons between aggregated survey expenditures and national account expenditures. Specifically, we performed the following comparisons:

1. HBS vs National Accounts (NA) for the same year;
2. EU-SILC year t matched with HBS year t: a) vs HBS year t, and b) vs NA year t

The first and most important validation will compare the results of the matched SILC-HBS to the original HBS to measure if and how aggregate expenditures by COICOP change following the matching procedure. This is the key result to understand how much the matching procedure distort original expenditures.

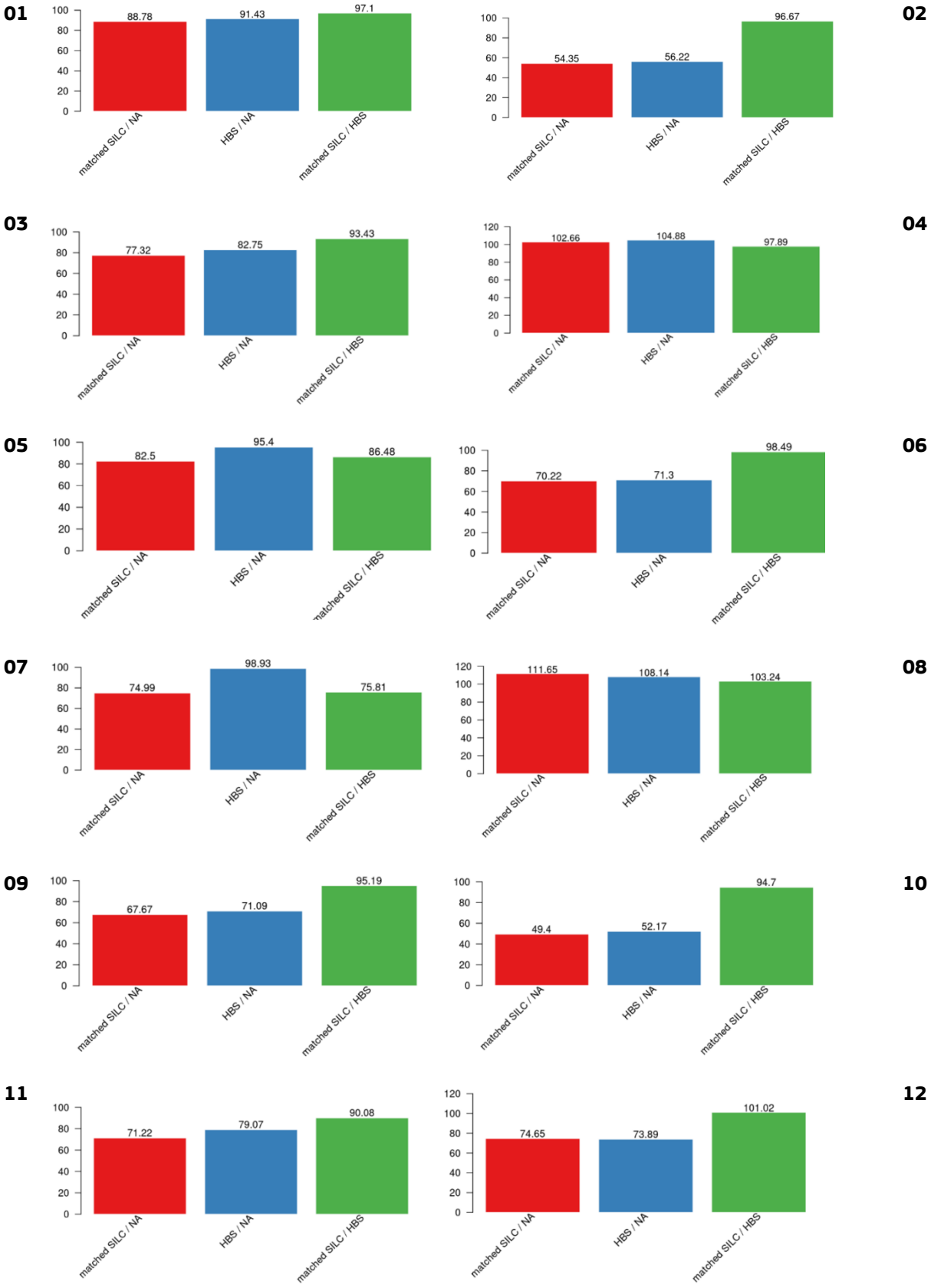
We also perform comparison of HBS vs NA as well as matched HBS vs NA evaluate how the error embedded in HBS influences total coverage compared to NA. This is done with the main aim to provide an overview of the quality of HBS data for the simulation of consumption taxes given that this matched dataset is aimed to be used for this purpose. These two comparisons will help us quantifying the error embedded in HBS. The discrepancies can derive from misreporting and mismeasurements included already in the HBS sample. We cannot correct this baseline difference because it is intrinsic to the HBS sample, but being aware of its existence it is important because it can help us interpret better our results. For example, if we observe that spending on a certain item is already under or over reported in HBS, this may translate in an under/over simulation of expenditure (and correspondent consumption taxes) also in our final simulated result. In this case, we would know that (at least part of the) mis-simulation for this specific category is due to original misrepresentation from HBS. A brief discussion of this macro validation results are included in Appendix A.2.

Box 1: Income down-rating for validation purposes (i.e., comparison with original data)

The year of comparison for the validation exercise is 2015, year of HBS, and to make the comparison between aggregated imputed expenditures and HBS and NA we performed an operation of income down-rating. This is needed because the input data that have been prepared involves matching of SILC data of recent years (2019, 2020, 2021 and 2022) with HBS data from 2015. To deal with differences across survey years we always perform down-rating exercises within statistical matching procedure so to match the income year of the HBS survey. As highlighted in sub-Section 2.2, the income shares that are imputed using our statistical matching methodology are calculated using income figures that gets down-rated to the income year of 2015. The shares are then automatically uprated to the SILC income year as the absolute expenditure values are calculated by multiplying income shares of expenditures with actual income of the reference year. Therefore, when performing modelling with CT we are using consumption data that follows consumption structure of 2015 (as original consumption data are from 2015 year) but under the assumption that the disposable income used by the individuals refers to the SILC reference years. Finally, for macro validation purposes, as we compare with National Account and HBS relative to year 2015, we down-rate income to 2015 year before calculating aggregated absolute expenditures.

In Figure 5, we show the results of macro validation for Denmark for COICOP level 1 categories.

Figure 5. Macro validation by COICOP categories (2 digits) 01 – 12 – DK 2022



Note: 01 Food and non-alcoholic beverages; 02 Alcoholic beverages, tobacco and narcotics, 03 Clothing and footwear; 04 Housing, water, electricity, gas and other fuels; 05 Furnishings, household equipment and routine household maintenance; 06 Health; 07 Transport; 08 Communication; 09 Recreation and culture; 10 Education; 11 Restaurants and hotels; 12 Miscellaneous goods and services.

The matching data do generally well and do not produce any big distortion compared to original HBS which can be noticed especially looking at the green diagram that measures the comparison between matched and HBS (ignoring the NA coverage). In Figure 5, looking at COICOP 02 expenditures (alcohol and tobacco), we notice that those expenses are fully under-reported in HBS and as such results under-reported also in the matched dataset but that does not depend on the matching procedure. Other problematic categories to be signalled regard transport expenses (CP07) where expenditures for CP071 are highly over-reported in HBS while results under-reported when looking at the matched dataset, this is the result of both imprecise starting data as well as distortions happening because of the matching procedure.

In Appendix A.2, we provide a summary of macro validation performed for all our datasets. More detailed macro validation is available upon request. The macro validation of simulated expenditures using simulated disposable income can be performed directly within the EUROMOD software with the macro validation tool.

4 Ad-hoc validation of the statistically matched against an administratively merged dataset for Czechia

We evaluate the relevance of distortions due to the statistical imputation procedure by performing a set of exploratory and comparative analysis using an administratively matched EU-SILC – HBS dataset for Czechia for the years 2019, 2020 and 2021. An administratively matched dataset is a joint dataset where EU-SILC household records can be matched directly with HBS household expenditure information as the sample of households interviewed is the same for both surveys and we can directly link information of the same household across the two datasets. Using this unique set of datasets, we can compare the ability of our statistically matched methodology to be able to match total expenditures to the "right" households.

In the following, we will first show some validation results based on comparison of administratively and statistically merged dataset. Note that SILC 2019 includes income that has been earned in 2018 while HBS expenses includes 2019 expenses, we take account of this chronological mismatch by uprating SILC incomes to 2019 using EUROMOD's uprating factors (average growth rates of wages, pensions, benefits, etc.) yearly updated.

In Table 4, we show the number of exact household matched by the statistical matching procedure. By comparing the administratively merged with the statistically matched we are able to quantify how many times expenditure from a household in HBS are imputed to the same household in SILC, where the best performance would be a 100% match indicating that expenditures from recipient survey, HBS in this case, are correctly assigned to the same household in SILC. Further, in Figure 6, we show the total % of correctly matched expenses at household level for the 12 main categories (COICOP 2-digit level) across the 3 years.

Table 4. Percentage of correctly matched households (comparison of households IDs) across three years

	Sample frequencies			**Population frequencies**		
	2019	**2020**	**2021**	**2019**	**2020**	**2021**
0 - not matching	234	77	194	126,542	45,886	113,770
1 - matching	1,652	1,692	1,283	800,436	846,174	635,983
Total sample (full HBS and SILC sub-sample)	1,886	1,769	1,477	926,978	892,060	749,753
% of matching	88%	96%	87%	86%	95%	85%

In Table 4, we show that the percentage of correctly matched household retrieved is quite high, ranging between 87% and 96% both for sample frequencies as well as population frequencies. This confirms that the distortion produced by the statistical matched procedure is relatively little and it does not prevent us from using the matched dataset for analysis based on expenditures data. Moreover, we show that on average the majority of expenses are exactly the same when comparing original HBS vs matched SILC with fitted expenditures. To be noted that except for COICOP 03 (clothing and footwear), COICOP 07 (transport), COICOP 10 (Education) and COICOP 11 (restaurants and hotels) – the ability of the statistical methodology to retrieve original total expenditures is over 80% indicating a good reliability of the procedure. The only expenditure performing very badly is education which is not subject to consumer taxation.

We further explored differences between administratively merged expenditures and statistically merged expenditures by examining first the differences between statistically and administratively merged aggregated expenses for specific points of distribution (mean and quantiles of expenditure data) in Figure 7. Next, we examined the same differences across income ventiles with the aim to understand whether specific distortions happen for specific ventiles (Figure 8, Figure 9 and Figure 10).

In Figure 7, we show average gaps, i.e., differences between imputed vs original expenses aggregated at COICOP 2-digit level, for specific points of the distribution (mean and quantiles), for selected categories (COICOP 01, 04 and 07).

Looking at Figure 7, we found that for COICOP 01 the distortion is higher when expenditure are little (p5) where the differences reach up to 17% while for higher absolute value of expenses does not seem that there is a big distortion. For COICOP 04 the distortion created by the matching does not seem to happen either for high or low expenses while for COICOP 07 the situation is more varied as there is no similar trend across the different years – this indicate that this category may be more subject to great distortion compared to others due to the expenditure infrequency problem.

Further, Figure 8, Figure 9 and Figure 10 show differences across ventiles for selected COICOP where we can appreciate that for COICOP 01 (food), a necessity good, the biggest gaps that can be noticed is around 5% indicating a low level of distortion. For COICOP 04 (housing and energy), we noticed that the biggest gap is around 10% in 2021 while for COICOP 07 (transport), we noticed biggest average gaps up to 20% however it is to be noticed that most of the biggest gaps are identified toward the end of the income distribution.

Figure 6. Percentage of correctly matched expenses for 12 COICOPs across 2019-2021

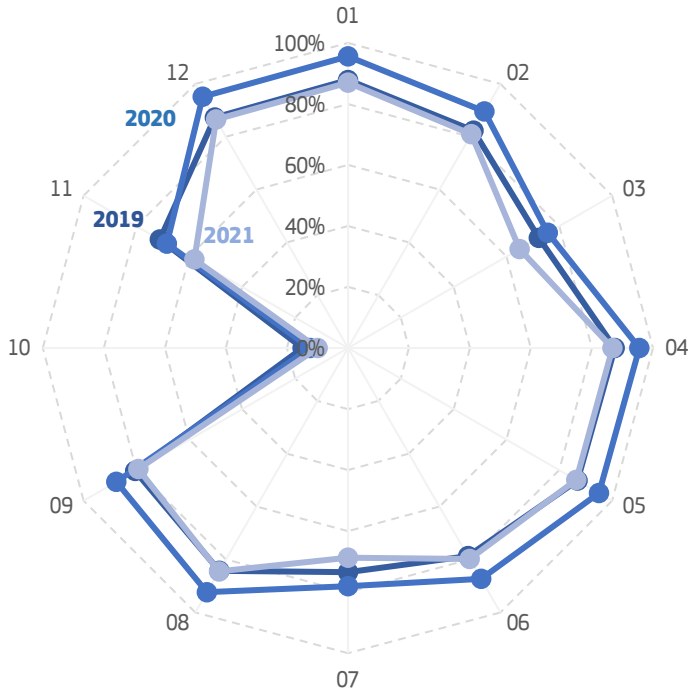
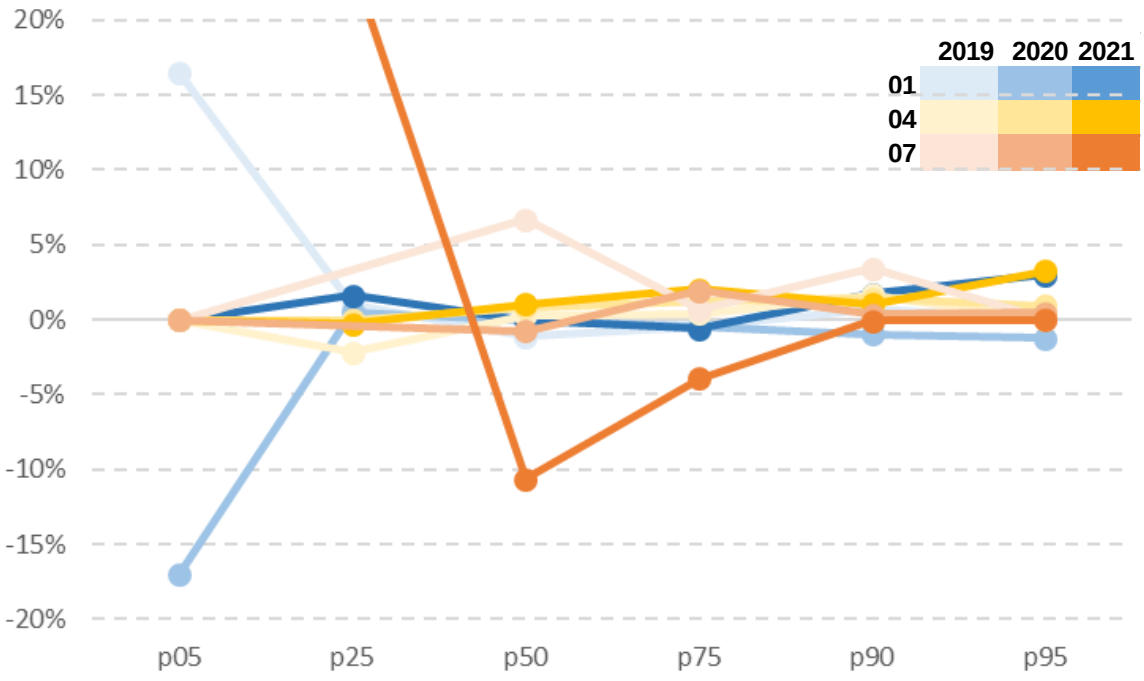


Figure 7. Average gaps across distribution of COICOP 01, 04 and 07 expenditures



Note: pX stands for percentile of expenditure (not household income which is the dimension averaged). Vertical axe truncated for enhance visualization (p05-CP07=57.73%)

Figure 8. Average % gaps across ventiles - COICOP 01, 04 and 07 - 2019

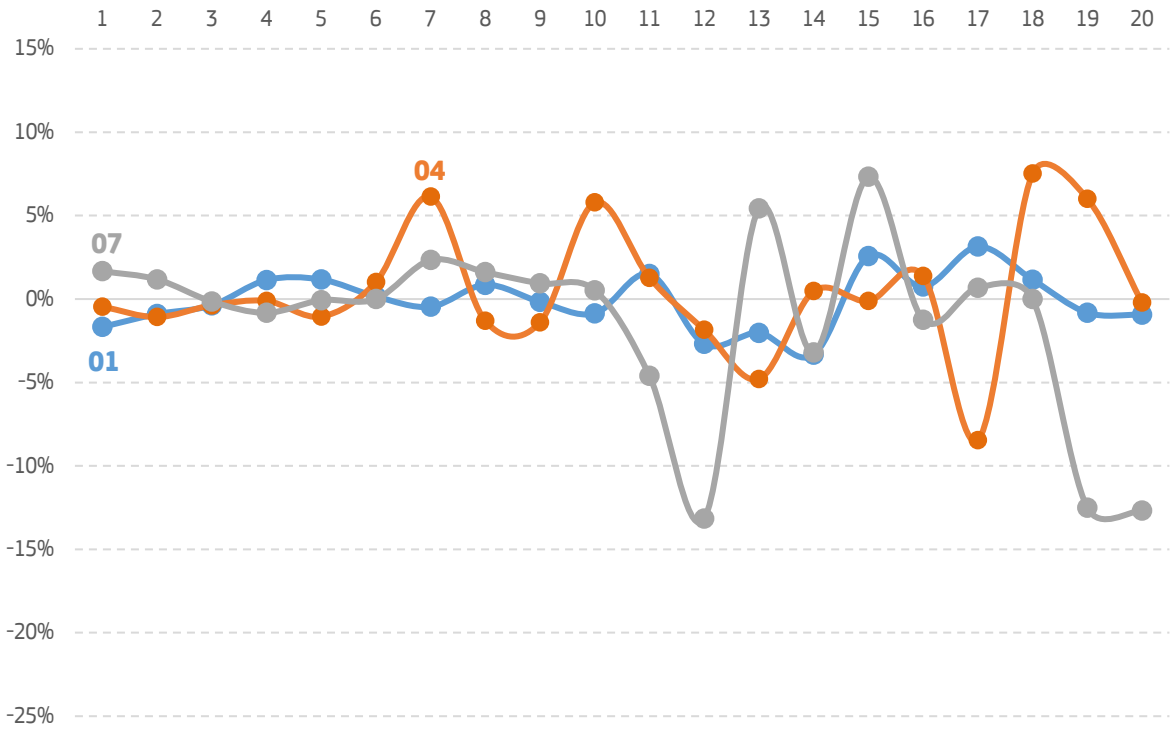


Figure 9. Average % gaps across ventiles - COICOP 01, 04 and 07- 2020

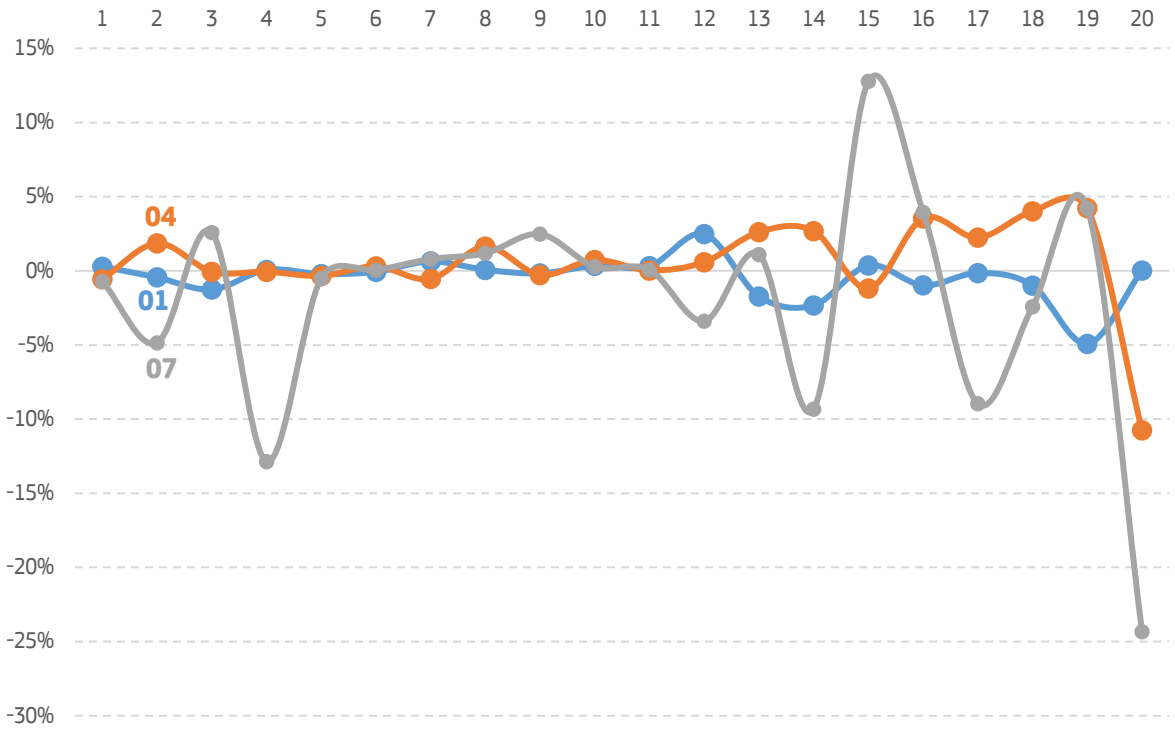
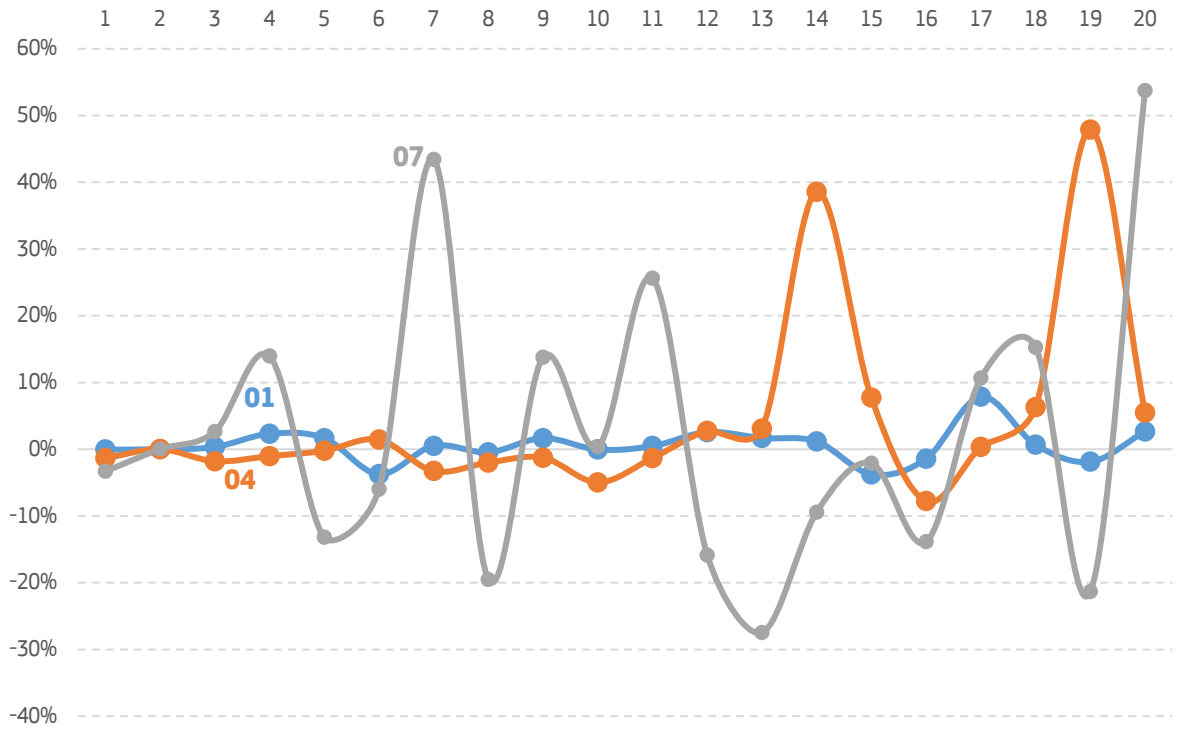


Figure 10. Average % gaps across ventiles - COICOP 01, 04 and 07 - 2021



Note: Please, notice the different vertical scale of this figure respect to two previous ones.

References

Akoğuz, E.C., Capéau, B., Decoster, A., De Sadeleer, L., Güner, D., Manios, K., Paulus, A. and Vanheukelom, T. (2020): A new indirect tax tool for EUROMOD, JRC Technical Report, https://europa.eu/!cbmgYh

Capeau, B., Decoster, A., Gunner, D. (2022) Extension of the EUROMOD-ITT Tool JRC Project no. JRC/SVQ/2020/OP/1373, Final Report.

D'Alberto, R., Raggi, M. (2023) "Integrating rather than collecting: statistical matching in the data flood era"- Stat Papers. Avaliable at: https://doi.org/10.1007/s00362-023-01468-3

Deb, P., Norton, E. C. (2018) "Modelling Health Care Expenditures and Use" – Annual Review of Public Health, Vol. 39. [Online] Available at: https://doi.org/10.1146/annurev-publhealth-040617-013517

Gao, D.; Srikukenthiran, S.; Habib, K. N.; & Miller, E. J. (2017) "Data Fusion: Techniques and Applications" – Report. University of Toronto – Faculty of Applied Science and Engineering [Online] Available at: https://dmg.utoronto.ca/wp-content/uploads/2023/03/Data_Fusion_Report_Final.pdf

Kaplan, D. & Turner, A. (2012) "Statistical Matching of PISA 2009 and TALIS 2008 Data in Iceland" – OECD Education Working Papers No. 78. Available at: https://doi.org/10.1787/5k97g3zzvg30-en

Leulescu, A. & Agafitei, M. (2013) "Statistical Matching: a model based approach for data integration" – EUROSTAT Methodologies and Working Papers, ISSN 1977-0375. Available at: https://ec.europa.eu/eurostat/web/products-statistical-working-papers/-/KS-RA-13-020

Rassler, S. (2004) "Data Fusion: Identification problems, Validity, and Multiple Imputation" – Austrian Journal of Statistics, Vol. 33, Number 1&2 153-171. Available at: https://doi.org/10.17713/ajs.v33i1&2.436

Miller, R.E. & Blair, P.D. (2009) "Input–Output Analysis: Foundations and Extensions". 2nd Edition. Cambridge University Press.

List of figures

List of tables

A.1 Appendix – details of the imputation methodology

A.1.1 Broad consumption categories

Table A.1.1. Broad categories description and correspondence with COICOP classification

	Broad categories description	**COICOP codes**
1	Food and non-alcoholic beverages	01
2	Actual rentals for housing	041
3	Water supply, refuse collection, sewerage collection, electricity, gas and other fuels	0441; 0442; 0443; 045
4	Communication	08
5	Personal care	121
6	Insurance	125
7	Alcoholic drinks	021
8	Tobacco	022
9	Operation of personal transport equipment	072
10	Education	010
11	Clothing and footwear, personal items	03; 123
12	Health products and services; social protection services	06; 124
13	Catering services	111
14	Maintenance and repair of the dwelling; other services relating to the dwelling; furnishings, household equipment and routine maintenance of the house	043; 0444; 0513; 052; 0533; 054; 0552; 056
15	Furniture and furnishings; carpets and other floor coverings; large household appliances; small electrical household appliances; big tools for garden	0511; 0512; 0531; 0532; 0551
16	Culture & Leisure; package holidays	09; 096
17	Transport services	073
18	Purchase of vehicles	071
19	Travel expenses; accommodation services	096; 112
20	Other services	126; 127

A.1.2 Covariates included in regression analysis

Table A.1.2. Set of covariates included in the matching methodology for all 26 countries

Variable	*AT*	*BE*	*BG*	*CY*	*CZ*	*DE*	*DK*	*EE*	*EL*	*ES*	*FI*	*FR*	*HR*	*HU*	*IE*	*LT*	*LU*	*LV*	*MT*	*NL*	*PL*	*PT*	*RO*	*SE*	*SI*	*SK*	*#*
HH disposable income - (3rd degree polynomial)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	26
n adult male HH members	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	26
n HH members age <=14	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	26
n HH members 15 – 29	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	26
n HH members 30 – 44	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	26
n HH members 45 – 59	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	26
n HH members age >= 60	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	25
n employed HH members	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	25
n unemployed HH members	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	24
n pensioned HH members	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	×	✓	✓	24
n disabled HH members	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	×	✓	×	✓	✓	21
n student HH members age >14	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	×	✓	✓	24
n with higher education	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	✓	25
n non–EU citizens	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	×	×	×	✓	×	✓	×	✓	19
reference person farmer	×	×	×	×	×	×	×	×	×	✓	×	×	×	×	×	×	×	×	×	×	✓	×	×	×	×	×	2
region dummies	✓	✓	×	×	×	×	×	×	×	×	×	×	×	✓	×	×	×	×	×	×	×	×	✓	×	×	×	4
Number of Variables	15	15	13	14	13	13	13	14	14	15	13	14	14	15	14	13	14	14	10	12	14	13	14	9	13	14	

A.2 Appendix – list of datasets

A.2.1 Datasets produced as EM input for the CT modelling

Table A.2.1. Datasets produced as EM input for the CT modelling

SILC year	2015	2019	2020	2021	2022
AT	AT_2015_a3_2015_03_n2	AT_2019_b2_2015_03_n2	AT_2020_b2_2015_03_n2	AT_2021_b1_2015_03_n2	AT_2022_b1_2015_03_n2
BE	BE_2015_a1_2015_03_e2	BE_2019_c3_2015_03_e2	BE_2020_c2_2015_03_e2	BE_2021_c6_2015_03_e2	BE_2022_c1_2015_03_e2
BG	BG_2015_b3_2015_03_e2	BG_2019_c2_2015_03_e2	BG_2020_c1_2015_03_e2	BG_2021_c1_2015_03_e2	BG_2022_c1_2015_03_e2
CY	CY_2015_a1_2015_03_e2	CY_2019_a1_2015_03_e2	CY_2020_b2_2015_03_e2	CY_2021_b3_2015_03_e2	CY_2022_b1_2015_03_e2
CZ	CZ_2015_a3_2015_03_e2	CZ_2019_b1_2015_03_e2	CZ_2020_b1_2015_03_e2	CZ_2021_b1_2015_03_e2	CZ_2022_b1_2015_03_e2
DE	DE_2015_a1_2015_03_e2	DE_2019_a2_2015_03_e2	DE_2020_b1_2015_03_e2	DE_2021_b2_2015_03_e2	DE_2022_b1_2015_03_e2
DK	DK_2015_a2_2015_03_e2	DK_2019_a2_2015_03_e2	DK_2020_b1_2015_03_e2	DK_2021_c1_2015_03_e2	DK_2022_c1_2015_03_e2
EE	EE_2015_c3_2015_03_e2	EE_2019_c1_2015_03_e2	EE_2020_f1_2015_03_e2	EE_2021_f2_2015_03_e2	EE_2022_f1_2015_03_e2
EL	EL_2015_a3_2015_03_e2	EL_2019_a2_2015_03_e2	EL_2020_c2_2015_03_e2	EL_2021_c1_2015_03_e2	EL_2022_c2_2015_03_e2
ES	ES_2015_a1_2015_03_e2	ES_2019_b1_2015_03_e2	ES_2020_b1_2015_03_e2	ES_2021_b1_2015_03_e2	ES_2022_b2_2015_03_e2
FI	FI_2015_a3_2015_03_e2	FI_2019_a1_2015_03_e2	FI_2020_b1_2015_03_e2	FI_2021_b1_2015_03_e2	FI_2022_b1_2015_03_e2
FR	FR_2015_a2_2015_03_e2	FR_2019_c2_2015_03_e2	FR_2020_c2_2015_03_e2	FR_2021_c2_2015_03_e2	FR_2022_c1_2015_03_e2
HR	HR_2015_a4_2015_03_e2	HR_2019_b1_2015_03_e2	HR_2020_b3_2015_03_e2	HR_2021_b2_2015_03_e2	HR_2022_b2_2015_03_e2
HU	HU_2015_a2_2015_03_e2	HU_2020_b1_2015_03_e2	HU_2020_b1_2015_03_e2	HU_2020_b1_2015_03_e2	HU_2020_b1_2015_03_e2
IE	IE_2015_a2_2015_03_e2	IE_2019_a1_2015_03_e2	IE_2020_b1_2015_03_e2	IE_2021_b1_2015_03_e2	IE_2022_b1_2015_03_e2

SILC year	2015	2019	2020	2021	2022
IT	[not matched]	[not matched]	[not matched]	[not matched]	[not matched]
LT	LT_2015_a1_2015_03_e2	LT_2019_a3_2015_03_e2	LT_2020_a1_2015_03_e2	LT_2021_c1_2015_03_e2	LT_2022_c1_2015_03_e2
LU	LU_2015_a1_2015_03_e2	LU_2019_a1_2015_03_e2	[SILC not available]	LU_2021_b1_2015_03_e2	LU_2022_b1_2015_03_e2
LV	LV_2015_a3_2015_03_e2	LV_2019_a1_2015_03_e2	LV_2020_b3_2015_03_e2	LV_2021_b2_2015_03_e2	LV_2022_b2_2015_03_e2
MT	MT_2015_a1_2015_03_e2	MT_2019_b1_2015_03_e2	MT_2020_b1_2015_03_e2	MT_2021_b1_2015_03_e2	MT_2022_b1_2015_03_e2
NL	NL_2015_a2_2015_03_e2	NL_2019_b3_2015_03_e2	NL_2020_b2_2015_03_e2	NL_2021_b1_2015_03_e2	NL_2022_b1_2015_03_e2
PL	PL_2015_b1_2015_03_e2	PL_2019_b3_2015_03_e2	PL_2020_b2_2015_03_e2	PL_2021_b1_2015_03_e2	[SILC not available]
PT	PT_2015_a1_2015_03_e2	PT_2019_a1_2015_03_e2	PT_2020_a1_2015_03_e2	PT_2021_b1_2015_03_e2	PT_2022_b2_2015_03_e2
RO	RO_2015_a1_2015_03_e2	RO_2019_a1_2015_03_e2	RO_2020_a1_2015_03_e2	RO_2021_b1_2015_03_e2	RO_2022_b1_2015_03_e2
SE	SE_2015_a1_2015_03_e2	SE_2019_a1_2015_03_e2	SE_2020_b1_2015_03_e2	SE_2021_b1_2015_03_e2	SE_2022_b1_2015_03_e2
SI	SI_2015_a2_2015_03_e2	SI_2019_a1_2015_03_e2	SI_2020_c2_2015_03_e2	SI_2021_c2_2015_03_e2	SI_2022_c1_2015_03_e2
SK	SK_2015_a1_2015_03_e2	SK_2019_a1_2015_03_e2	SK_2020_b1_2015_03_e2	[not matched]	SK_2022_b1_2015_03_e2

A.3 Appendix – Country-specific macro validation summaries

A.3.1 SILC year – 2022

Austria

The HBS data seems to underestimate consumption in National Accounts for Austria, but matching improves coverage by twenty percentage points. Matching enhances the overall coverage of NA consumption, but also introduces overestimation in comparison to HBS across various subcategories.

There are no instances of underestimation when examining two-digit COICOP codes, but there are cases of underestimation at the three-digit subcategory level, particularly for CP104 and CP121.

The matching procedure inflates expenses across almost all COICOP categories compared to the original HBS, but given the overall underreporting of expenses in HBS, matched SILC generally remains below NA and approaches its values in several cases.

There is severe overestimation in matched SILC/HBS (exceeding 30%) for CP03, CP04, CP05, CP08, CP12. This seems to be driven by overestimation in the following subcategories: CP031, CP045, CP051, CP054, CP081, CP083, CP124, CP125.

Belgium

The matching performs relatively well for expenditures at the one-digit aggregate COICOP categories. Expenditures imputed from HBS to SILC do not depart more than 20% from the original source survey (HBS) for all COICOP one-digit aggregate categories except for CP06 (health) and CP12 (others), where the imputed expenditures are about 122% of those reported in HBS, followed by CP07 (114.5%).

When we compare these matched expenditures in SILC (mSILC) with expenditures from National Accounts (NA), gaps are larger. The mSILC/NA coverage ranges from 45.7% in CP02 -alcohol and tobacco– to 119/121% in CP07/08 -i.e. transport and communication.

The low coverage of category CP02 (alcoholic beverages and tobacco) comes entirely from the under-reporting in HBS (the ratio HBS/NA is 43%, even slightly lower than mSILC/NA). This is not a Belgium-specific problem, as these types of products tend to be under-reported in household surveys. In fact, the HBS/NA coverage in Belgium is quite above the EU average (35.6%). When we look at the specific products within this category, we see that the consumption of tobacco (CP02.2) is the main driver of this under-reported consumption in HBS with respect to NA (27.3%), whereas the coverage for alcoholic beverages (CP02.1) is still below 100% but much higher, of 72.2%.

The overestimation of expenditures in category 08 (communication) is also mainly driven by the original over-reporting (an HBS/NA ratio of 114%), which is exacerbated after the matching. The distortion introduced by the matching is mainly observed in category CP 08.1 (postal services): expenditures in the matched SILC double those reported in the original HBS file. The rest (CP08.2 and CP08.3 - i.e., telephone equipment and services) are both over-reported in HBS w.r.t NA and get slightly reinforced with the match.

In contrast, the over-estimation of expenditures in CP07 with respect to NA is mainly driven by the matching, as the original HBS/NA ratio is 104%. Looking at the components of CP07, we can clearly see that this over-reporting of HBS w.r.t NA comes from category CP07.1 (purchase of vehicles).

Bulgaria

Overall, on average, the matching deviates significantly from HBS in terms of aggregate expenses (%). The deviation is about 24%. There is a general and severe underestimation of total expenses by HBS over NA (%). The underestimation is deemed severe (> 30%) for all first-level COICOP groups except CP03, CP04, CP05, and CP12.

At the 3-digit COICOP level, the underestimation in HBS over NA remains strong. We observe large deviations (greater than 30%) in subcategories of CP012, CP021, CP022, CP031, CP041, CP051, CP052, CP053, CP054, CP055, CP071, CP072, CP073, CP081, CP082, CP091, CP092, CP093, CP094, CP095, CP096, CP104, CP105, CP111, CP112, CP123, CP125, CP126, CP127.

At the 3-digit COICOP level, the matching performs well (deviations below 15%) except in categories CP041, CP054, CP055, CP063, CP072, CP081, CP082, CP091, CP092, CP093, CP096, CP102, CP104, CP105, CP111, CP112, CP123, CP126, CP127.

Cyprus

Overall, it seems that HBS underreports consumption in NA for Cyprus (weighted average of the shares of aggregate expenses of HBS/NA is 76.5%). The matching process generally results in improved coverage of NA consumption (85.3%) with respect to the performance of original HBS across some COICOP groups and their two-digit level subcategories.

Notably, in groups CP06 and CP12, the matching inflates the original HBS figures. This trend is particularly evident in specific three-digit subcategories such as CP061, CP123, CP126 and CP127, where the reported expenses after matching exceed those originally reported by HBS by varying extents.

Similarly, in CP10, subcategory CP102 also shows a decrease in reported expenses post-matching. This case of underestimation highlights that the matching process does not uniformly increase expense figures and can differentially impact detailed product groups and subcategories.

Overall, the matching procedure yields results that are relatively close to those of the HBS.

Czechia

The HBS data appears to underreport consumption in National Accounts (NA) for CZ (weighted average of the shares of aggregate expenses of HBS/NA is 59.7%). Matching improves coverage of NA consumption (64.7%). The matched SILC/HBS figures tend to overestimate consumption relative to the original HBS figures (111.4%).

Several subcategories within CP01 to CP12 show underreporting in SILC/NA: Severe in CP126 (14.1%), CP124 (18.6%), CP022 (15.8%), CP112 (28.1%), while moderate in many subcategories (30-50% range). These highlight areas with potentially significant underreporting in the original HBS data, however keeping in mind that HBS significantly underreports NA overall.

The matching procedure inflates expenses compared to the original HBS, evident in several subcategories. However, given the overall underreporting of expenses in HBS, SILC generally remains below NA.

Notable Overestimation in SILC/HBS: CP127 (241.1%), CP104 (196.1%), CP081 (195.0%), CP051 (131.8%) present the most significant increases post-matching. Yet overestimation in SILC/NA is almost negligible for these subcategories, with the highest (128.6%) in CP127.

Moderate Overestimation: Many subcategories fall in the 10-30% increase range, indicating a widespread impact of the matching process. It is evident in categories CP10, CP11 and CP07. Yet SILC/NA remains underestimated.

The matching procedure improves the overall coverage of NA consumption, but it also introduces overestimation relative to HBS across many subcategories. Yet the mismatches between HBS and NA are more significant and spread across all categories, highlighting possible inaccuracies in the underlying HBS data collection.

Denmark

The matching data do generally well and do not produce any big distortion compared to original HBS (most of them are <30%). There are few exceptions to this which includes 3-digit level of CP063, CP071, CP081 and CP126 where the matching produces a more severe distortion. COICOP 02 expenditures (alcohol and tobacco) are fully under-reported in HBS and as such results under-reported also in the matched dataset but that does not depend on the matching procedure. Other problematic categories to be signalled regard transport expenses (CP07) where expenditures for CP071 are highly over-reported in HBS while results under-reported when looking at the matched dataset, this is the result of both imprecise starting data as well as distortions happening because of the matching procedure.

General issues related to HBS original data regards: over-reporting for CP125, CP043 and under-reporting (more frequent) for CP022, CP063, CP091, CP092, CP094, CP105, CP124, CP126.

Estonia

The matching data do generally well and do not produce any big distortion compared to original HBS. At the level one, distortions are generally of the order of 10%, whereas at the level 2 they are typically below 30%. A few exceptions to this includes CP081, CP102 and CP127 where the matching produces severe distortions (beyond 50%). When compared to national account, COICOP 02 (alcohol and tobacco), COICOP 03 (clothing and footwear), COICOP 11 (restaurant and accommodation services) and CP12 (miscellaneous good and service) are significantly under-reported in HBS and as such results under-reported also in the matched dataset but that does not depend on the matching procedure.

Greece

Overall, HBS seems to be underreporting the consumption in NA for Greece (in all categories apart from CP06). Underreporting becomes even more pronounced when we compare the matched SILC with NA, with seven categories exceeding the acceptable range of 30% (CP02, CP03, CP07, CP08, CP09, CP11 and CP12).

The matched SILC/HBS rates are all in the range of 90%-101% for all categories apart from CP06.

CP06 seems to be the most worrisome category: the overall consumption reported in HBS is very close to the one in NA, but the level 2 analysis shows that in item CP063 HBS there is a significant overestimation (HBS/NA ≈ 300%). The matching process creates a significant overestimation of this category's overall expenditure (matched SILC/HBS ratio is at 171.6%).

Spain

Overall, HBS seems to underreport consumption compared with National Accounts data for all COICOP categories.

However, overall, underreporting becomes less pronounced with the statistically matched dataset of the EU-SILC with HBS, compared to National Accounts. Four COICOP categories have deviations larger than 30% (CP02, CP09, CP11 and CP12). From the remaining eight categories, two closely match National Accounts (CP04 and CP08).

The statistically matched dataset of the EU-SILC with HBS is similar to HBS data in this regard, with all COICOP level 1 categories over reported between 5 to 25%, except for CP02 for which there is a strong underreporting.

Finland

In general, the HBS appear to be slightly underestimating consumption figures compared to NA, with a weighted average of aggregate expenses shares reaching 82.1%. The matching process generally results in an additional underrepresentation of NA consumption coverage compared to the original HBS data across various three-digit subcategories, exacerbating the issue of consumption underreporting for some specific categories.

Particularly noteworthy is the substantial overestimation of consumption in specific three-digit subcategories such as CP081 and CP127, where reported expenses after matching significantly exceed those originally reported by HBS.

Despite the overall underestimation in coverage, there are instances of increased coverage. In COICOP groups CP06 and CP12, all subcategories report higher expenses after matching compared to HBS. Furthermore, in other COICOP groups, three-digit subcategories exhibit a decrease in reported expenses post-matching. Notably, in CP123 the matching adjustment leads to a significantly higher expenditure than in HBS. These cases of underestimation underscore the non-uniform impact of the matching process on expense figures across detailed groups and subcategories.

France

The original HBS data seems to be generally underestimating consumption figures over NA, with a weighted average of aggregate expenses shares equal to 74.8%. However, the only first level COICOP groups deemed as severe underreported (> 30%) are: CP02, CP06 and CP11.

The matching process generally results in an additional underestimation of NA consumption coverage compared to the original HBS data (70.01%). The severe underreporting persist for the mentioned three

COICOP categories (particularly worsening for CP02 but remaining similar for CP06 and CP11), as well as CP09, when comparing the HBS matched to SILC dataset to the NA.

Moreover, the matched SILC/HBS rates are in the acceptable range of 80.03%-118.31% for all COICOP level 1 categories. On the 3-digit COICOP level, the matching performs well, where we do not have deviations larger than 30%.

Croatia

Overall, HBS seems to be underreporting the consumption in NA in all COICOP categories.

Underreporting becomes even more pronounced when we compare the matched SILC with NA, with only three categories being within the acceptable range of 30% (CP01, CP03 and CP04). From the remaining nine categories, the most underreported is CP11, followed by CP02, CP09 and CP06.

The matched SILC/HBS rates are in the (acceptable) range of 92%-112% for all COICOP level 1 categories.

Moving to level 2, COICOP categories CP054, CP081, CP102, CP104, CP124 and CP127 are showing matched SILC/HBS rates above the acceptable range of 120%.

Hungary

Overall, on average the matching is extremely close to HBS in terms of aggregate expenses (98%). There is a general underestimation of total expenses by HBS over NA (59%). The underreporting from HBS is greater than 30% (severe) for most first level COICOP groups except CP04 (Housing, water, electricity, gas and other fuels) that is over-reported by about 46% and CP08 (communication) that results accurate compared to the national account figures.

The matching process reproduces the same patterns of consumption as for HBS in SILC and aggregates expenditures coverage at COICOP level 1 is very similar (matched SILC on average captures 61% of NA).

The matching performs well also at 3-digit COICOP level: it replicates closely HBS coverage with an average overestimation of 11%. However, users should be aware that because of the underreporting of many expenditures in HBS representing only 60% of NA expenditures, also matched SILC+HBS is underestimating expenditures compare to NA by 35% on average (representing 65% of NA on average). Although most deviation from NA represent underestimation, some categories are over-reported such as CP041 CP043 CP082 CP102 and CP127.

Ireland

Overall, the HBS data appears to be underreporting consumption in NA (weighted average of the shares of aggregate expenses of HBS/NA is 82.2%). Further, the matching process results in enhanced coverage of NA consumption (92.7%). Conversely, the matched SILC/HBS is over reporting consumption concerning the HBS original figures (111.3%).

There are several underreported categories in HBS/NA. Notably, groups such as CP02, CP05, CP06 and CP11 are all below the 30% threshold. This trend is evident within specific three-digit categories such as CP021, CP022, CP054, CP056, CP062, CP063, CP111 and CP112 where reported expenses in HBS in relation to NA are all below the threshold.

Overall, the matched SILC/HBS marginally inflates all categories (except CP09) in relation to the original HBS figures. However, only CP04, CP08 and CP10 are above the acceptable threshold. Expanding the analysis to specific 3-digit categories, we verify that CP041, CP043, CP045, CP081 and CP104 are all overestimated by the matching procedure. The subcategories CP081 and CP104 are particularly overestimated with 225.9% and 156.8% respectively.

From this analysis, we can conclude that the matching procedure increases expense figures for almost all categories when compared to HBS. This outcome offsets differences in underreported categories initially observed between HBS and NA. However, it worsens comparability for categories initially overestimated in HBS.

Lithuania

The HBS data seems to severely underestimate consumption in National Accounts, with matching improving coverage by almost fourteen percentage points. Matching enhances the overall coverage of NA consumption, but also introduces overestimation in comparison to HBS across various subcategories.

There are no instances of underestimation when examining two-digit COICOP codes, but there are cases of underestimation at the three-digit subcategory level, particularly for CP063, CP092 and CP124.

The matching procedure inflates expenses across all COICOP categories compared to the original HBS, but given the overall underreporting of expenses in HBS, matched SILC generally remains below NA in all categories except CP10.

There is severe overestimation in matched SILC/HBS (exceeding 30%) for CP03, CP05, CP06, CP07, CP10. This seems to be driven by overestimation in the following subcategories: CP031, CP032, CP051, CP053, CP054, CP055, CP062, CP071, CP102 and CP104.

Luxembourg

Overall, HBS seems to be underreporting the consumption in NA (weighted average of the shares of aggregate expenses of HBS/NA is 69.8%). Further, the matching process results in enhanced coverage of NA consumption (83.3%). Conversely, the matched SILC/HBS is over reporting consumption concerning the HBS original figures (120.3%).

For HBS/NA, nearly all categories (except CP08) are underestimated. Notably, groups such as CP02, CP06, CP10 and CP12 are all below the 30% threshold. This trend is evident within specific three-digit categories such as CP021, CP022, CP062, CP063, CP102, CP103, CP105, CP123, CP124, CP126 and CP127 where reported expenses in HBS in relation to NA are all below the threshold.

Conversely, for matched SILC/HBS, all categories (except CP10) are inflated with relation to original HBS figures. In particular, CP01, CP02, CP03, CP04, CP05, CP06, CP07, CP11 and CP12 are all above the 15% threshold. Expanding the analysis to specific three-digit categories, we verify that CP011, CP012, CP021, CP022, CP031, CP032, CP041, CP043, CP045, CP051, CP052, CP053, CP054, CP055, CP056, CP061, CP062, CP063, CP071, CP072, CP111, CP112, CP121, CP123, CP124, CP125 and CP127 are all overestimated by the matching procedure.

From this analysis, we can conclude that the matching procedure consistently (and almost uniformly) increases expense figures when compared to HBS. However, since HBS was severely underreported with relation to NA, this offsets the initial differences between HBS and NA, bringing the matched SILC/HBS figures closer to NA for nearly all categories at the 2 digit-level.

Latvia

Overall, on average the matching is close to HBS in terms of aggregate expenses (106%). There is a general underestimation of total expenses by HBS over NA (73%). The underestimation is deemed as severe (> 30%) for all first level COICOP groups except CP01, CP04 and CP08.

On the 3-digit COICOP level, the underestimation in HBS over NA is very strong. We observe large deviations (larger than 30%) in subcategories of CP012, CP021, CP022, CP031, CP032, CP044, CP051, CP053, CP054, CP056, CP062, CP071, CP072, CP073, CP081, CP082, CP091, CP092, CP093, CP094, CP095, CP096, CP101, CP102, CP104, CP105, CP111, CP112, CP121, CP122, CP123, CP124, CP125, CP126 and CP127.

On the 3-digit COICOP level, matching performs well (deviations below 15%) in all categories, except CP031, CP032, CP041, CP052, CP053, CP054, CP056, CP062, CP071, CP081, CP082, CP091, CP093, CP094, CP096, CP101, CP105, CP111, CP112, CP121, CP123, CP125, CP126, CP127.

Malta

Overall, HBS seems to underreport consumption compared with National Accounts for all COICOP categories except CP04, in which there is over reporting.

However, overall, underreporting becomes less pronounced with the statistically matched dataset of the EU-SILC with HBS, compared to National Accounts. Three COICOP categories have underreporting deviations larger than 30% (CP02, CP09 and CP11). On the contrary, CP04 over reports by almost 40%. From the remaining eight categories, four closely match National Accounts (CP01, CP03, CP06 and CP07).

The statistically matched dataset of the EU-SILC with HBS is similar to HBS data in this regard, with all COICOP level 1 categories over reported between 10 to 40%, except for CP02 for which there is a strong underreporting.

Netherlands

The HBS data underreports expenditures compared to National Accounts (NA) on average by11%. Exceptions are spending on CP04 (Housing, water, electricity, gas and other fuels) which are over reported by 30% and CP10 (Education), over-reported by 67%. Matching enhances coverage of NA expenditures (average underreporting estimated to 2% compared to NA). Spending on CP04 and CP10 are also a bit higher in the matched SILC compared to original HBS, which means they overestimate NA significantly. The matched SILC+HBS figures tend to overestimate consumption relative to the original HBS figures by about 11%.

The matching performs well also at 3-digit COICOP level: it replicates closely HBS coverage with an average overestimation of 13%. When comparing to the NA, expenditures in matched SILC slightly overestimate aggregate consumption on average by 8%. Although most deviation from NA represent underestimation, some categories are severely overestimated following over-reporting in HBS. Such groups are CP043 (55%) CP044 (almost triple compared to NA, but only 6% higher in matched SILC than HBS), CP104 and CP105 (similar to HBS, but much higher than NA).

Portugal

The original HBS data seems to be generally underestimating consumption figures over NA, with a weighted average of aggregate expenses shares equal to 64.4%. The only first level COICOP groups that are within the 30% range are CP04, CP06, CP07, CP08, and CP10.

The matching process generally worsen the coverage of NA consumption (61.5%) respect to the performance of original HBS, specifically for the COICOP groups CP03, CP04, CP05, CP07, CP09, CP10, CP11 and CP12.

Moreover, the matched SILC/HBS rates are in the acceptable range of 80.13%-106.39% for all COICOP level 1 categories. On the 3-digit COICOP level, the matching performs good, where we do not have deviations larger than 30%, except for CP081 (193.64%), and CP103 (64.93%).

Romania

The matching data do generally well and do not produce any big distortion compared to original HBS (most of them are <30%). There are few exceptions to this which includes 3-digit level of 063, 071, 081, 082, 091, 092, 096, 101, 102 and 126 where the matching produces a more severe distortion. Many expenditures categories (almost all except COICOP 04) are fully under-reported in HBS and as such results under-reported also in the matched dataset but that does not depend on the matching procedure. Other problematic categories to be signalled regard transport expenses (CP07) where expenditures for CP081 are highly over-reported in HBS while results under-reported when looking at the matched dataset, this is the result of both imprecise starting data as well as distortions happening because of the matching procedure. COICOP 10 also seem to be quite distorted by the matching procedure especially for CP101 and CP105.

Sweden

Overall the matching performs well, where the average SILC / HBS share is 107 percent. COICOP categories that perform less good (>120% or <80%) are the following COICOP categories:

- CP02, where SILC underreports compared to HBS (77%).
- CP04, where SILC overestimates compared to HBS, which is mostly due to CP041 and CP045.
- CP06, where SILC severely overestimates compared to HBS (154%), due to CP062.
- CP08, where SILC overestimates compared to HBS (130%), consistently over its' subcategories.
- CP09, where SILC overestimates compared to HBS (123%), mainly driven by CP092 and CP093.
- CP10, where SILC overestimates compared to HBS (131%), mainly driven by CP101 and CP105.
- CP11, where SILC overestimates compared to HBS (135%), consistently for the underlying subcategories.

Compared to NA, HBS underestimates significantly. On average 67% of the expenditures are captured by HBS. Because of the relative overestimation in SILC compared to HBS (70 %).

Slovenia

Overall, HBS seems to be underreporting consumption in NA (weighted average of the shares of aggregate expenses of HBS/NA is 67.9%). The matching process generally results in slightly enhanced coverage of NA consumption, at 71.3%, compared to the performance of the original HBS across some COICOP groups and their two-digit level subcategories.

Notably, in groups CP06 and CP10, the matching inflates the figures slightly more than 15% from the original HBS figures. This trend is evident within specific two-digit subcategories such as CP063 and CP101, where the reported expenses after matching exceed those originally reported by HBS by varying degrees.

Overall, the matching procedure for Slovenia yields results that are relatively close to those of the HBS.

Slovakia

Overall, HBS data appears to underreport consumption in National Accounts (weighted average of the shares of aggregate expenses of HBS/NA is 54.5%). The matching process results in a slight decrease in the coverage of NA consumption (52%). The matched SILC/HBS figures show modest decline of expenses compared to original HBS data (94%).

Several categories in HBS/NA are underreported most severe in CP10 (23.3%) led by CP105 (6.7%), CP02 (26.6%), CP09 (33.7%) led by CP095 (14.6%), while moderate for many other categories show underreporting in the 30-50% range compared to NA figures.

Overall, SILC/HBS matching produces very close expenses, on average they are slightly below HBS, hence the resulting SILC/NA patterns are very similar to described above. The matching process (SILC/HBS) leads to increased expense figures only in CP06 (110.1%). On level 2, results are more nuanced.

CP071 (Operation of personal transport equipment), CP125 (Personal care products) present the most significant overestimation (over 130%) post-matching (SILC/NA). Some subcategories within CP04, CP05, and CP06 show a slight increase in reported expenses after the matching process.

The matching procedure follows HBS closely and does not consistently improve the poor HBS coverage of NA consumption. Matching introduces overestimation in rare subcategories.

A.3.2 SILC year – 2021

Austria

The HBS data seems to underestimate consumption in National Accounts, but matching improves coverage by almost nineteen percentage points. Matching enhances the overall coverage of NA consumption, but also introduces overestimation in comparison to HBS across various subcategories.

There are no instances of underestimation when examining two-digit COICOP codes, but there are cases of underestimation at the three-digit subcategory level, particularly for CP104 and CP121.

The matching procedure inflates expenses across almost all COICOP categories compared to the original HBS, but given the overall underreporting of expenses in HBS, matched SILC generally remains below NA and approaches its values in several cases.

There is severe overestimation in matched SILC/HBS (exceeding 30%) for CP03, CP04, CP05, CP08. This seems to be driven by overestimation in the following subcategories: CP031, CP043, CP045, CP051, CP054, CP081 and CP083.

Belgium

The matching performs relatively well in terms for expenditures at the one-digit aggregate COICOP categories. Expenditures imputed from HBS to SILC do not depart more than 20% from the original source survey (HBS) for all COICOP one-digit aggregate categories except for CP06 (health) and CP12 (others), where the imputed expenditures are about 122% of those reported in HBS, followed by CP07 (114.5%).

When we compare these matched expenditures in SILC (mSILC) with expenditures from National Accounts (NA), gaps are larger. The mSILC/NA coverage ranges from 45.7% in CP02 -alcohol and tobacco– to 119/121% in CP07/08 -i.e. transport and communication.

The low coverage of category CP02 (alcoholic beverages and tobacco) comes entirely from the under-reporting in HBS (the ratio HBS/NA is 43%, even slightly lower than mSILC/NA). This is not a Belgium-specific problem, as these types of products tend to be under-reported in household surveys. In fact, the HBS/NA coverage in Belgium is quite above the EU average (35.6%). When we look at the specific products within this category, we see that the consumption of tobacco (CP02.2) is the main driver of this under-reported consumption in HBS with respect to NA (27.3%), whereas the coverage for alcoholic beverages (CP02.1) is still below 100% but much higher, of 72.2%.

The overestimation of expenditures in category 08 (communication) is also mainly driven by the original over-reporting (an HBS/NA ratio of 114%), which is exacerbated after the matching. The distortion introduced by the matching is mainly observed in category CP 08.1 (postal services): expenditures in the matched SILC double those reported in the original HBS file. The rest (CP08.2 and CP08.3 - i.e., telephone equipment and services) are both over-reported in HBS w.r.t NA and get slightly reinforced with the match.

In contrast, the over-estimation of expenditures in CP07 with respect to NA is mainly driven by the matching, as the original HBS/NA ratio is 104%. Looking at the components of CP07, we can clearly see that this over-reporting of HBS w.r.t NA comes from category CP07.1 (purchase of vehicles).

Bulgaria

Overall, on average, the matching deviates significantly from HBS in terms of aggregate expenses (%). The deviation is about 26%. There is a general and severe underestimation of total expenses by HBS over NA (%). The underestimation is deemed severe (> 30%) for all first-level COICOP groups except CP01, CP04, CP05, and CP06.

On the 3-digit COICOP level, the underestimation in HBS over NA remains strong. We observe large deviations (greater than 30%) in subcategories of CP012, CP021, CP022, CP031, CP041, CP051, CP052, CP053, CP054, CP055, CP061, CP062, CP071, CP072, CP073, CP081, CP082, CP091, CP092, CP093, CP094, CP095, CP096, CP103, CP104, CP105, CP111, CP112, CP123, CP125, CP126, CP127.

On the 3-digit COICOP level, the matching performs well (deviations below 15%) except in categories CP041, CP054, CP055, CP063, CP072, CP081, CP082, CP091, CP092, CP093, CP096, CP102, CP104, CP111, CP112, CP124, CP126, CP127.

Cyprus

Overall, it seems that HBS underreports consumption in NA (weighted average of the shares of aggregate expenses of HBS/NA is 76.5%). The matching process generally results in improved coverage of NA consumption (84.7%) with respect to the performance of original HBS across some COICOP groups and their two-digit level subcategories.

Notably, in groups CP06 and CP12, the matching inflates the original HBS figures. This trend is particularly evident in specific three-digit subcategories such as CP061, CP123, CP126 and CP127, where the reported expenses after matching exceed those originally reported by HBS by varying extents.

Similarly, in CP10, subcategory CP102 also shows a decrease in reported expenses post-matching. This case of underestimation highlights that the matching process does not uniformly increase expense figures and can differentially impact detailed product groups and subcategories.

Overall, the matching procedure yields results that are relatively close to those of the HBS.

Czechia

The HBS data 2021 appears to significantly underreport consumption in National Accounts (weighted average of the shares of aggregate expenses of HBS/NA is 57.4%). The matching process moderately improves the coverage of NA consumption (64.3%). The matched SILC/HBS figures show an overall increase in expenses compared to the original HBS data (111.5%).

HBS/NA: Several categories exhibit severe underreporting: CP124 (11.3%), CP104 (28.8%), CP112 (24.6%), CP126 (14.7%), and CP02 (18.7%). Many other categories fall in the moderate underestimation range (30-50% below NA).

SILC/NA: Matching lessens underestimation in some categories. However, several remain severely underreported, including CP103 (1.8% - down significantly from 43.17% in HBS/NA), CP104 (25.9%), CP126 (16.3%), and CP091 (30.5%). On level 1, CP02 (19.8%) and CP11 (39.4%) remain severely underestimated.

SILC/HBS: The matching process inflates expenses across many categories. Notable overestimations are present in CP127 (255.8%), CP081 (233.3%), CP102 (128.9%), and CP041 (122.9%).

SILC/NA: Despite some increase compared to HBS, the matched SILC figures only exceed NA for a limited number of subcategories, including CP102 (162.2%), CP041 (134.2%) and CP127 (136.4%).

The matching procedure for Czechia 2021 improves the coverage of NA consumption but falls short of fully addressing the underreporting in HBS. While it introduces overestimation in many categories, significant underestimation persists in others.

Denmark

The matching data do generally well and do not produce any big distortion compared to original HBS (most of them are <30%). There are few exceptions to this which includes 3-digit level of 063, 071, 081 and 126 where the matching produces a more severe distortion. COICOP 02 expenditures (alcohol and tobacco) are fully under-reported in HBS and as such results under-reported also in the matched dataset but that does not depend on the matching procedure. Other problematic categories to be signalled regard transport expenses (CP07) where expenditures for CP071 are highly over-reported in HBS while results under-reported when looking at the matched dataset, this is the result of both imprecise starting data as well as distortions happening because of the matching procedure.

General issues related to HBS original data regards: over-reporting for CP125, 043; and under-reporting (more frequent) for 022, 063, 091, 092, 094, 105, 124, 126

Germany

The matching performs relatively well in terms for expenditures at the one-digit aggregate COICOP categories. Expenditures imputed from HBS to SILC do not depart more than 20% from the original source survey (HBS) for all COICOP one-digit aggregate categories except for CP12 ("Other", where matched expenditures are 138% of original HBS, driven by CP12.5, insurance services) and CP10 ("Education", with an extremely low ratio of 24%).

When we compare these matched expenditures in SILC (mSILC) with expenditures from National Accounts (NA), gaps are larger. The categories with the smallest coverage are CP10 ("Education", 18%), CP02 ("Alcoholic beverages and tobacco", 41%) and CP05 ("Furnishing", 57%).

The very small share of expenditures in mSILC with respect to NA for the CP10 category is mainly driven by the distortion of the matching, as the original HBS/NA ratio is much larger (74.5%).

In contrast, the low coverage in CP02 and CP05 is driven by the original low HBS/NA rates. In the case of CP02 ("Alcoholic beverages and tobacco") the HBS/NA ratio is 38%. Although this ratio looks quite low, it is even slightly above the EU average. In the case of CP05 ("Furnishings"), the HBS/NA ratio is 58%.

Estonia

The matching data do generally well and do not produce any big distortion compared to original HBS. At the level one, distortions are generally of the order of 10%, whereas at the level 2 they are typically below 30%. A few exceptions to this include 081 and 127 where the matching produces severe distortions (beyond 50%). When compared to national account, COICOP 02 (alcohol and tobacco), COICOP 03 (clothing and footwear), COICOP 11 (restaurant and accommodation services) and COICOP 12 (miscellaneous good and service) are significantly under-reported in HBS and as such results under-reported also in the matched dataset but that does not depend on the matching procedure.

In the following case the comparison with NA is especially poor: 126 and 124 are dramatically underestimated, while 101 is largely overestimated.

Greece

Overall, HBS seems to be underreporting the consumption in NA (in all categories apart from CP06). Underreporting becomes even more pronounced when we compare the matched SILC with NA, with seven categories exceeding the acceptable range of 30% (CP02, CP03, CP07, CP08, CP09, CP11, CP12).

The matched SILC/HBS rates are all in the range of 90%-101% for all categories apart from CP06.

CP06 seems to be the most worrisome category: the overall consumption reported in HBS is very close to the one in NA, but the level 2 analysis shows that in item CP063 HBS there is a significant overestimation (HBS/NA ≈ 300%). The matching process creates a significant overestimation of this category's overall expenditure (matched SILC/HBS ratio is at 155.3%).

Spain

Overall, HBS seems to underreport consumption compared with National Accounts data for all COICOP categories.

However, overall, underreporting becomes less pronounced with the statistically matched dataset of the EU-SILC with HBS, compared to National Accounts. Four COICOP categories have deviations larger than 30% (CP02, CP09, CP11 and CP12). From the remaining eight categories, three closely match National Accounts (CP03, CP04 and CP08).

The statistically matched dataset of the EU-SILC with HBS is similar to HBS data in this regard, with all COICOP level 1 categories over reported between 5 to 25%, except for CP02 for which there is a strong underreporting.

Finland

In general, the HBS appear to be overstating consumption figures in the NA, with a weighted average of aggregate expenses shares reaching 119.6%. The matching process generally results in an additional overrepresentation of NA consumption coverage compared to the original HBS data across various three-digit subcategories, exacerbating the issue of consumption over reporting.

Particularly noteworthy is the substantial overestimation of consumption in specific three-digit subcategories such as CP071, CP081 and CP127, where reported expenses after matching significantly exceed those originally reported by HBS.

Despite the overall overestimation in coverage, there are instances of underestimated coverage. In COICOP groups CP01, CP03, CP09, and CP10, all subcategories report lower expenses after matching compared to HBS. Furthermore, in other COICOP groups, three-digit subcategories exhibit a decrease in reported expenses post-matching. Notably, in CP022, CP043, CP123, and CP124, the matching adjustment leads to a significantly lower expenditure than in HBS. These cases of underestimation underscore the non-uniform impact of the matching process on expense figures across detailed groups and subcategories.

France

The original HBS data seems to be generally underestimating consumption figures over NA for France, with a weighted average of aggregate expenses shares equal to 74%. However, the only first level COICOP groups deemed as severe underreported (>30%) are: CP02, CP06 and CP11.

The matching process generally results in an additional underestimation of NA consumption coverage compared to the original HBS data (69.9%). The severe underreporting persist for the mentioned three COICOP categories (particularly worsening for CP02 but remaining similar for CP06 and CP11), as well as CP09, when comparing the HBS matched to SILC dataset to the NA.

Moreover, the matched SILC/HBS rates are in the acceptable range of 76.58%-101.15% for all COICOP level 1 categories. On the 3-digit COICOP level, the matching performs good, where we do not have deviations larger than 30% except for CP092 (152.89%) and CP104 (68.79%)

Croatia

Overall, HBS seems to be underreporting the consumption in NA in all COICOP categories.

Underreporting becomes even more pronounced when we compare the matched SILC with NA, with only five categories being within the acceptable range of 30% (CP01, CP03, CP04, CP07 and CP08). From the remaining nine categories, the most underreported is CP11, followed by CP02, CP09 and CP06.

The matched SILC/HBS rates are in the (acceptable) range of 95%-112% for all COICOP level 1 categories.

Hungary

Overall, on average the matched SILC is extremely close to HBS in terms of aggregate expenses (100% on average). There is a general underestimation of total expenses by HBS over NA (59%). The underreporting from HBS is greater than 30% (severe) for most first level COICOP groups except CP04 (Housing, water, electricity, gas and other fuels) that is over-reported by about 46% and CP08 (communication) that results accurate compared to the national account figures.

The matching process reproduces the same patterns of consumption as for HBS in SILC and aggregates expenditures coverage at COICOP level 1 is very similar (matched SILC on average captures 60% of NA).

The matching performs well also at 3-digit COICOP level: it replicates closely HBS coverage with an average underestimation of 1%. However, users should be aware that because of the underreporting of many expenditures in HBS representing only 65% of NA expenditures. Consequently also matched SILC+HBS underestimates expenditures compare to NA by 34% on average (representing 66% of NA on average). Although most deviation from NA represent underestimation, some categories are over-reported such as CP041 CP043 CP044 CP045 CP096 CP101 CP125 and CP127.

Ireland

Overall, the HBS data appears to be underreporting consumption in NA (weighted average of the shares of aggregate expenses of HBS/NA is 82.2%). Further, the matching process results in enhanced coverage of NA consumption (88.5%). Conversely, the matched SILC/HBS is over reporting consumption concerning the HBS original figures (108.3%).

There are several underreported categories in HBS/NA. Notably, groups such as CP02, CP05, CP06 and CP11 are all below the 30% threshold. This trend is evident within specific three-digit categories such as CP021, CP022, CP054, CP056, CP062, CP063, CP103, CP111, CP112, CP121 and CP126, where reported expenses in relation to NA are below the threshold.

Overall, the matched SILC/HBS marginally inflates all categories (except CP10) in relation to the original HBS figures. However, only CP06 and CP08 are above the acceptable threshold. Expanding the analysis to specific 3-digit categories, we verify that CP043, CP045, CP061, CP072, CP081, CP123 and CP127 are all overestimated by the matching procedure.

From this analysis, we can conclude that the matching procedure increases expense figures for almost all categories when compared to HBS. This outcome offsets differences in underreported categories initially observed between HBS and NA. However, it worsens comparability for categories initially overestimated in HBS.

Lithuania

The HBS data seems to severely underestimate consumption in National Accounts, with matching improving coverage by fourteen percentage points. Matching enhances the overall coverage of NA consumption, but also introduces overestimation in comparison to HBS across various subcategories.

There are no instances of underestimation when examining two-digit COICOP codes, but there are cases of underestimation at the three-digit subcategory level, particularly for CP092, CP105 and CP124.

The matching procedure inflates expenses across all COICOP categories compared to the original HBS, but given the overall underreporting of expenses in HBS, matched SILC generally remains below NA in all categories except CP10.

There is severe overestimation in matched SILC/HBS (exceeding 30%) for CP03, CP05, CP06, CP07, CP10. This seems to be driven by overestimation in the following subcategories: CP031, CP032, CP051, CP052, CP053, CP054, CP062, CP071, CP102 and CP104.

Luxembourg

Overall, HBS seems to be underreporting the consumption in NA (weighted average of the shares of aggregate expenses of HBS/NA is 69.8%). Further, the matching process results in enhanced coverage of NA consumption (78.6%). Conversely, the matched SILC/HBS is over reporting consumption concerning the HBS original figures (113.1%).

For HBS/NA, nearly all categories (except CP08) are underestimated. Notably, groups such as CP02, CP06, CP10 and CP12 are all below the 30% threshold. This trend is evident within specific three-digit categories such as CP021, CP022, CP062, CP063, CP102, CP103, CP105, CP123, CP124, CP126 and CP127 where reported expenses in HBS in relation to NA are all below the threshold.

Conversely, for matched SILC/HBS, all categories (except CP10) are inflated with relation to original HBS figures. In particular, CP03, CP05, CP06, CP09 and CP012 are all above the 15% threshold. Expanding the analysis to specific three-digit categories, we verify that CP022, CP031, CP041, CP051, CP052, CP054, CP055, CP056, CP063, CP081, CP092, CP093, CP096, CP103, CP124 and CP127 are all overestimated by the matching procedure.

From this analysis, we can conclude that the matching procedure consistently (and almost uniformly) increases expense figures when compared to HBS. However, since HBS was severely underreported with relation to NA, this offsets the initial differences between HBS and NA, bringing the matched SILC/HBS figures closer to NA for nearly all categories at the 2 digit-level.

Latvia

Overall, on average the matching is close to HBS in terms of aggregate expenses (115%). There is a general underestimation of total expenses by HBS over NA (67%). The underestimation is deemed as severe (>30%) for all first level COICOP groups except CP01, CP04 and CP08.

On the 3-digit COICOP level, the underestimation in HBS over NA is very strong. We observe large deviations (larger than 30%) in subcategories of CP012, CP021, CP022, CP031, CP032, CP044, CP051, CP053, CP054, CP056, CP062, CP071, CP072, CP073, CP081, CP082, CP091, CP092, CP093, CP094, CP095, CP096, CP101, CP102, CP104, CP105, CP111, CP112, CP121, CP122, CP123, CP124, CP125, CP126, CP127.

On the 3-digit COICOP level, matching performs well (deviations below 15%) in all categories, except CP031, CP032, CP041, CP052, CP053, CP054, CP056, CP062, CP071, CP081, CP082, CP091, CP093, CP094, CP096, CP101, CP105, CP111, CP112, CP121, CP123, CP125, CP126, CP127.

Malta

Overall, HBS seems to underreport consumption compared with National Accounts data for all COICOP categories except CP04, in which there is over reporting.

However, overall, underreporting becomes less pronounced with the statistically matched dataset of the EU-SILC with HBS, compared to National Accounts. Three COICOP categories have underreporting deviations larger than 30% (CP02, CP09 and CP11). On the contrary, CP04 over reports by almost 30%. From the remaining eight categories, three closely match National Accounts (CP01, CP05 and CP07).

The statistically matched dataset of the EU-SILC with HBS is similar to HBS data in this regard, with all COICOP level 1 categories over reported between 10 to almost 50%, except for CP02 for which there is a strong underreporting.

Netherlands

The HBS data underreports expenditures compared to National Accounts (NA) on average by 11%. Exceptions are spending on CP04 (Housing, water, electricity, gas and other fuels) which are over reported by 30% and CP10 (Education), over-reported by 67%. Matching enhances coverage of NA expenditures (average underreporting estimated to 1% compared to NA). Spending on CP04 and CP10 are also a bit higher in the matched SILC compared to original HBS, which means they overestimate NA significantly. The matched SILC+HBS figures tend to overestimate consumption relative to the original HBS figures by about 11%.

The matching performs well also at 3-digit COICOP level: it replicates closely HBS coverage with an average overestimation of 10%. When comparing to the NA, expenditures in matched SILC slightly overestimate aggregate consumption on average by 9%. Although most deviation from NA represent underestimation,

some categories are severely overestimated following over-reporting in HBS. Such groups are CP043 (45%) CP044 (almost triple compared to NA, but only 9% higher in matched SILC than HBS), CP055, CP083, CP104 and CP105 (similar to HBS, but much higher than NA, respectively 4 and 6 times higher).

Poland

Overall the matching performs very well, where the average SILC / HBS share is 105 percent. COICOP categories that perform less good are the following COICOP categories:

None of the COICOP categories of the first level have a deviation that >120% or <80%.

Compared to NA, HBS underestimates significantly. On average 50.7 percent of the expenditures are captured by HBS. Similarly SILC/NA captures this underestimation with an average share of 52.7%.

Portugal

The original HBS data seems to be generally underestimating consumption figures over NA, with a weighted average of aggregate expenses shares equal to 72.7%. The only groups that are within the 30% range are CP04, CP06, CP07, CP08, CP10.

The matching process generally results in an additional underestimation of NA consumption coverage compared to the original HBS data (69.4%), except for the first level COICOP groups CP02, CP06 and CP08.

Moreover, the matched SILC/HBS rates are in the acceptable range of 83.35%-109.16% for all COICOP level 1 categories. On the 3-digit COICOP level, the matching performs well, where we do not have deviations larger than 30%, except for CP081 (198.29%), CP102 (62.69%), CP112 (67.57%), and CP127 (160.05%).

Romania

The matching data do generally well and do not produce any big distortion compared to original HBS (most of them are <30%). There are few exceptions to this which includes 3-digit level of 063, 071, 081, 082, 091, 092, 096, 101, 102 and 126 where the matching produces a more severe distortion. Many expenditures categories (almost all except COICOP 04) are fully under-reported in HBS and as such results under-reported also in the matched dataset but that does not depend on the matching procedure. Other problematic categories to be signalled regard transport expenses (CP07) where expenditures for CP081 are highly over-reported in HBS while results under-reported when looking at the matched dataset, this is the result of both imprecise starting data as well as distortions happening because of the matching procedure. COICOP 10 also seem to be quite distorted by the matching procedure especially for CP101 and CP105.

Sweden

Overall the matching performs well, where the average SILC / HBS share is 101%. COICOP categories that perform less good (>120% or <80%) are the following COICOP categories:

- CP02, where SILC underreports compared to HBS (77%).
- CP06, where SILC overestimates compared to HBS (142%), this is due to overestimation for CP062.
- CP08, where SILC overestimates compared to HBS (122%), which is due to overestimation for CP081 and CP083.

CP10, where SILC underestimates compared to HBS (77.89%), the overestimation seems to be mainly driven by CP102 and CP107.

Compared to NA, HBS underestimates significantly. On average 64.5% of the expenditures are captured by HBS. Because of the slight relative overestimation in SILC compared to HBS, this increases to 65.9 % for SILC vs NA.

Slovenia

Overall, HBS seems to be underreporting consumption in NA (weighted average of the shares of aggregate expenses of HBS/NA is 67.9%). The matching process generally results in slightly enhanced coverage of NA consumption, at 71.8%, compared to the performance of the original HBS across some COICOP groups and their two-digit level subcategories.

Notably, in group CP10, the matching inflates the figures slightly more than 15% from the original HBS figures. This trend is evident within specific two-digit subcategories such as CP101, CP104, and CP105, where the reported expenses after matching exceed those originally reported by HBS by varying degrees.

Overall, the matching procedure for Slovenia yields results that are relatively close to those of the HBS.

A.3.3 SILC year – 2020

Austria

The HBS data seems to underestimate consumption in National Accounts, but matching improves coverage by eighteen percentage points. Matching enhances the overall coverage of NA consumption, but also introduces overestimation in comparison to HBS across various subcategories.

There are no instances of underestimation when examining two-digit COICOP codes, but there is one case of underestimation at the three-digit subcategory level, for CP121.

The matching procedure inflates expenses across almost all COICOP categories compared to the original HBS, but given the overall underreporting of expenses in HBS, matched SILC remains below NA in most categories, with the most notable exception being CP10.

There is severe overestimation in matched SILC/HBS (exceeding 30%) for CP03 and CP04. This seems to be driven by overestimation in the following subcategories: CP031 and CP043.

Belgium

The matching performs relatively well in terms for expenditures at the one-digit aggregate COICOP categories. Expenditures imputed from HBS to SILC do not depart more than 20% from the original source survey (HBS) for all COICOP one-digit aggregate categories except from CP06 (health) and CP12 (others), where the imputed expenditures are about 122% of those reported in HBS, followed by CP07 (114.5%).

When we compare these matched expenditures in SILC (mSILC) with expenditures from National Accounts (NA), gaps are larger. The mSILC/NA coverage ranges from 45.7% in CP02 -alcohol and tobacco– to 119/121% in CP07/08 -i.e. transport and communication.

- The low coverage of category CP02 (alcoholic beverages and tobacco) comes entirely from the under-reporting in HBS (the ratio HBS/NA is 43%, even slightly lower than mSILC/NA). This is not a Belgium-specific problem, as these types of products tend to be under-reported in household surveys. In fact, the HBS/NA coverage in Belgium is quite above the EU average (35.6%). When we look at the specific products within this category, we see that the consumption of tobacco (CP02.2) is the main driver of this under-reported consumption in HBS with respect to NA (27.3%), whereas the coverage for alcoholic beverages (CP02.1) is still below 100% but much higher, of 72.2%.
- The overestimation of expenditures in category 08 (communication) is also mainly driven by the original over-reporting (an HBS/NA ratio of 114%), which is exacerbated after the matching. The distortion introduced by the matching is mainly observed in category CP 08.1 (postal services): expenditures in the matched SILC double those reported in the original HBS file. The rest (CP08.2 and CP08.3 - i.e., telephone equipment and services) are both over-reported in HBS w.r.t NA and get slightly reinforced with the match.
- In contrast, the over-estimation of expenditures in CP07 with respect to NA is mainly driven by the matching, as the original HBS/NA ratio is 104%. Looking at the components of CP07, we can clearly see that this over-reporting of HBS w.r.t NA comes from category CP07.1 (purchase of vehicles).

Bulgaria

Overall, on average the matching is far away from HBS in terms of aggregate expenses (%). The deviation is about 24%. There is a general and severe underestimation of total expenses by HBS over NA (%). The underestimation is deemed as severe (> 30%) for all first level COICOP groups except CP01 and CP04. On the 3-digit COICOP level, the underestimation in HBS over NA is very strong.

We observe large deviations (larger than 30%) in subcategories of CP012, CP021, CP022, CP031, CP041, CP051, CP052, CP053, CP054, CP055, CP061, CP062, CP063, CP071, CP072, CP073, CP081, CP082, CP083,

CP091, CP092, CP093, CP094, CP095, CP102, CP103, CP104, CP105, CP111, CP112, CP123, CP125, CP126 and CP127.

On the 3-digit COICOP level, matching performs well (deviations below 15%) except in categories CP041, CP043, CP051, CP054, CP055, CP063, CP072, CP081, CP082, CP091, CP092, CP093, CP096, CP102, CP104, CP112, CP124 and CP127.

Cyprus

Overall, it seems that HBS underreports consumption in NA (weighted average of the shares of aggregate expenses of HBS/NA is 76.5%). The matching process generally results in improved coverage of NA consumption (84.6%) with respect to the performance of original HBS across some COICOP groups and their two-digit level subcategories.

Notably, in groups CP06 and CP12, the matching inflates the original HBS figures. This trend is particularly evident in specific three-digit subcategories such as CP063, CP126 and CP127, where the reported expenses after matching exceed those originally reported by HBS by varying extents.

Similarly, in CP12, subcategory CP124 also shows a decrease in reported expenses post-matching. This case of underestimation highlights that the matching process does not uniformly increase expense figures and can differentially impact detailed product groups and subcategories.

Overall, the matching procedure yields results that are relatively close to those of the HBS.

Czechia

The HBS data appears to underreport consumption in National Accounts (weighted average of the shares of aggregate expenses of HBS/NA is 57.4%). The matching process slightly improves the coverage of NA consumption (64.6%). The matched SILC/HBS figures show an overall increase in expenses compared to the original HBS data (113%).

HBS/NA: Several categories exhibit underreporting, some severely: CP124 (11.3%), CP104 (28.8%), CP112 (24.6%), CP126 (14.7%), and CP02 (18.7%). A number of additional categories fall in the moderate underestimation range (30-50% below NA).

SILC/NA: The matching process lessens the underestimation for many categories. However, several remain significantly underestimated, most notably CP103 (10.1%), CP104 (50.5%), CP091 (31.4%), and CP063 (45.8%). CP02 category is underestimated at 20%.

SILC/HBS: The matching process inflates expenses across many subcategories. Notable overestimations are present in CP102 (185.8%), CP081 (217.5%) CP127 (204.3%), and CP051 (123.8%).

SILC/NA: While significant increase is observed compared to HBS, the matched SILC figures now exceed NA values for only a few subcategories: CP102 (233.8%), CP041 (112.7%), and CP127 (108.9%).

The matching procedure for Czechia improves the coverage of NA consumption compared to the original HBS but does not fully eliminate underreporting. While overestimation is introduced in some areas, all level 1 categories remain below NA levels in the matched SILC dataset.

Denmark

The matching data do generally well and do not produce any big distortion compared to original HBS (most of them are <30%). There are few exceptions to this which includes 3-digit level of CP063, CP071, CP081, CP092, CP096 and CP126 where the matching produces a more severe distortion. CP02 expenditures (alcohol and tobacco) are fully under-reported in HBS and as such results under-reported also in the matched dataset but that does not depend on the matching procedure. Other problematic categories to be signalled regard transport expenses (CP07) where expenditures for CP071 are highly over-reported in HBS while results under-reported when looking at the matched dataset, this is the result of both imprecise starting data as well as distortions happening because of the matching procedure. CP10 also seem to be quite distorted by the matching procedure especially for CP101 and CP105.

General issues related to HBS original data regards: Over-reporting for CP125, CP043; and under-reporting (more frequent) for CP022, CP063, CP091, CP092, CP094, CP105, CP124 and CP126

Germany

The matching performs relatively well in terms for expenditures at the one-digit aggregate COICOP categories. Expenditures imputed from HBS to SILC do not depart more than 20% from the original source survey (HBS) for all COICOP one-digit aggregate categories except for CP12 ("Other", where matched expenditures are 123% of original HBS, driven by CP125, insurance services) and CP10 ("Education", with an extremely low ratio of 20%).

Of lower order, under-simulation (when comparing SILC to HBS) of about 15% is also present in category CP04 (housing).

When we compare these matched expenditures in SILC (mSILC) with expenditures from National Accounts (NA), gaps are larger. The categories with the smallest coverage are CP10 ("Education", 15%), CP02 ("Alcoholic beverages and tobacco", 38%) and CP05 ("Furnishing", 50%).

The very small share of expenditures in mSILC with respect to NA for the CP10 category is mainly driven by the distortion of the matching, as the original HBS/NA ratio is much larger (74.5%).

In contrast, the low coverage in CP02 and CP05 is driven by the original low HBS/NA rates. In the case of CP02 ("Alcoholic beverages and tobacco") the HBS/NA ratio is 38%. Although this ratio looks quite low, it is even slightly above the EU average. In the case of CP05 ("Furnishings"), the HBS/NA ratio is 58%.

Estonia

The matching data do generally well and do not produce any big distortion compared to original HBS. At the level one, distortions are generally of the order of 10%, whereas at the level 2 they are typically below 30%. A few exceptions to this include CP081, CP092 and CP127 where the matching produces severe distortions (beyond 50%). When compared to national account, CP02 (alcohol and tobacco), CP03 (clothing and footwear), CP11 (restaurant and accommodation services) and CP12 (miscellaneous good and service) are significantly under-reported in HBS and as such results under-reported also in the matched dataset but that does not depend on the matching procedure.

In the following cases the comparison with NA is especially poor: CP112, CP126 and CP124 are dramatically underestimated while CP101 is largely overestimated.

Greece

Overall, HBS seems to be underreporting the consumption in NA (in all categories apart from CP06). Underreporting becomes even more pronounced when we compare the matched SILC with NA, with six categories exceeding the acceptable range of 30% (CP02, CP07, CP08, CP09, CP11, CP12).

The matched SILC/HBS rates are all in the range of 90%-101% for all categories apart from CP06, CP09 and CP10.

CP06 seems to be the most worrisome category: the overall consumption reported in HBS is very close to the one in NA, but the level 2 analysis shows that in item CP063 HBS there is a significant overestimation (HBS/NA ≈ 300%). The matching process creates a significant overestimation of this category's overall expenditure (matched SILC/HBS ratio is at 161.42%).

Spain

Overall, HBS seems to underreport consumption compared with National Accounts data for all COICOP categories.

However, overall, underreporting becomes less pronounced with the statistically matched dataset of the EU-SILC with HBS, compared to National Accounts. Four COICOP categories have deviations larger than 30% (CP02, CP09, CP11 and CP12). From the remaining eight categories, two closely match National Accounts (CP04 and CP08).

The statistically matched dataset of the EU-SILC with HBS is similar to HBS data in this regard, with all COICOP level 1 categories over reported between 5 to 25%, except for CP02 for which there is a strong underreporting.

Finland

In general, the HBS appear to be overstating consumption figures in the NA, with a weighted average of aggregate expenses shares reaching 119.6%. The matching process generally results in an additional overrepresentation of NA consumption coverage compared to the original HBS data across various three-digit subcategories, exacerbating the issue of consumption over reporting.

Particularly noteworthy is the substantial overestimation of consumption in specific three-digit subcategories such as CP071, CP081 and CP127, where reported expenses after matching significantly exceed those originally reported by HBS.

Despite the overall overestimation in coverage, there are instances of underestimated coverage. In COICOP groups CP01, CP03, CP09, and CP10, all subcategories report lower expenses after matching compared to HBS. Furthermore, in other COICOP groups, three-digit subcategories exhibit a decrease in reported expenses post-matching. Notably, in CP022, CP043, CP123, and CP124, the matching adjustment leads to a significantly lower expenditure than in HBS. These cases of underestimation underscore the non-uniform impact of the matching process on expense figures across detailed groups and subcategories.

France

The original HBS data seems to be underestimating consumption figures over NA, with a weighted average of aggregate expenses shares equal to 74%. However, the only first level COICOP groups deemed as severe underreported (>30%) are: CP02, CP06 and CP11.

The matching process results in a high additional underestimation of NA consumption coverage compared to the original HBS data (51.1%). The severe underreporting affects all the first level COICOP groups, except for CP03 and CP04, when comparing the HBS matched to SILC dataset to the NA.

Additionally, the matched SILC/HBS rates are significantly underestimated, particularly for the following COICOP groups with severe levels: CP011, CP021, CP051, CP052, CP062, CP071, CP072, CP073, CP82, CP09: CP093, CP094, CP101, CP102, CP104, CP105, CP111, CP111, CP121 and CP123.

Croatia

Overall, HBS seems to be underreporting the consumption in NA in all COICOP categories.

Underreporting becomes even more pronounced when we compare the matched SILC with NA, with only three categories being within the acceptable range of 30% (CP01, CP03 and CP04). From the remaining nine categories, the most underreported is CP11, followed by CP02, CP09 and CP06.

The matched SILC/HBS rates are in the (acceptable) range of 89%-112% for all COICOP level 1 categories apart from CP10 (83%).

Hungary

Overall, on average the matched SILC is extremely close to HBS in terms of aggregate expenses (99% on average). There is a general underestimation of total expenses by HBS over NA (59%). The underreporting from HBS is greater than 30% (severe) for most first level COICOP groups except CP04 (Housing, water, electricity, gas and other fuels) that is over-reported by about 46% and CP08 (communication) that results accurate compared to the national account figures.

The matching process reproduces the same patterns of consumption as for HBS in SILC and aggregates expenditures coverage at COICOP level 1 is very similar (matched SILC on average captures 53% of NA).

The matching performs well also at 3-digit COICOP level: it replicates closely HBS coverage with an average underestimation of 1%. However, users should be aware that because of the underreporting of many expenditures in HBS representing only 65% of NA expenditures. Consequently also matched SILC+HBS underestimates expenditures compare to NA by 45% on average (representing 99% of NA on average). Although most deviation from NA represent underestimation, some categories are over-reported such as CP041 CP043 CP044 CP045 CP096 CP101 CP125 and CP127.

Ireland

Overall, the HBS data appears to be underreporting consumption in NA (weighted average of the shares of aggregate expenses of HBS/NA is 82.2%). Further, the matching process results in enhanced coverage of NA

consumption (88.6%). Conversely, the matched SILC/HBS is over reporting consumption concerning the HBS original figures (107.9%).

There are several underreported categories in HBS/NA. Notably, groups such as CP02, CP05, CP06 and CP11 are all below the 30% threshold. This trend is evident within specific three-digit categories such as CP021, CP022, CP054, CP056, CP062, CP063, CP073, CP103, CP111, CP112, CP121, CP126 and CP127, where reported expenses in relation to NA are below the threshold.

Overall, the matched SILC/HBS marginally inflates all categories (except CP09) in relation to the original HBS figures. However, only CP08 is above the acceptable threshold. Expanding the analysis to specific 3-digit categories, we verify that CP022, CP043, CP045, CP081, CP121 and CP127 are all overestimated by the matching procedure.

From this analysis, we can conclude that the matching procedure increases expense figures for almost all categories when compared to HBS. This outcome offsets differences in underreported categories initially observed between HBS and NA. However, it worsens comparability for categories initially overestimated in HBS.

Lithuania

The HBS data seems to severely underestimate consumption in National Accounts, with matching improving coverage by sixteen percentage points. Matching enhances the overall coverage of NA consumption, but also introduces overestimation in comparison to HBS across various subcategories.

There are no instances of underestimation when examining two-digit COICOP codes, but there are cases of underestimation at the three-digit subcategory level, particularly for CP041 and CP105.

The matching procedure inflates expenses across all COICOP categories compared to the original HBS, but given the overall underreporting of expenses in HBS, matched SILC generally remains below NA in all categories except CP10.

There is severe overestimation in matched SILC/HBS (exceeding 30%) for CP03, CP07 and CP10. This seems to be driven by overestimation in the following subcategories: CP031, CP032, CP071, CP101, CP102 and CP104. The overestimation of CP102 is particularly severe (by more than one-thousand percent).

Latvia

Overall, on average the matching is close to HBS in terms of aggregate expenses (%). There is a general underestimation of total expenses by HBS over NA (%). The underestimation is deemed as severe (> 30%) for all first level COICOP groups except CP01, CP04 and CP08.

On the 3-digit COICOP level, the underestimation in HBS over NA is very strong. We observe large deviations (larger than 30%) in subcategories of CP012, CP021, CP022, CP031, CP032, CP044, CP051, CP053, CP054, CP056, CP062, CP071, CP072, CP073, CP081, CP082, CP091, CP092, CP093, CP094, CP095, CP096, CP101, CP102, CP104, CP105, CP111, CP112, CP121, CP122, CP123, CP124, CP125, CP126 and CP127.

On the 3-digit COICOP level, matching performs well (deviations below 15%) in all categories, except CP031, CP032, CP041, CP052, CP053, CP054, CP056, CP062, CP071, CP081, CP082, CP091, CP093, CP094, CP096, CP101, CP105, CP111, CP112, CP121, CP123, CP125, CP126 and CP127.

Poland

Overall the matching of Poland performs very well, where the average SILC/HBS share is 106.5%. COICOP categories that perform less good are the following COICOP categories:

None of the COICOP categories of the first level have a deviation that >120% or <80%.

Compared to NA, HBS underestimates significantly. On average 50.7% of the expenditures are captured by HBS. Similarly SILC/NA captures this underestimation with an average share of 53.4%.

Portugal

The original HBS data seems to be generally underestimating consumption figures over NA, with a weighted average of aggregate expenses shares equal to 72.7% The only first level COICOP groups that are within the 30% range are CP04, CP06, CP07, CP08 and CP10.

The matching process generally worsen the coverage of NA consumption (67.9%) respect to the performance of original HBS, specifically for the COICOP groups CP03, CP04, CP05, CP07, CP09, CP10, and CP11.

Moreover, the matched SILC/HBS rates are in the acceptable range of 77.66%-107.89% for all COICOP level 1 categories. On the 3-digit COICOP level, the matching performs well, where we do not have deviations larger than 30%, except for CP081 (194.78%), CP092 (63.8%), CP124 (134.35%), and CP127 (213.42%).

Romania

The matching data do generally well and do not produce any big distortion compared to original HBS (most of them are <30%). There are few exceptions to this which includes 3-digit level of CP063, CP071, CP081, CP082, CP091, CP092, CP096, CP101, CP102 and CP126 where the matching produces a more severe distortion. Many expenditures categories (almost all except CP04) are fully under-reported in HBS and as such results under-reported also in the matched dataset but that does not depend on the matching procedure. Other problematic categories to be signalled regard transport expenses (CP07) where expenditures for CP081 are highly over-reported in HBS while results under-reported when looking at the matched dataset, this is the result of both imprecise starting data as well as distortions happening because of the matching procedure. CP10 also seem to be quite distorted by the matching procedure especially for CP101 and CP105.

Sweden

Overall the matching of Sweden performs well, where the average SILC / HBS share is 105.6 %. COICOP categories that perform less good (>120% or <80%) are the following COICOP categories:

- CP02, where SILC underestimates compared to HBS (75.5%).
- CP04, where SILC overestimates compared to HBS (120.1%).
- CP06, where SILC overestimates compared to HBS (140.6.4%), due to CP062.
- CP08, where SILC overestimates compared to HBS (126.8%), due CP081 and CP083.
- CP10, where SILC severely underestimates compared to HBS (39.7%), mainly driven by CP102 and CP107.
- CP11, where SILC overestimates compared to HBS (127.2%).

Compared to NA, HBS underestimates significantly. On average 60.4% of the expenditures are captured by HBS. Because of the slight relative overestimation in SILC compared to HBS, this increases to 65.5% for SILC vs NA.

Slovenia

Overall, HBS seems to be underreporting consumption in NA (weighted average of the shares of aggregate expenses of HBS/NA is 67.9%). The matching process generally results in slightly enhanced coverage of NA consumption, at 70.5%, compared to the performance of the original HBS across some COICOP groups and their two-digit level subcategories.

Notably, in group CP06, the matching inflates the figures slightly more than 15% from the original HBS figures. This trend is evident within specific two-digit subcategories such as CP062, and CP063, where the reported expenses after matching exceed those originally reported by HBS by varying degrees.

Overall, the matching procedure for Slovenia yields results that are relatively close to those of the HBS.

Slovakia

The HBS data (2020) appears to underreport consumption in National Accounts (weighted average of the shares of aggregate expenses of HBS/NA is 54.5%). The matched SILC/HBS figures show a slight decrease in expenses compared to the original HBS data (96.3%). The matching process results in a decrease in the coverage of NA consumption (53%).

HBS/NA: Severe underreporting exists in multiple categories: CP124 (14%), CP126 (3.2%), CP081 (8.4%) and CP052 (5.5%). Many other categories exhibit moderate underestimation (30-50% below NA).

SILC/NA: The matching lessens underestimation for several categories. However, severe underestimation persists in CP124 (8.8%), CP101 (15.3%) and CP126 (2.9%). Significant underestimation is evident in various categories and level 1 categories CP10 (21.7%), CP02 (26.8%), CP09 (30.3%), CP11 (43.6%) and CP12 (41%).

SILC/HBS: The matching inflates expenses in several categories. The most notable overestimations occur in CP127 (193.6%), CP125 (124.4%), CP104 (114.8%), and CP073 (111.3%).

SILC/NA: While significant overestimation is present compared to HBS, only a few subcategories very slightly exceed NA figures: CP071 (127%) and CP125 (124.4%). None of the level 1 categories demonstrate overestimation.

The matching procedure for Slovakia 2020 does not consistently improve the coverage of NA consumption compared to the original HBS data. While introducing overestimation in certain areas, underreporting (particularly in CP10 and CP02 categories) persists across multiple categories in the matched SILC dataset.

A.3.4 SILC year – 2019

Austria

The HBS data seems to underestimate consumption in National Accounts, but matching improves coverage by more than seventeen percentage points. Matching enhances the overall coverage of NA consumption, but also introduces overestimation in comparison to HBS across various subcategories.

There are no instances of underestimation when examining two-digit COICOP codes, but there is one case of underestimation at the three-digit subcategory level, for CP121.

The matching procedure inflates expenses across almost all COICOP categories compared to the original HBS, but given the overall underreporting of expenses in HBS, matched SILC remains below NA in most categories, with the most notable exception being CP10.

There is severe overestimation in matched SILC/HBS (exceeding 30%) for CP03 and CP04. This seems to be driven by overestimation in subgroup CP031, while no subcategory of CP04 appears to be overestimated in a severe way.

Belgium

The matching performs relatively well in terms for expenditures at the one-digit aggregate COICOP categories. Expenditures imputed from HBS to SILC do not depart more than 20% from the original source survey (HBS) for all COICOP one-digit aggregate categories except from CP06 (health) and CP12 (others), where the imputed expenditures are about 132% and 122% of those reported in HBS, respectively.

When we compare these matched expenditures in SILC (mSILC) with expenditures from National Accounts (NA), gaps are larger. The mSILC/NA coverage ranges from 45% in CP02 -alcohol and tobacco– to 120% in CP08 -communication-.

- The low coverage of category CP02 (Alcoholic beverages and tobacco) comes entirely from the under-reporting in HBS (the ratio HBS/NA is 43%, even slightly lower than mSILC/NA). This is not a Belgium-specific problem, as these types of products tend to be under-reported in household surveys. In fact, the HBS/NA coverage in Belgium is quite above the EU average (36.3%). When we look at the specific products within this category, we see that the consumption of tobacco (CP02.2) is the main driver of this under-reported consumption in HBS with respect to NA (27.3%), whereas the coverage for alcoholic beverages (CP02.1) is still below 100% but much higher, of 72.2%.
- The overestimation of expenditures in CP08 (Communication) is also mainly driven by the original over-reporting (an HBS/NA ratio of 114%), which is exacerbated after the matching. The distortion introduced by the matching is mainly observed in category CP 08.1 (postal services): expenditures in the matched SILC double those reported in the original HBS file. The rest (CP08.2 and CP08.3 - i.e., telephone equipment and services) are both over-reported in HBS w.r.t NA and get slightly reinforced with the match.
- In contrast, the over-estimation of expenditures in CP07 with respect to NA is mainly driven by the matching, as the original HBS/NA ratio is 104%. Looking at the components of CP07, we can clearly see that this over-reporting of HBS w.r.t NA comes from category CP071 (purchase of vehicles).

Bulgaria

Overall, on average the matching is far away from HBS in terms of aggregate expenses (%). The deviation is about 24%. There is a general and severe underestimation of total expenses by HBS over NA (%). The underestimation is deemed as severe (>30%) for all first level COICOP groups except CP01 and CP04. On the 3-digit COICOP level, the underestimation in HBS over NA is very strong.

We observe large deviations (larger than 30%) in subcategories of CP012, CP021, CP022, CP031, CP041, CP051, CP052, CP053, CP054, CP055, CP061, CP062, CP063, CP071, CP072, CP073, CP081, CP082, CP083, CP091, CP092, CP093, CP094, CP095, CP102, CP103, CP104, CP105, CP111, CP112, CP123, CP125, CP126 and CP127.

On the 3-digit COICOP level, matching performs well (deviations below 15%) except in categories CP041, CP043, CP051, CP054, CP055, CP063, CP072, CP081, CP082, CP091, CP092, CP093, CP096, CP102, CP104, CP112, CP124 and CP127.

Cyprus

Overall, it seems that HBS underreports consumption in NA (weighted average of the shares of aggregate expenses of HBS/NA is 76.5%). The matching process generally results in improved coverage of NA consumption (80.2%) with respect to the performance of original HBS across some COICOP groups and their two-digit level subcategories.

Notably, in group CP06, the matching inflates the original HBS figures. This trend is particularly evident in specific three-digit subcategories such as CP063, where the reported expenses after matching exceed those originally reported by HBS by varying extents.

Similarly, in CP10, subcategory CP104 also shows a decrease in reported expenses post-matching. This case of underestimation highlights that the matching process does not uniformly increase expense figures and can differentially impact detailed product groups and subcategories.

Overall, the matching procedure for Cyprus yields results that are relatively close to those of the HBS.

Czechia

The HBS data (2019) appears to underreport consumption in National Accounts (weighted average of the shares of aggregate expenses of HBS/NA is 57.4%). The matching process slightly improves the coverage of NA consumption (64.6%). The matched SILC/HBS figures show an overall increase in expenses compared to the original HBS data (112.9%).

HBS/NA: Several categories and subcategories exhibit underreporting, some severely: CP124 (11.3%), CP104 (28.8%), CP112 (24.6%), CP126 (14.7%), and CP02 (18.7%). A number of additional categories fall in the moderate underestimation range (30-50% below NA).

SILC/NA: The matching process lessens the underestimation for many categories. However, a few remain significantly underestimated, most notably CP103 (58.4%), CP104 (33.2%), CP091 (32.7%), and CP063 (56.9%). Additionally, all level 1 categories remain underestimated in the matched SILC dataset, with the severe underestimation for CP02 (20%).

SILC/HBS: The matching process inflates expenses across many categories and subcategories. Notable overestimations are present in CP127 (223.9%), CP082 (217.5%), CP102 (98.3%) and CP051 (121.2%).

SILC/NA: While significant inflation is observed compared to HBS, the matched SILC figures exceed NA values for only a few categories: CP082 (127.9%), CP127 (119.4%), CP041 (129.5%) and in CP102 (123.7%). There is no overestimation in level 1 categories.

The matching procedure for improves the coverage of NA consumption compared to the original HBS but does not fully eliminate underreporting. While overestimation is introduced in many areas, most categories remain below NA levels in the matched SILC dataset.

Germany

The matching performs relatively well in terms for expenditures at the one-digit aggregate COICOP categories. Expenditures imputed from HBS to SILC do not depart more than 20% from the original source survey (HBS) for all COICOP one-digit aggregate categories except for CP12 ("Other", where matched expenditures are

143% of original HBS, driven by CP12.5, insurance services) and CP10 ("Education", with an extremely low ratio of 25%).

Of lower order, under-simulation (when comparing SILC to HBS) of about 10% is also present in category CP04 (housing), whereas over-simulation of about 10%-15% takes place in CP06 (health) and CP08 (communication).

When we compare these matched expenditures in SILC (mSILC) with expenditures from National Accounts (NA), gaps are larger. The categories with the smallest coverage are CP10 ("Education", 19%), CP02 ("Alcoholic beverages and tobacco", 41%) and CP05 ("Furnishing", 56%).

The very small share of expenditures in mSILC with respect to NA for the CP10 category is mainly driven by the distortion of the matching, as the original HBS/NA ratio is much larger (74.5%).

In contrast, the low coverage in CP02 and CP05 is driven by the original low HBS/NA rates. In the case of CP02 ("Alcoholic beverages and tobacco") the HBS/NA ratio is 38%. Although this ratio looks quite low, it is even slightly above the EU average. In the case of CP05 ("Furnishings"), the HBS/NA ratio is 58%.

Denmark

The matching data do generally well and do not produce any big distortion compared to original HBS (most of them are <30%). There are few exceptions to this which includes 3-digit level of 063, 071, 081, 092, 096 and 126 where the matching produces a more severe distortion. COICOP 02 expenditures (alcohol and tobacco) are fully under-reported in HBS and as such results under-reported also in the matched dataset but that does not depend on the matching procedure. Other problematic categories to be signalled regard transport expenses (CP07) where expenditures for CP071 are highly over-reported in HBS while results under-reported when looking at the matched dataset, this is the result of both imprecise starting data as well as distortions happening because of the matching procedure. COICOP 10 also seem to be quite distorted by the matching procedure especially for CP101 and CP105

General issues related to HBS original data regards over-reporting for CP125 and CP043, and under-reporting (more frequent) for CP022, CP063, CP091, CP092, CP094, CP105, CP124 and CP126.

Estonia

The matching data do generally well and do not produce any big distortion compared to original HBS. At the level one, distortions are generally of the order of 10%, whereas at the level 2 they are typically below 30%. A few exceptions to this include 081, 103 and 127 where the matching produces severe distortions (beyond 50%). When compared to national account, COICOP 02 (alcohol and tobacco), COICOP 03 (clothing and footwear), COICOP 11 (restaurant and accommodation services) and COICOP 12 (miscellaneous good and service) are significantly under-reported in HBS and as such results under-reported also in the matched dataset but that does not depend on the matching procedure.

In the following case the comparison with NA is especially poor: CP112, CP126 and CP124 are dramatically underestimated, while CP101 is largely overestimated.

Greece

Overall, HBS seems to be underreporting the consumption in NA (in all categories apart from CP06). Underreporting becomes even more pronounced when we compare the matched SILC with NA, with six categories exceeding the acceptable range of 30% (CP02, CP07, CP08, CP09, CP11 and CP12).

The matched SILC/HBS rates are all in the range of 90%-110% for all categories apart from CP06 and CP10.

CP06 seems to be the most worrisome category: the overall consumption reported in HBS is very close to the one in NA, but the level 2 analysis shows that in item CP063 HBS there is a significant overestimation (HBS/NA≈300%). The matching process creates a significant overestimation of this category's overall expenditure (matched SILC/HBS ratio is at 165%).

Spain

Overall, HBS seems to underreport consumption compared with National Accounts data for all COICOP categories.

However, overall, underreporting becomes less pronounced with the statistically matched dataset of the EU-SILC with HBS, compared to National Accounts. Four COICOP categories have deviations larger than 30% (CP02, CP09, CP11 and CP12). From the remaining eight categories, three closely match National Accounts (CP03, CP04 and CP08).

The statistically matched dataset of the EU-SILC with HBS is similar to HBS data in this regard, with all COICOP level 1 categories over reported between 5 to 25%, except for CP02 for which there is a strong underreporting.

Finland

In general, the HBS appear to be overstating consumption figures in the NA, with a weighted average of aggregate expenses shares reaching 131%. The matching process generally results in an additional overrepresentation of NA consumption coverage compared to the original HBS data across various three-digit subcategories, exacerbating the issue of consumption over reporting.

Particularly noteworthy is the substantial increase of consumption in specific three-digit subcategories such as CP043, CP044, CP071, CP081, CP112, CP125 and CP127, where reported expenses after matching significantly exceed those originally reported by HBS.

Despite the overall overestimation in coverage, there are instances of decreased coverage. In COICOP groups CP02, CP03, CP09, CP10, and CP11 all subcategories report lower expenses after matching compared to HBS. Furthermore, in other COICOP groups, three-digit subcategories exhibit a decrease in reported expenses post-matching. Notably, in CP022, CP031, CP032, CP054, CP062, CP092, CP094, CP105, CP111, CP123, and CP124, the matching adjustment leads to a significantly lower expenditure than in HBS. These cases of underestimation underscore the non-uniform impact of the matching process on expense figures across detailed groups and subcategories.

France

The original HBS data seems to be underestimating consumption figures over NA, with a weighted average of aggregate expenses shares equal to 74%. However, the only first level COICOP groups deemed as severe underreported (>30%) are: CP02, CP06 and CP11.

The matching process generally results in an additional underestimation of NA consumption coverage compared to the original HBS data (71.2%). The severe underreporting persist for the mentioned three COICOP categories (particularly worsening for CP02 but remaining similar for CP06 and CP11), as well as CP09, when comparing the HBS matched to SILC dataset to the NA.

Moreover, the matched SILC/HBS rates are in the acceptable range of 82.7%-101.82% for all COICOP level 1 categories. On the 3-digit COICOP level, the matching performs well, where we do not have deviations larger than 30%, except for CP121 (64.54%).

Croatia

Overall, HBS seems to be underreporting the consumption in NA in all COICOP categories.

Underreporting becomes even more pronounced when we compare the matched SILC with NA, with only three categories being within the acceptable range of 30% (CP01, CP03 and CP04). From the remaining nine categories, the most underreported is CP11, followed by CP02, CP09 and CP06.

The matched SILC/HBS rates are in the (acceptable) range of 94%-112% for all COICOP level 1 categories for all categories apart from CP10 (84%).

Hungary

Overall, on average the matched SILC is extremely close to HBS in terms of aggregate expenses (107% on average). There is a general underestimation of total expenses by HBS over NA (59%). The underreporting from HBS is greater than 30% (severe) for most first level COICOP groups except CP04 (Housing, water, electricity, gas and other fuels) that is over-reported by about 46% and CP08 (Communication) that results accurate compared to the national account figures.

The matching process reproduces the same patterns of consumption as for HBS in SILC and aggregates expenditures coverage at COICOP level 1 is very similar (matched SILC on average captures 54% of NA).

The matching performs well also at 3-digit COICOP level: it replicates closely HBS coverage with an average overestimation of 9%. However, users should be aware that because of the underreporting of many expenditures in HBS representing only 65% of NA expenditures. Consequently also matched SILC+HBS underestimates expenditures compare to NA by 48% on average (representing 52% of NA on average). Although most deviation from NA represent underestimation, some categories are over-reported such as CP043 CP044 CP045 CP083 and CP127.

Ireland

Overall, the HBS data appears to be underreporting consumption in NA (weighted average of the shares of aggregate expenses of HBS/NA is 82.2%). Further, the matching process results in enhanced coverage of NA consumption (89.7%). Conversely, the matched SILC/HBS is over reporting consumption concerning the HBS original figures (110.3%).

There are several underreported categories in HBS/NA. Notably, groups such as CP02, CP05, CP06 and CP11 are all below the 30% threshold. This trend is evident within specific three-digit categories such as CP021, CP022, CP054, CP056, CP062, CP063, CP073, CP103, CP111, CP112, CP121, CP126 and CP127, where reported expenses in relation to NA are below the threshold.

Overall, the matched SILC/HBS marginally inflates all categories in relation to the original HBS figures. However, only CP02 and CP06 are above the acceptable threshold. Expanding the analysis to specific 3-digit categories, we verify that CP022, CP043, CP044, CP081, CP092, CP102, CP103, CP126 and CP127 are all overestimated by the matching procedure.

From this analysis, we can conclude that the matching procedure increases expense figures for almost all categories when compared to HBS. This outcome offsets differences in underreported categories initially observed between HBS and NA. However, it worsens comparability for categories initially overestimated in HBS.

Lithuania

The HBS data seems to severely underestimate consumption in National Accounts, with matching improving coverage by more than fifteen percentage points. Matching enhances the overall coverage of NA consumption, but also introduces overestimation in comparison to HBS across various subcategories.

There are no instances of underestimation when examining two-digit COICOP codes, but there are cases of underestimation at the three-digit subcategory level, particularly for CP041, CP092, CP105 and CP124.

The matching procedure inflates expenses across all COICOP categories compared to the original HBS, but given the overall underreporting of expenses in HBS, matched SILC generally remains below NA in all categories except CP10.

There is severe overestimation in matched SILC/HBS (exceeding 30%) for CP03, CP07 and CP10. This seems to be driven by overestimation in the following subcategories: CP031, CP032, CP071, CP102 and CP104. The overestimation of CP102 is particularly severe (by more than one-thousand percent).

Luxembourg

Overall, HBS seems to be underreporting the consumption in NA (weighted average of the shares of aggregate expenses of HBS/NA is 69.8%). Further, the matching process results in enhanced coverage of NA consumption (77.4%). Conversely, the matched SILC/HBS is over reporting consumption concerning the HBS original figures (110.7%).

For HBS/NA, nearly all categories (except CP08) are underestimated. Notably, groups such as CP02, CP06, CP10 and CP12 are all below the 30% threshold. This trend is evident within specific three-digit categories such as CP012, CP021, CP022, CP052, CP054, CP055, CP056, CP062, CP063, CP072, CP081, CP094, CP095, CP102, CP103, CP105, CP123 CP124, CP126 and CP127, where reported expenses in HBS in relation to NA are all below the threshold.

Conversely, for matched SILC/HBS, all categories (except CP10) are inflated with relation to original HBS figures. In particular, CP06 and CP012 are all above the 15% threshold. Expanding the analysis to specific three-digit categories, we verify that CP012, CP041, CP043, CP053, CP061, CP081 and CP093 are all overestimated by the matching procedure.

From this analysis, we can conclude that the matching procedure consistently (and almost uniformly) increases expense figures when compared to HBS. However, since HBS was severely underreported with relation to NA, this offsets the initial differences between HBS and NA, bringing the matched SILC/HBS figures closer to NA for nearly all categories at the 2 digit-level.

Latvia

Overall, on average the matching is close to HBS in terms of aggregate expenses (%). There is a general underestimation of total expenses by HBS over NA (%). The underestimation is deemed as severe (>30%) for all first level COICOP groups except CP01, CP04 and CP08.

On the 3-digit COICOP level, the underestimation in HBS over NA is very strong. We observe large deviations (larger than 30%) in subcategories of CP012, CP021, CP022, CP031, CP032, CP044, CP051, CP053, CP054, CP056, CP062, CP071, CP072, CP073, CP081, CP082, CP091, CP092, CP093, CP094, CP095, CP096, CP101, CP102, CP104, CP105, CP111, CP112, CP121, CP122, CP123, CP124, CP125, CP126 and CP127.

On the 3-digit COICOP level, matching performs well (deviations below 15%) in all categories, except CP031, CP032, CP041, CP052, CP053, CP054, CP056, CP062, CP071, CP081, CP082, CP091, CP093, CP094, CP096, CP101, CP105, CP111, CP112, CP121, CP123, CP125, CP126 and CP127.

Malta

Overall, HBS seems to underreport consumption compared with National Accounts data for all COICOP categories except CP04, in which there is over reporting.

However, overall, underreporting becomes less pronounced with the statistically matched dataset of the EU-SILC with HBS, compared to National Accounts. Three COICOP categories have underreporting deviations larger than 30% (CP02, CP09 and CP11). On the contrary, CP04 over reports by almost 50%. From the remaining eight categories, five closely match National Accounts (CP01, CP03, CP06, CP10 and CP12).

The statistically matched dataset of the EU-SILC with HBS is similar to HBS data in this regard, with all COICOP level 1 categories over reported between 6 to 40%, except for CP02 for which there is a strong underreporting.

Netherlands

The HBS data underreports expenditures compared to National Accounts (NA) on average by 11%. Exceptions are spending on CP04 (Housing, water, electricity, gas and other fuels) which are over reported by 30% and CP10 (Education), over-reported by 67%. Matching enhances coverage of NA expenditures (average underreporting estimated to 6% compared to NA). Spending on CP04 and CP10 are also a bit lower in the matched SILC compared to original HBS, which means they still overestimate NA. The matched SILC+HBS figures tend to overestimate consumption relative to the original HBS figures by about 18%.

The matching performs well also at 3-digit COICOP level: it replicates closely HBS coverage with an average overestimation of 18%. When comparing to the NA, expenditures in matched SILC slightly overestimate aggregate consumption on average by 1%. Although most deviation from NA represent underestimation, some categories are severely overestimated following over-reporting in HBS. Such groups are CP043 (39%) CP044 (almost triple compared to NA, but only 22% higher in matched SILC than HBS), CP055, CP083, CP104 and CP105 (similar to HBS, but much higher than NA, respectively 4 and 6 times higher).

Poland

Overall the matching of Poland performs very well, where the average SILC/HBS share is 105%. COICOP categories that perform less good are the following COICOP categories:

None of the COICOP categories of the first level have a deviation that >120% or <80%.

Compared to NA, HBS underestimates significantly. On average 50.7 percent of the expenditures are captured by HBS. Similarly SILC/NA captures this underestimation with an average share of 54.4%.

Portugal

The original HBS data seems to be generally underestimating consumption figures over NA, with a weighted average of aggregate expenses shares equal to 72.7%. The only first level COICOP groups that are within the 30% range are CP04, CP06, CP07, CP08, and CP10.

The matching process generally worsen the coverage of NA consumption (66%) respect to the performance of original HBS, specifically for the COICOP groups CP03, CP04, CP05, CP07, CP09, CP10, and CP11.

Moreover, the matched SILC/HBS rates are in the acceptable range of 81.53%-106.28% for all COICOP level 1 categories. On the 3-digit COICOP level, the matching performs good, where we do not have deviations larger than 30%, except for CP081 (191.11%), CP103 (55.35%), and CP127 (206.51%).

Romania

The matching data do generally well and do not produce any big distortion compared to original HBS (most of them are <30%). There are few exceptions to this which includes 3-digit level of CP063, CP071, CP081, CP082, CP091, CP092, CP096, CP101, CP102 and CP126 where the matching produces a more severe distortion. Many expenditures categories (almost all except CP04) are fully under-reported in HBS and as such results under-reported also in the matched dataset but that does not depend on the matching procedure. Other problematic categories to be signalled regard transport expenses (CP07) where expenditures for CP081 are highly over-reported in HBS while results under-reported when looking at the matched dataset, this is the result of both imprecise starting data as well as distortions happening because of the matching procedure. COICOP 10 also seem to be quite distorted by the matching procedure especially for CP101 and CP105.

Sweden

Overall the matching of Sweden performs well, where the average SILC / HBS share is 104.7%. COICOP categories that perform less good (>120% or <80%) are the following COICOP categories:

- CP02, where SILC underestimates compared to HBS (77.7%).
- CP04, where SILC overestimates compared to HBS (122%).
- CP06, where SILC severely overestimates compared to HBS (150.4%), due to CP062.
- CP08, where SILC overestimates compared to HBS (126.5%), due to CP081 and CP083.
- CP10, where SILC severely underestimates compared to HBS (27.8%), driven by CP102 and CP107.

Compared to NA, HBS underestimates significantly. On average 60.4% of the expenditures are captured by HBS. Because of the slight relative overestimation in SILC compared to HBS, this increases to 65.1% for SILC vs NA.

Slovenia

Overall, HBS seems to be underreporting consumption in NA for Cyprus (weighted average of the shares of aggregate expenses of HBS/NA is 67.9%). The matching process generally results in slightly enhanced coverage of NA consumption, at 69.3%, compared to the performance of the original HBS across some COICOP groups and their two-digit level subcategories. The matched dataset deviate no more than 15% from the original HBS figures for any COICOP. Overall, the matching procedure for Slovenia yields results that are relatively close to those of the HBS.

Slovakia

The HBS data appears to underreport consumption in National Accounts (weighted average of the shares of aggregate expenses of HBS/NA is 54.5%). The matched SILC/HBS figures show a slight decrease in expenses compared to the original HBS data (97.8%). The matching process results in a decrease in the coverage of NA consumption (53.6%).

HBS/NA: Severe underreporting exists in multiple categories: CP10 (23.3%), CP124 (14%), CP126 (3.2%), and CP052 (5.5%). Many other categories exhibit moderate underestimation (30-50% below NA).

SILC/NA: The matching lessens underestimation for several categories. However, severe underestimation persists in CP10 (18.3%), and significant underestimation is evident in various categories and level 1 categories CP02 (26.3%), CP09 (32.8%) and CP11 (49.9%).

SILC/HBS: The matching inflates expenses in many categories. The most notable overestimations occur in CP041 (167.1%), CP127 (170.6%), CP125 (129.9%), and CP081 (255.5%).

SILC/NA: While overestimation is present compared to HBS, only a few subcategories exceed NA figures: CP125 (129.9%), CP071 (112.5%), and CP044 (114%). None of the level 1 categories demonstrate overestimation.

The matching procedure for does not consistently improve the coverage of NA consumption compared to the original HBS data. While introducing overestimation in certain areas, underreporting (particularly in the CP10 category) persists across multiple categories in the matched SILC dataset.

Getting in touch with the EU

In person

All over the European Union there are hundreds of Europe Direct centres. You can find the address of the centre nearest you online (european-union.europa.eu/contact-eu/meet-us_en).

On the phone or in writing

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

— by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),

— at the following standard number: +32 22999696,

— via the following form: european-union.europa.eu/contact-eu/write-us_en.

Finding information about the EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website (european-union.europa.eu).

EU publications

You can view or order EU publications at op.europa.eu/en/publications. Multiple copies of free publications can be obtained by contacting Europe Direct or your local documentation centre (european-union.europa.eu/contact-eu/meet-us_en).

EU law and related documents

For access to legal information from the EU, including all EU law since 1951 in all the official language versions, go to EUR-Lex (eur-lex.europa.eu).

EU open data

The portal data.europa.eu provides access to open datasets from the EU institutions, bodies and agencies. These can be downloaded and reused for free, for both commercial and non-commercial purposes. The portal also provides access to a wealth of datasets from European countries.

Science for policy

The Joint Research Centre (JRC) provides independent, evidence-based knowledge and science, supporting EU policies to positively impact society

EU Science Hub
Joint-research-centre.ec.europa.eu