JRC Project no. JRC/SVQ/2018/B.2/0021/OC

A new indirect tax tool for EUROMOD

Final Report

Elif Cansu Akoğuz, Bart Capéau, André Decoster, Liebrecht De Sadeleer, Duygu Güner, Kostas Manios, Alari Paulus, & Toon Vanheukelom

November 2020

Contents

1 Introduction

This report is the third and final deliverable of the project JRC/SVQ/2018/B.2/0021/OC, in which we develop a new Indirect Tax Tool for EUROMOD (ITTv3). The objective of the project is to modify the already existing indirect tax tool (ITTv2) as follows:

- increasing the number of countries for which indirect tax simulation can be performed;
- performing the imputation of expenditure variables from the Household Budget Surveys (HBS) to the European Union Statistics on Income and Living Conditions (EU–SILC) datasets at the most detailed level of aggregation available (roughly 200 good categories). This enables the simulation of tax rate changes on narrowly defined good or service categories;
- integrating the ITT, which is currently an add–on, into EUROMOD to increase the model's transparency and ease of use.

In order to achieve these objectives, the following tasks have been executed throughout the project:

- the latest releases of the Eurostat versions of the national HBS micro–data (2010) and the EU–SILC micro–data for the corresponding (or closest available) year are gathered for all member countries of the EU. A selection of 18 countries was made on the basis of relevance of the country and a first quality check of the available data. The datasets of these countries are prepared for imputation by means of standard data cleaning and harmonisation procedures;
- an imputation method is developed in order to meet the challenges of imputing expenditure variables at highly disaggregated levels. A tool kit for evaluating the imputation results of this method is developed and applied to the imputation results for the 18 selected countries;
- the ITT is integrated into the EUROMOD microsimulation model. Three alternative behavioural assumptions for simulating expenditure reactions to price and income fluctuations and associated changes in individual indirect tax burdens, are available: constant quantities, constant income shares, and constant expenditure shares.

The remainder of this report is structured as follows.

Section 2 explains the imputation method. First, the challenges inherent to imputing expenditure variables are underlined. Then, state–of–the–art imputation methods are examined in the light of this discussion and each method's advantages and disadvantages are highlighted. Lastly, an imputation method which attempts to combine the for our purpose attractive features of each state–of–the–art method, is proposed. Section 3 introduces the evaluation tools we developed for assessing the quality of the imputation.

Section 4 describes the process of data preparation prior to imputation. It also presents the criteria for selecting the 18 countries for which the new ITT is developed. During that process a number of inconsistencies were spotted in the HBS expenditure data and Section 4 also explains how we

resolved them. Section 5 discusses the practical implementation of the imputation method, such as the specification of the regression models and the treatment of outliers. In Section 6, imputation results are evaluated with the aid of the tools developed in Section 3. As a final quality check, this section also validates the baseline 2010 results by a comparison with national accounts statistics.

In section 7 a detailed description of the implementation of the indirect tax simulation tool in EUROMOD is provided. Section 8 illustrates the model by simulating a policy reform under the three available behavioural assumptions: constant quantities, constant income shares, and constant expenditure shares. Section 9 concludes the report.

2 Selection of imputation method

A comprehensive fiscal policy simulation often requires accounting for multiple aspects of the fiscal system, including both direct and indirect taxation. This requires data on, among others, both income and expenditures. Unfortunately, it is uncommon to have good quality records on both income and expenditures within one dataset. One possible way to overcome this problem is to merge two distinct datasets, each containing information on one of the desired variables, by taking advantage of the common variables in both datasets. This type of merge is called an imputation, and it can be achieved by various methods, each with their own strengths and weaknesses.

It is not possible to claim that either one of these methods is strictly better than another. The best performing method differs depending on the structure of the data at hand and the intended use of the merged dataset. This section's objective is to introduce and compare state–of–the–art imputation methods, in the light of the qualities expected from a successful imputation as well as the challenges inherent to imputing expenditures. It focuses particularly on imputing expenditure information into a dataset containing income information, as our primary aim is to expand the default dataset on which the direct tax–benefit microsimulation model EUROMOD runs (EU–SILC) and which contains income information, with information on expenditures in order to develop an integrated indirect tax tool.

Generally, one best keeps the dataset which is more comprehensive in other dimensions as the *recipient data*, *i.e.* the dataset *into* which the missing values are imputed, while the other dataset then serves as the *source data*, i.e. the dataset *from* which the missing values are imputed. In the case at hand, the Eurostat version of national Household Budget Surveys (HBS) will serve as the source data from which expenditure information will be drawn to impute into EU–SILC.

The objective of the imputation is to obtain values for the missing information, which would look similar to what can be expected for those values on the basis of the information on the joint distribution of household characteristics (the common variables) and the variables to be imputed within the source dataset. If both datasets are representative for the same population this amounts to the requirement that the imputed values should be such that the inferences regarding the joint population distribution derived from common and imputed variables in the recipient dataset would be equal to that derived from the common and observed variables in the source dataset. Put differently, the source dataset and the recipient dataset (augmented with the imputed values), should be conceived of as stemming from the same data generating process. We will develop below some visual and algebraic tools to inspect in how far an imputation method reaches this objective (see Section 3). We first highlight some specific challenges for imputing expenditures data.

One particular challenge, however, is not specific to imputing expenditures. It occurs when common variables in source and recipient dataset have non–overlapping distributions, more specifically, when a significant portion of the recipient data falls out of the range defined by the support of their distribution in the source. In that case, an imputation method should go further than looking

for observations in the source that closely resemble the deviant (out of range) observations in the recipient. It might then be useful to have a model that predicts how a household's missing values would look like on the basis of functional relations between common variables and variables to be imputed, exhibited in the source dataset. However, as long as it is reasonable to assume that both source and recipient data are generated by the same data generating process, the out of range problem should not be a major issue. Out of range values of certain variables are then true outliers, in the sense of having an almost zero probability to stem from the data generating process (the population). An imputation method would perform quite well if it is robust to the presence of outliers.

2.1 Challenges inherent to imputing expenditure information

There are specific challenges inherent to imputing expenditure data due to the nature of expenditure behaviour, survey methods, and final use of the completed dataset. The challenges introduced below render conventional imputation methods unfavourable and necessitate the construction of a new method that combines the most attractive aspects of each method.

2.1.1 Correlation among variables to be imputed

Economic theory tells that specific interaction patterns might occur between expenditures on several goods or good categories, depending on the degree of substitutability and complementarity such goods exhibit in the consumer's preferences. An imputation method should then try to preserve these interaction patterns as much as possible.

Existing methods coping with this problem (Raghunathan *et al.*, 2001, and van Buuren and Groothuis–Oudshoorn, 2010) are designed for imputation problems where not all of the records for the variables to be imputed are missing. One then uses information of the records with non–missing observations on (some) of these variables to impute the missing values for other records in the dataset. These methods cannot cope with cases such as ours, where all values of the variables in the recipient data are originally missing.

Thus, other methods to safeguard the observed interaction patterns between expenditures on different goods in the source data are needed. The interaction pattern between expenditures can for example be partly determined by (some of) the household characteristics commonly available in source and recipient datasets. An imputation method would then succeed in preserving the interrelation patterns of the variables in so far as it benefits from this information regarding the relation between the household characteristics and expenditures in the source dataset.

2.1.2 Zero expenditures problem

The next challenge, referred to as the 'zero expenditures problem', is the high fraction of zero expenditures observed for certain goods and services that are consumed by only a fraction of the population. Examples include expenditures on alcohol, tobacco, or education. It is not possible to estimate such expenditures accurately by fitting a one–step linear regression model. One can instead use a two–step regression model where the first model with a binary dependent variable determines whether the expenditure is positive or not, while the second model with a continuous dependent variable determines the level of expenditure given that it is positive. However, such a two–step modelling approach poses its own problems. For instance, when the variables for which values have to be imputed are shares which necessarily add up to one, the fitted values obtained by two–step regression models do not necessarily satisfy this requirement, or, there is no unique way to make sure such a restriction holds.

2.1.3 Infrequent expenditures problem

The third and final challenge can be referred to as the 'infrequent expenditures problem' and becomes increasingly important as the expenditures to be imputed become more detailed. This problem arises due to the fact that the more specific an expenditure becomes, the more infrequent and volatile its purchase gets. Food expenditure of a household is not expected to vary greatly from one month to another, though expenditure on Gouda cheese may vary significantly over time due to personal and temporal factors that can not be observed in the data (*e.g.* love for variety, the supermarket running out of another type of cheese, or the last time that the product is consumed). Additionally, at such a detailed level, decisions to buy a particular good might be affected by temporarily variations in preferences (usually I drink coffee, but every now and then I want a tea).

Other reasons to why we observe infrequent expenditures might be bulk purchasing (*e.g.* toilet paper) and durability of certain goods. So–called durable goods deliver services over a longer period of time and have to be replaced therefore only at long intervals. Usually the time period between renewals is much longer than the time window of the survey. Surveys capture only a snapshot of households' year–long expenditures and depending on the timing of the snapshot, the observed expenditure of a household will vary considerably.

Thus, regression models, even two–step models, using household characteristics as covariates are not well suited to explain these type of observed expenditures, let alone predict unobserved behaviour.

2.2 State–of–the–art imputation methods

This section provides a summary of state of the art imputation methods along with their advantages and disadvantages in imputing expenditure information in the light of the aforementioned challenges.

2.2.1 Imputation by means of predicted values of a regression model

This approach involves estimating a regression model on the source data where the common variables, $\mathbf{x}$, are the covariates and the variable to be imputed, say z, is the dependent variable. The missing values are imputed with their fitted values obtained from the estimated regression model. This model can be parametric, semi–parametric or non–parametric. An example of a parametric approach consists of estimating the following linear regression model:

$$z_{sh} = \boldsymbol{\beta}' \mathbf{x}_{sh} + \varepsilon_h \tag{1}$$

where $\mathbf{x}_{sh}$ are the characteristics of the household h in the source dataset s, such as disposable income, number of workers in the household, or the region that the household resides in, while z_{sh} corresponds to the observed value of the variable for household h in the source dataset s, to be imputed later on into the recipient dataset. The model is estimated with observations in the source dataset. The estimated model is then used to obtain fitted values for observations in the recipient dataset. Thus, it is important to emphasize that the household characteristics that constitute these covariates, need to be available in *both* source *and* recipient datasets. There are two variations to this approach. One can either impute a value of z_{rg} for household g in the recipient dataset, by using only the deterministic part of the regression model, that is:

$$z_{rg} = \widehat{\boldsymbol{\beta}}' \mathbf{x}_{rg}, \tag{2}$$

or one can add a random term to this value. It is common practice to use a draw from the error terms of the estimated model in the source data for that purpose.

One of the useful features of the regression based approach is that it allows the user to perform imputation with respect to a behavioural model. The existing EUROMOD Indirect Tax Tool (ITTv2) relies on such an approach. The regression equations consist of Engel curves, explaining expenditure shares on the basis of disposable income and household characteristics, and there is ample economic theory which derives such functional relationship from preference maximisation subject to a budget constraint (Gorman, 1981). Assuming that the behavioural model, *i.e.* the regression model, is correctly specified, it can successfully impute the expenditures of households whose characteristics lay outside the limits of the observed values, such as *e.g.* on other dataset, a quality other imputation methods fail to meet. Another advantage of imputing with a behavioural model is that it enables to simulate behavioural reactions as a response to changing household characteristics (for instance, a change in disposable income). However, the experience with ITTv2 shows that the out of sample range performance and behavioural simulation of the Engel curve method often performs poorly in practice.

The Engel curve approach is in principle also well suited to preserve the correlation between the observed expenditures, at least to the extent that it can be explained by the observable household characteristics in the data. For instance, if the number of cars owned by a household is used as explanatory variable, and it is positively correlated with expenditures on private transportation

and negatively correlated with expenditures on public transportation, the negative correlation between these expenditures arising from the availability of a car will be preserved in the imputed dataset. However, if there exists another channel that correlates these two expenditures which is not observed in the data, such as the proximity of one's house to one's workplace or the availability of public transportation infrastructure in the neighbourhood of the house, that correlation will not be reflected in the imputed dataset. Therefore, the performance of the Engel curve approach in maintaining the correlation structure between expenditures depends on the explanatory power of the common household characteristics. A system of equations approach that uses implicit cross–sectional restrictions (*e.g.* that budget shares should add–up to one) might further improve the capacity of this method to preserve the observed correlation structure of the vector of expenditures.

The main disadvantage of this method is that it does not allow imputing expenditures at the most detailed level. Expenditure data on goods recorded at a very low level of aggregation contain many zeroes. This has been attributed earlier to infrequent expenditures behaviour. Continuous models for this type of observations are misspecified and therefore do not produce very reliable results. Usually the problem is avoided by aggregating expenditures into broader categories of goods. However, for some durables which have a rather long life cycle, this will not solve the problem.

The zero expenditure problem is in this framework usually approached by fitting a two–step regression model, where in the first step a prediction is made whether the household consumes the good or not. For those consuming the good, a normal continuous regression is fitted. However, also this solution works only at a sufficiently high level of aggregation. At a more detailed level, the proportion of observations with positive expenditures might be too small to allow for estimating the continuous regression model. Nevertheless, some tax–benefit systems may contain taxes imposed on very specific categories of products. It is then necessary to have information on the expenditures of households on those specific categories in order to simulate the effects of changes in the design of such specific taxes successfully. For instance, not all alcoholic beverages face the same tax rate. So, information on the consumption of different alcohol products is required to successfully simulate tax liabilities on alcoholic beverages.

2.2.2 Hot Deck Matching (HDM)

This approach corresponds to matching each record in the recipient dataset with a record from the source dataset. The missing values of the household in the recipient dataset are then filled in by the observed values of the matched record. The method of matching determines the type of hot deck. There are thus numerous hot deck imputation methods, of which the distance hot deck is the most familiar one.[1]

[1] Non distance–based HDM methods use other criteria than distance measures to determine one or more similar observations in the source from which values for variables missing in corresponding observations of the recipient can be obtained.

The distance hot deck approach measures the proximity of a pair of records by using a distance metric that takes common variables of these records as input. For instance, if there exists only one common variable, say income, using any standard distance metric will amount to taking the absolute value of the difference between the incomes of these records. Each record in the recipient file is then matched with either the closest record in the source data or it is matched randomly to one of the K closest records, depending on the choice of method.

In the case of more than one common variable, the choice of the distance measure matters. There are many different distance measures available, however among them the Mahalanobis distance is the most widely used. Its primary advantage is that it accounts for the variances and covariances of the variables used to construct the distance measure. Recall that $\mathbf{x}_{sh}$ is the vector of values of the common variables for household h in the source dataset, and, similarly, $\mathbf{x}_{rg}$ are the observations for the corresponding vector for household g in the recipient dataset. The Mahalanobis distance is then defined as

$$d\left(\mathbf{x}_{rg}, \mathbf{x}_{sh}\right) = \sqrt{\left(\mathbf{x}_{rg} - \mathbf{x}_{sh}\right)' \Sigma^{-1} \left(\mathbf{x}_{rg} - \mathbf{x}_{sh}\right)} \tag{3}$$

where Σ stands for the covariance matrix of the common variables. The latter is usually calculated using observations from both source and recipient datasets.

The Mahalanobis distance is calculated for each pair of observations consisting of one observation from the source and one observation from the recipient data. Then, for each household g in the recipient dataset, the household h with the smallest distance with respect to g, is selected as the match (or the K closest households are selected, and one is chosen at random as the match). On then uses the observed values for the household h in the source of the variables to be imputed, as imputed values of those variables for the recipient household g.

The HDM approach matches records based only on the similarities between their common variables. The differences between all common variable pairs are either treated as equally important (implicit weights of 1), or they are assigned externally determined weights. In both cases, the weights of the common variables in the distance function are not determined in proportion to their capacity to explain observed expenditure patterns. Therefore, this method fails to benefit from the relationships between household characteristics and expenditures in the source data. In other words, it fails to recognise in which characteristics the matched households should bear similarities with each other in order to have similar expected expenditure profiles. Yet, it is favourable in other aspects. Its greatest advantage is that it allows expenditures to be imputed in as much detail as needed, since it involves no regressions. This implies that imputing zero, or infrequent, expenditures poses no problem.

Despite failing to benefit from the potential relation between household characteristics and expenditures in the source data, the correlation between expenditures can potentially be largely preserved with this method. That is because all missing expenditures of a household in the recipient data are imputed from a single household in the source data. In such a setting, the correlation between expenditures can be largely preserved as long as the distributions of households characteristics in the source and the recipient datasets overlap well enough. Otherwise, there is no guarantee that the

correlation will be preserved since the records in the donor data would be matched disproportionately: while the expenditures belonging to some records in the source data will be imputed many times, some others will be imputed only once or not at all. Naturally this will result in changing the sample correlation between expenditures.

2.2.3 Predictive Mean Matching (PMM)

The predictive mean matching approach combines regression based imputation and hot deck matching approaches. Its most basic application (as for all other methods) pertains to imputing a single variable (or multiple independent variables) with missing values. This univariate case is described step by step below.

1) First, a regression model is estimated on the source data where the variable to be imputed is the dependent variable and the common variables are the covariates, as in regression based imputation (see Equation 1).

2) Then, the variable to be imputed is fitted for *both* recipient *and* source data. Note that, in contrast to regression based methods, fitted values from the estimated model are also produced for observations in the source data.

3) The distances between households in source and recipient datasets are constructed on the basis of these fitted values. As there is only one variable to impute, the absolute value of the difference between fitted values of an observation h in the source, and an observation g in the recipient constitute this distance. The pair (g, h) with the closest distance (or a random selection in the set of pairs $(g, h_1), (g, h_2), \ldots, (g, h_K)$ with the K closest distances) form a match.

4) Missing values in the recipient data are filled in with the *observed* values of the matching records in the source data.

Essentially, the PMM approach can be defined as a specific type of HDM which uses a distance metric that assigns corresponding regression coefficients as weights to the differences between the values of the variables entering the distance function. By doing so, it succeeds to benefit from the information regarding the relation between household characteristics and the variable to be imputed in the source data. Thus, it matches households with respect to characteristics that have a strong relation with the variable to be imputed rather than trying to match with respect to the full set of characteristics some of which are not informative in predicting missing values.

Since it is based on regressions, the method does not perform well when it comes to imputing values for expenditures on a detailed level of aggregation.

2.3 Imputation method

We now explain in detail the imputation method we have developed in the course of this project (Section 2.3.1) and discuss some of its limitations (Section 2.3.2).

The imputation method we propose tries to combine the best properties of existing methods in order to meet the specific challenges listed above when imputing expenditures. The idea is to use fitted values of regressions of broad categories of expenditures on the common variables (household characteristics) as inputs in a distance measure to obtain a match.

Even for such broadly defined expenditure categories, a (substantial) amount of households may indicate zero expenditures. At this level of aggregation we interpret these as true zeros, *i.e.* goods not consumed by the household, not as the result of infrequent expenditures. Therefore, a standard two–step procedure is followed for modelling such true zeros. A binary regression model determines the probability of positive expenditures and an ordinary continuous regression equation is used for modelling means of the dependent variable (expenditures) conditional upon household characteristics and expenditures being positive.

After estimating this two–step model we can fit expected expenditures on those broad categories conditional on household characteristics, *i.e.* the estimated probability of positive expenditures (resulting from the binary regression) times the expected amount of expenditures conditional on expenditures being positive (resulting from the continuous regression). Such fitted values were calculated for both households in the source and recipient dataset. In line with PMM, these fitted expenditures are then used as inputs for calculating the distance between pairs of households. As there is more than one fitted variable (one for each of the broad expenditure categories), a distance measure needs to be chosen, and we stick to the Mahalanobis distance, which is most commonly used one among distance based HDM's. We then use a HDM approach to match each household in the recipient dataset to a household in the source dataset. A household h in the source dataset is matched to a household g in the recipient dataset with the smallest distance to h. We then use the observed values for household h, of expenditures at the lowest level of commodity aggregation as imputed values for the recipient household g's expenditures on those commodities.

By using observed rather than fitted values resulting from a regression, this method can successfully impute values at a detailed level of aggregation. Moreover, the regressions exploit information on the relation between household characteristics (that is, the explanatory variables in the regression) and expenditures (dependent variables) in the dataset, which is neglected by the traditional HDM.

2.3.1 Imputation procedure

We now give a step–by–step description of our imputation method.

1) A household h's expenditure on a good i in the source dataset (the HBS in our case), indexed by

s), denoted by e_{shi}, is converted into a share, w_{shi}, of disposable income (y_{sh}):

$$w_{shi} = \frac{e_{shi}}{y_{sh}}, \qquad i \in \mathcal{N}, \tag{4}$$

with $\mathcal{N}$ the set of indices of goods at the most detailed level in the HBS.

2) These income shares of expenditures on detailed goods are aggregated under broader categories. These categories should be big enough to reduce the infrequent expenditure problem (see Section 2.1.3), but small enough to allow household characteristics to explain differences in allocations of income across these goods. We index these categories by $A, B, \ldots$. Consequently, the indices $A, B, \ldots$ denote non–overlapping and non–empty subsets of $\mathcal{N}$, denoted by $\mathcal{N}_A, \mathcal{N}_B, \ldots$, whose union equals $\mathcal{N}$. Thus, the income share of expenditure category X for $X = A, B, \ldots$, say W_{shX} , equals:

$$W_{shX} \equiv \sum_{i \in \mathcal{N}_X} w_{shi}. \tag{5}$$

3) The purpose is to develop a multidimensional PMM method by constructing a distance function that takes values of income shares of aggregate expenditure categories fitted from a regression model as inputs. Some of the broad categories may still contain a significant number of zero observations. For a sufficiently high level of aggregation, we can consider these as true zeros, i.e. they are not a consequence of the infrequent expenditures problem.[2] Therefore, we propose a two–step approach for modelling these aggregates. The probability that a household exhibits positive expenditures on commodity aggregate X ($X = A, B, \ldots$) is modelled by a binary model, more specifically, a probit model, using the common variables in the source and destination data as explanatory variables. Formally,

$$\Pr\left(W_{shX} > 0\right) = 1 - \Phi\left(-\boldsymbol{\gamma}_X' \mathbf{x}_{sh}\right) = \Phi\left(\boldsymbol{\gamma}_X' \mathbf{x}_{sh}\right), \tag{6}$$

where Φ denotes the standard normal distribution function, $\mathbf{x}_{sh}$ is the vector of values of explanatory variables for household h in the source dataset s, and the vector $\boldsymbol{\gamma}_X$ contains parameters to be estimated.
Next, an ordinary continuous regression model is formulated for assessing the relation of *positive income shares of broad expenditure categories with the common variables*:

$$W_{shX} = \boldsymbol{\beta}_X' \mathbf{x}_{sh} + \varepsilon_{hX}, \qquad \text{for } W_{shX} > 0. \tag{7}$$

As we are only after correlations, no sample selection correction terms are added to this equation.

The model is estimated on the source dataset, and estimated parameters are denoted with a hat: $\widehat{\boldsymbol{\gamma}}_X$ for the probit models, and $\widehat{\boldsymbol{\beta}}_X$ for the linear regression models.

4) Using the estimated models, values are fitted for the income shares of expenditures on the broad categories $A, B, \ldots$, for all households in *both* the source *and* the recipient datasets, indexed by s,

[2] See Sections 2.1.3 and 2.1.2 for a explanation of the zero and infrequent expenditures problems .

respectively r. These fitted values are denoted by $\widehat{W}_{dhX}$ and defined as follows:

$$\widehat{W}_{dhX} = \Phi\left(\widehat{\boldsymbol{\gamma}}'_X \mathbf{x}_{dh}\right) \widehat{\boldsymbol{\beta}}'_X \mathbf{x}_{dh} \qquad \text{for } d = s, r, \tag{8}$$

where the first factor on the RHS corresponds to the estimated probability that household h has positive expenditure on the aggregate category X, while the second factor corresponds to the estimated income share that household h spends on aggregate expenditure category X, given that this share is positive.

Before we construct a distance function on the basis of these fitted values, we want to assess the extent to which the estimated two–step model is able to explain households' expenditure behaviour. Thereto we construct a pseudo–R^2 value which indicates how much of the variance in the dependent variable is explained by the fitted values our model generates:

$$\text{pseudo}{-}R^2\left(X\right) = 1 - \frac{\sum_h \left(W_{shX} - \widehat{W}_{shX}\right)^2}{\sum_h \left(W_{shX} - \overline{W}_{sX}\right)^2}, \tag{9}$$

where $\overline{W}_{sX} = \sum_h W_{shX}/H_s$, and H_s is the number of households in the source dataset, and sums run over all observations in the source dataset.

Only categories exhibiting a 'reasonable' fit according to this pseudo–R^2 are retained as inputs for calculating the distance between two households. Including variables with a low fit would imply comparing households on the basis of only a small fraction of their true expenditure behaviour. So we could consider two households to be close to each other on that basis, while in reality they might fall far apart, or *vice versa*. We discuss the choice of the threshold level for the pseudo–R^2's in Section 5.3.

5) Denoting a vector of fitted shares retained as input for the distance by $\widehat{\mathbf{W}}_{dh} \equiv \left(\widehat{W}_{dhA}, \widehat{W}_{dhB}, \ldots\right)$ $(d = s, r)$, and 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:

$$dist\left(h, g\right) = d\left(\widehat{\mathbf{W}}_{rg}, \widehat{\mathbf{W}}_{sh}\right) = \sqrt{\left(\widehat{\mathbf{W}}_{rg} - \widehat{\mathbf{W}}_{sh}\right)' \Sigma^{-1} \left(\widehat{\mathbf{W}}_{rg} - \widehat{\mathbf{W}}_{sh}\right)}, \tag{10}$$

where Σ here stands for the variance covariance matrix of the vector $\widehat{\mathbf{W}}$, 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 this distance is measured in terms of Equation (10).

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

$$w_{rgi} = w_{shi}. \tag{11}$$

2.3.2 Limitations of the method

The method described above comes with some limitations. Firstly, it does not enable to impute expenditures of households with non–positive income. That is because households are matched with each other with respect to a distance function that takes fitted values of expenditures as inputs. These fitted values are obtained by means of a regression model that takes the logarithm of income as input. In fact our approach makes only sense for households with a sufficiently high and positive income. Expenditure behaviour of agents with negative or extremely small positive income, do not fit into our model. Indeed, the concept of an income share in terms of which our model is specified, makes not much sense in case of negative incomes, is not defined in case of zero incomes, and may yield extreme values in case of incomes close to zero.

Furthermore, we can only make reasonable predictions on expenditures on sufficiently aggregated, broad categories, so that only such aggregates enter into the distance function. Therefore, there is no guarantee that matched households will bear very similar characteristics. There might very well be two households with very different sets of characteristics both of which indicate similar levels of expenditure on several broad categories. Consequently, when two households with very different characteristics are matched, there is no reason to expect them to have similar expenditure behaviour when it comes to allocating their budget for such a broad category to the commodities belonging to that category.

3 Evaluation of the imputation method

The development of our evaluation methodology rests upon the basic assumption that the SILC and HBS are representative for the same population. Let $\mathbf{x}$ be the vector of common variables in SILC and HBS containing information on socio–demographic characteristics[3] and $\mathbf{w}$ the vector of income shares of expenditures at the most detailed level available. The HBS data contain information on both type of variables, which allows to do inference on the joint distribution of these variables. Denote the estimated joint distribution of $(\mathbf{x}, \mathbf{w})$ by $\hat{f}_{\text{HBS}}(\mathbf{x}, \mathbf{w})$. After imputation, also the SILC will contain values for $(\mathbf{x}, \mathbf{w})$ so that we can do the same exercise, to arrive at an estimate of their joint distribution, say $\hat{f}_{\text{SILC}}(\mathbf{x}, \mathbf{w})$. Ideally, both estimates should be close to each other. Differences might be caused by two reasons. First, differences between $\hat{f}_{\text{HBS}}(\mathbf{x}, \mathbf{w})$ and $\hat{f}_{\text{SILC}}(\mathbf{x}, \mathbf{w})$ can originate from differences in the marginal distribution of the income shares of expenditures inferred from the HBS dataset and the one derived from the SILC with imputed income shares of expenditure. This suggests that the imputation method performs not that well for the considered country. Second, differences between $\hat{f}_{\text{HBS}}(\mathbf{x}, \mathbf{w})$ and $\hat{f}_{\text{SILC}}(\mathbf{x}, \mathbf{w})$ can be caused by differences in the inference drawn on the distribution of socio–demographic characteristics of the population by the two datasets. This is something beyond our control, and might explain why imputations perform better for one country than another. In that case it has less to do with our proposed imputation method, but rather with the data at hand.

In order to disentangle what stems from the SILC and HBS datasets exhibiting different inference on the socio-demographic characteristics of the population, and what is due to the imputation method performing better in one country than another, we propose two devices, which we discuss more in detail in the next two subsections.

Given the high dimensionality of the distributions to be compared, the problem becomes intractable from a practical point of view. We therefore simplify things, first by considering only the performance of the imputation with respect to the 20 broad aggregates which we constructed for running the regressions, and, second, by concentrating on the conditional distributions for each socio–demographic characteristic separately, neglecting the impact of other dimensions.

3.1 Ventile tables and graphs

The objective of the first tool is to visualise the distribution of the income shares in both datasets, conditional upon the socio–demographic characteristics. As income is for the purpose of this project the main socio–demographic characteristic, we concentrate on this variable. The information on the income distribution in both SILC and HBS is summarised by calculating the weighted mean disposable household income per ventile of incomes.[4] Differences in these estimated means are then an

[3] That is, the variables used as explanatory variables in our regressions for imputation.

[4] The ventiles are also constructed using population weights. From the HBS dataset net disposable incomes smaller than 100 € are excluded, and for SILC non–positive incomes are excluded. The reason for the difference is explained in

example of the fact that both datasets can make different inference about the population distribution of socio-demographic characteristics, income in this case, despite them being representative for the same population.

We then computed per income ventile a number of statistics for the imputed income shares of expenditures on 20 broad good categories in SILC, and for the corresponding observed income shares in HBS. The same is done for the income shares of total expenditure (the sum of income shares of those 20 categories of goods) and saving (which equals one minus the income share of total expenditure). The results of these computations can be found in the `ventile tables and graphs XX` tab of the summary file of each country (`XX` stands for the country code, and the legend of the country codes can be found in Appendix I).[5] The tables on the left contain the results for SILC, those on the right apply to the HBS. More specifically, the following statistics are computed:

- weighted mean, minimum and maximum values of income shares of expenditures, overall and per income ventile,
- overall and per income ventile, the 5th, 25th, 50th, 75th and 90th population percentiles of income shares,
- weighted mean household disposable income overall and per income ventile, and
- an alternative mean income share, calculated as population total expenditure population (overall or per ventile) divided by total income (overall or per ventile), denoted by `mean2` in the sheet, and therefore potentially different from the mean of the individual household shares per ventile.

To the right of each couple of tables a graph is produced which plots observed mean income shares against mean income per ventile according to the HBS data (red circles), and imputed income shares against mean income per ventile according to the SILC data (blue triangles). An example of such a graph is given in Figure 1.

Ideally, both plots should coincide for a perfect imputation. Deviations can however occur for two reasons: either because imputed values for given income differ from observed for the same income (which is low quality of imputation), or because estimated mean incomes differ (the estimated mean income share is evaluated at a different income level). The latter has to do with differences in inference about the population income distribution from both datasets. As a crude measure of disentangling what is due to differences in inference on the population income distribution and what is due to poor imputation quality, we linearly interpolated both series. The quality of the imputation can then be assessed by the degree to which both interpolated lines coincide.[6]

Section 5.3.

[5] See Section 6 and Appendix II for a description of the information on the imputation and the evaluation of the imputation that we render available as an *on line* addendum to this report.

[6] A *caveat* to be made is that even if the linear interpolation is a good way to fit the relation of the income shares with incomes, vertical distances between both lines might still be due not to the imputation having poor quality but to other socio–demographic characteristics being differently distributed in both SILC and HBS.

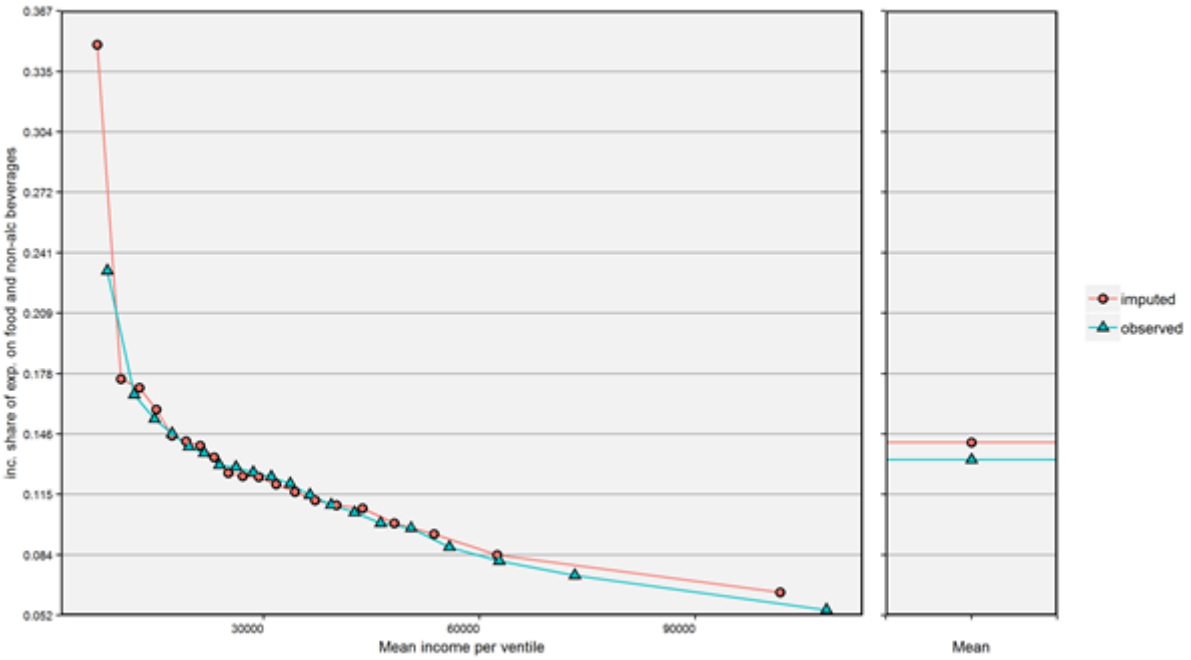

Figure 1: Example: mean imputed (SILC) and observed (HBS) income share of *food* against mean income per ventile – Germany

While discussing these graphs we will concentrate on the following three factors.

1) *The imputation's performance with respect to preserving the observed relation between net household income and income share of expenditures on a certain category.*

   This performance can be evaluated by comparing the general patterns of the red and blue plots, disregarding their volatility. If the plots display the same ascending, descending or stable patterns around the same income ventiles, we can infer that the relationship between the two variables is preserved successfully in the imputed data.

2) *Similarity of the expectations of observed and imputed income shares of expenditures on a category, conditional on income.*

   This similarity can be evaluated by looking at the vertical distance between the means per ventile of the imputed shares (read circles) and the interpolation line (in blue) between the HBS observed mean income shares. The smaller the distance is, the more similar mean income shares conditional on income are.[7]

   When inspecting these graphs, the reader should keep in mind that the graphs are scaled to the range of the variable on the vertical axis. When this range is small, the vertical distances between the two plots might look bigger than in a graph of a category with a larger range.

[7] Assuming linear interpolation is a good way to estimate conditional means. An alternative would be to produce non-parametric regressions of income shares on income for both SILC and HBS and compare their vertical distance.

Therefore, it might be misleading to evaluate imputations by comparing these distances visually. It might also be misleading to compare absolute measures of these distances, since a difference of one percentage point between the observed and imputed mean income shares of expenditure can be quite acceptable when the observed mean share is 15%, whereas it would be more alarming if this value is instead 1%.

The second important factor to account for is the volatility of the blue plot. If the values of observed means per ventile are much more volatile for one category than for another one, it is only natural to expect higher vertical distances between the plots for the former category. That does not necessarily mean that the former category is imputed less successfully. Instead, it implies that the relation between household income and expenditure on this category is less strong, and thus there is more room for randomness.

3) *Unbiasedness of the imputation.*

This can be evaluated by examining whether the mean shares of imputations per ventile are consistently higher or lower than the mean shares of observations per ventile. Over– or under–imputation of a category, either for the overall income distribution or for certain parts of the income distribution indicate bad quality of the imputation.

3.2 Difference in correlation matrices

One property of the multivariate population distribution $f(\mathbf{x}, \mathbf{w})$ is its correlation matrix. Again we can make inference on this correlation matrix from a sample by calculating weighted sample correlations between each couple of variables in the vector $(\mathbf{x}, \mathbf{w})$. We do this for both, the SILC with imputed values and the HBS (again, instead of income shares at the most detailed level, we use income shares of the 20 broad categories), say $\widehat{C}_{\text{SILC}}(\mathbf{x}, \mathbf{w})$ and $\widehat{C}_{\text{HBS}}(\mathbf{x}, \mathbf{w})$. The absolute value of the entries of the matrix obtained by taking the difference between $\widehat{C}_{\text{SILC}}(\mathbf{x}, \mathbf{w})$ and $\widehat{C}_{\text{HBS}}(\mathbf{x}, \mathbf{w})$ is equal to:

$$\Delta C(\mathbf{x}, \mathbf{w}) = \text{abs}\left[\widehat{C}_{\text{SILC}}(\mathbf{x}, \mathbf{w}) - \widehat{C}_{\text{HBS}}(\mathbf{x}, \mathbf{w})\right], \tag{12}$$

where $\text{abs}[\mathbf{A}]$ indicates the operator which takes the absolute value of the entries of the matrix $\mathbf{A}$.

If both HBS and SILC allow to make inference for the same population, this matrix should be 'close to' zero.

Now, let X be the number of socio–demographic characteristics in the vector $\mathbf{x}$ characteristics and let Y be the number of broad categories of expenditure, then the matrix $\Delta C(\mathbf{x}, \mathbf{w})$ can be partitioned as follows:

1) the $X \times X$ sub–matrix on the upper left corner, which compares the correlations within the common socio–demographic characteristics in HBS and SILC datasets,

2) the $X \times Y$ sub–matrix on the upper right corner, which compares the correlations between the common socio–demographic characteristics and income shares of expenditures on different categories in HBS and SILC datasets, and

3) the $Y \times Y$ sub–matrix on the lower right corner, which compares the correlations within the income shares of expenditures on different categories in HBS and SILC datasets.

The first sub–matrix does not signal anything regarding the performance of the imputation. Instead, it indicates whether inference on the socio–demographic household characteristics contained in the vector $\mathbf{x}$ of the population, is similar if based on the HBS or on SILC. If there are substantial differences between both (as there are sometimes) then this might cause differences in the second and third sub–matrices too, which do not indicate bad quality of the imputation *per se*.

The second sub–matrix gives insight on how well the relation between household characteristics and income shares of expenditure is preserved in the imputed dataset. If these values are substantially lower for one country than an other, it means that, if the same regression model estimated on the HBS data were to redone on the SILC data augmented with imputed values for income shares of expenditures, the two resulting model estimations would be more similar for the former country than for the latter.

Finally, the third sub–matrix gives insight on how well the relation among income shares of expenditures on broad categories is preserved in the imputed dataset.

We will thus evaluate the imputation on the basis of differences in the correlation matrices as follows. First mean values are calculated of the entries of each of the three sub–matrices. Then we rank the countries in terms of the magnitude of that mean for the $X \times X$–matrix. Larger values mean that it is less likely that both datasets stem from the same population (or are representative for the same population). An imputation is then said to be of relatively worse quality if the means for the other two sub–matrices, $X \times Y$ and $Y \times Y$ are relatively high when compared to countries with similar values for the mean of the first sub–matrix (see Section 6.3).

We report the $\Delta C(\mathbf{x}, \mathbf{w})$–matrix in the `correlation differences XX` tab of the summary files of the imputation for each country. Cells are shaded more darkly, the higher their values, that is the larger the absolute value of the the difference of the corresponding correlations inferred from HBS as compared to SILC. This visual aid helps us making a quick comparison between the tables that belong to different countries. On the top left corner of this sheet, one can find the mean values of the three sub–matrices discussed previously.

4 Data

The objective of the data collection was to obtain the most recent versions of the EUROSTAT national household budget surveys (HBS) and to impute expenditure information from those data into the EUROMOD version of EU–SILC data of the closest corresponding year. The currently most recently HBS's refer to the year 2010. More information can be found on the EUROSTAT website on these budget surveys (`https://ec.europa.eu/eurostat/web/microdata/household-budget-survey`). We were able to obtain permission to use all the 26 available surveys for that year. For 22 countries we could obtain the EU–SILC of the corresponding year (2010), and for another two countries we obtained the EU–SILC of 2012. The reason for looking after SILC–datasets of the corresponding years, is that HBS and EU–SILC both cover the same populations (representative samples for the private, non–collective households), a property which we heavily relied upon when developing our tools to evaluate the imputation (see Section 3).

In this section, we describe first the data harmonisation process that SILC and HBS micro–datasets are subjected to prior to imputation. Then, we explain our criteria to select 18 countries for which we performed the imputation and developed the new ITT. Next, we describe the process of the imputation of income for the Italian HBS micro–data which lacks income values.

Finally, we give a detailed report on the expenditure data in the HBS's. Several inconsistencies were found in these data. We discuss how we tried to resolve these issues. We then briefly discuss the descriptive statistics of the expenditure variables resulting from our extensive data processing work that was necessary in order to make the data ready for imputation.

4.1 Data preparation

Prior to imputation, the two datasets EU–SILC and HBS were harmonised for each country. Variable names and units as well as the scope and detail of information available, differ. Therefore, the two datasets are initially subjected to a harmonisation process. For this harmonisation, we follow the work procedures proposed by D'Orazio et al (2006) and advocated by EUROSTAT (Leulescu & Agafitei, 2013; Lamarche, 2017; Serafino & Tonkin, 2017).

During this process, the common variables on household characteristics are identified and their names and units are standardised. Where necessary, information is converted from the individual to the household level for both datasets. Following the harmonisation process, the common variables of the SILC and HBS micro–data are listed and their distributions are compared for each country (detailed information on this comparison can found in Deliverable 1 of this project, see Akoğuz *et al.*, 2019).

4.1.1 Harmonisation of the definition of units

Both datasets, HBS and SILC, contain information at the individual level (such as age, education level, gender etc.) and at the household level (such as household type, disposable income, the region the household resides in). The definitions of 'household' in both datasets draw on EUROSTAT's household concept.[8] Thus, there is no need for harmonisation of units. However, since the expenditure information in HBS is recorded, and thus will be imputed, at the household level, the information at the individual level need to be transformed into household level too.

We convert individual characteristics into household characteristics. When a person has a certain individual characteristic (age, say), the household she belongs to is characterised by the presence of a person with that characteristic. For instance, one can record the number of individuals in the household that fall under a certain group (such as the number of male household members, the number of household members with higher education and so on).

For the countries not belonging to the euro area, monetary variables in EU–SILC were converted into Euro, as the corresponding values in the HBS were reported in Euro. The exchange rates are stored in the tab `Parameters` of the file `country parameters.xlsx` that is delivered together with the imputation code. These exchange rates are downloaded from the annual Euro/ECU exchange rates database of EUROSTAT (`https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=ert_bil_eur_a&lang=en`). For Lithuanian litas the conversion rate pegged as of May 1st of 2004 has been used: 1 ltl = 0.28962 euro (since January 1st 2015, Lithuania entered into the euro area).

4.1.2 Harmonisation of reference period

While the HBS data officially refer to 2010, not all countries did collect a HBS in that year. EUROSTAT allows the national statistical offices of each country to deliver household budget survey data collected at most two years earlier than the reference year 2010. In addition, it was allowed to all countries to increase sample sizes of the delivered data by merging budget surveys of two preceding years. Information on how data were processed to be representative for the 2010 is scarce.[9]

The reference period for the income variables in the 2010 SILC datasets is 2009. Therefore, the income levels in the SILC 2010 datasets were adjusted using 2010 inflation rates. You can find these inflation rates in the tab `Parameters` of the file `country parameters.xlsx` that is delivered together with the imputation code. These inflation rates are downloaded from the HICP–inflation rates

[8] According to the glossary of *EUROSTAT Statistics Explained*, this concept corresponds to 'either one person living alone or a group of people, not necessarily related, living at the same address with common housekeeping, i.e. sharing at least one meal per day or sharing a living or sitting room (`https://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Household_-_social_statistics`).

[9] For more information, see also the "Quality report of the 'Household Budget Surveys' 2010" downloadable from `https://ec.europa.eu/eurostat/web/household-budget-surveys/publications`.

database of EUROSTAT (`https://ec.europa.eu/eurostat/tgm/table.do?tab=table&init=1&language=en&pcode=tec00118&plugin=1`).

For Denmark, the closest year to 2010 for which a SILC dataset is available, is 2012. For that country, the incomes of SILC were rescaled such that the mean income of the SILC data equals that of the HBS of 2010.

4.1.3 Adjustment for missing data

Countries for which the data lack important variables for the performance of imputation, are preferably excluded from the project. One exception to this rule is Italy. Income values in HBS data are missing for that country. But Italy is one of the four countries for which imputation needed to be accomplished according to the technical offer of the project. Since income is one of the key variables for which we wanted to preserve correlation with expenditure patterns as much as possible, it is not possible to perform the imputation for Italy as such. An external data source that includes income information was used to impute missing income variables in HBS prior to performing the imputation of expenditure variables from HBS to EU–SILC. A description of this income imputation process is given below in Section 4.3.

4.1.4 Harmonisation of existing variables and derivation of new variables

Harmonisation of existing variables and derivation of new variables are the most vital steps in preparing the datasets for imputation. In order to achieve that, we first identified the common variables in both datasets.

1) Common variables at the household level.

- The region the households resides in: determined with respect to NUTS1 code.
- Disposable income.

  In SILC, total disposable household income is computed as follows.
  Gross personal incomes of all household members are added together. Components of gross personal income are:

  - gross employee cash or near cash income,
  - company car,
  - gross cash benefits or losses from self-employment (including royalties),
  - pensions received from individual private plans (other than those covered under ESSPROS),
  - unemployment benefits,
  - old–age benefits,
  - survivor' benefits,
  - sickness benefits,

  - disability benefits, and
  - education–related allowances.

  To this sum, gross income components at household level are added. These consist of:

  - income from rental of a property or land,
  - family/children related allowances,
  - social exclusion not elsewhere classified,
  - housing allowances,
  - regular inter-household cash transfers received,
  - interests, dividends, profit from capital investments in unincorporated business, and
  - income received by people aged under 16.

  From that the following amounts are subtracted:

  - regular taxes on wealth,
  - regular inter–household cash transfer paid,
  - tax on income,
  - the tax adjustments–repayment/receipt received or paid during the income reference period, and
  - the social insurance contributions paid during the income reference period.

  For the HBS, no information on the construction and components of the disposable income variable is available (EUROSTAT, 2012). We only know that it concerns monetary income. Therefore, we subtract from the SILC disposable income variable (`yds`) the fringe benefits for company cars (`kfbcc`). We have no variables on other non–monetary income possibly included in the SILC disposable income concept. We proceeded by assuming that the thus corrected concept of disposable income in SILC (`yds-kfbcc`) coincides with that of the HBS. Whether this is a reasonable assumption or not is determined for each country, separately, by comparing the distributions of disposable income inferred from both datasets. These comparisons were reported and discussed in Section 5.3 of Deliverable 1 of this project (Akoğuz *et al.*, 2019).

2) Common Variables at the individual level.

- Citizenship: national, other EU, non–EU.
- Gender: male, female.
- Education level (completed).
- Current education level.

  The two last variables are determined with respect to the ISCED1 code:

  a. less than primary education,
  b. primary education,
  c. lower secondary education,

   d. upper secondary education,
   e. post–secondary non–tertiary education,
   f. tertiary education.

- Employment status:

   a. manual worker except agriculture,
   b. worker except agriculture,
   c. self–employed person or agricultural worker ,
   d. unemployed,
   e. retired,
   f. other inactive,
   g. not applicable (legal age to work not attained).

- Age.

  The SILC dataset contains the exact age of each individual whereas the HBS dataset contains only the intervals in which the age of the individual falls. These intervals are defined as follows:

   a. between 0 and 14,
   b. between 15 and 29,
   c. between 30 and 44,
   d. between 45 and 59,
   e. 60 or more.

  Another set of age–related variables that are available in the HBS dataset are at the household level. They are as follows:

   a. number of household members between the ages of 0 and 4,
   b. number of household members between the ages of 5 and 13,
   c. number of household members between the ages of 14 and 15,
   d. number of household members between the ages of 16 and 24,
   e. number of household members between the ages of 16 and 24 who are students,
   f. number of household members between the ages of 25 and 64,
   g. number of household members between who are 65 years old or older.

  However, values of this second set of variables are not always available for all households in all countries. Therefore, in order to maintain as much as possible the same regression model across countries, we will use the first set of age–related variables.

The variables above are harmonised and used for deriving new common variables at household level. The final set of common variables that will be used for the imputation is as follows:

- disposable household income (EUR/year),
- the region the households resides in (NUTS1 level),

- whether the reference person is a farmer or not,[10]
- number of male household(HH) members over the age of 14,
- number of HH members under the age of 14,
- number of HH members between the ages of 15 and 29,
- number of HH members between the ages of 30 and 44,
- number of HH members between the ages of 45 and 59,
- number of HH members over the age of 60,
- number of disabled HH members,
- number of employed HH members,
- number of unemployed HH members,
- number of non–EU citizen HH members,
- number of pensioner HH members,
- number of student HH members over the age of 14,
- number of HH members with higher education.

A set of descriptive statistics of these variables for both the HBS and SILC datasets of each country, is contained in the sheet `descriptive statistics XX` of the file `summary XX.xlsx` (see Appendix II for the complete content of these `Excel` files) where `XX` stands for the country code (see Appendix I for the country codes).

4.2 Data evaluation and country selection

We committed to develop the ITTv3 for 18 EUROMOD countries. 18 countries among 28 are needed to be chosen to perform imputations with. Four of these countries, namely Germany, Italy, France and Spain, were already agreed upon in the technical offer of the project. To select the remaining 14 countries we used the following criteria:

- availability of, and ease of access to, both SILC and HBS datasets,
- availability of common variables in HBS and SILC datasets, and the overlap between the distributions of these variables.

For Austria and the Netherlands, 2010 HBS micro–data are not available. We were not able to gain permission to access the SILC micro–data of Luxembourg and the UK. For Denmark and Croatia, we could not obtain the 2010 SILC micro–data, but we have gained access to their SILC datasets of

[10] This is considered to be an indicator that is concerns an agricultural household.

2012. When imputing expenditure information into SILC–data from an HBS dataset of another year we rescaled incomes of the SILC–data (see section 4.1.2).

For some countries for which we had both, SILC and HBS data of 2010, some variables crucial for the imputation where absent in one or both datasets. Croatia, Malta, Italy and Sweden are examples to such countries. We therefore did not select those countries, except for Italy (for which imputation was required by the technical offer).

Availability of datasets and common variables are presented for all 28 countries in Table 1. After the eliminations of countries for which data are not complete or not available, we are left with 17 countries from which we need to choose 14. Both HBS and SILC datasets claim to be representative of the same population. Therefore, we expect them to perform similarly when it comes to making inference on the population distribution of common variables. In Section 5.3 of Deliverable 1 of this project, we reported in detail on the comparison of these distributions (Akoğuz *et al.*, 2019). On the basis of this comparison, we decided not to select Bulgaria, Estonia, and Latvia from remaining list. The final list of 18 countries for which we performed imputation, and develop the new ITTv3, is indicated in the last column of Table 1.

4.3 Imputation of missing income information in Italian HBS data

As previously discussed, net household income is missing in the 2010 Italian HBS data. Since income is one of the key variables for which we wanted to preserve correlation with expenditure patterns as much as possible, it is not possible to perform the imputation for Italy as such. Therefore, we first impute these missing values in HBS by means of an external data source, and then perform the imputation of expenditures from HBS to EU–SILC.

The dataset we use to impute incomes into the HBS is the 2010 Survey on Household Income and Wealth (SHIW). SHIW is published regularly by the Bank of Italy since 1960s. The 2010 survey comprises 7,951 households (19,836 individuals), distributed over about 300 Italian municipalities. This data represents the same population and reference year as the 2010 Italian HBS data, but unlike HBS, it contains income information.

Prior to the imputation, the two datasets need to be harmonised as the names and units of the variables they contain, as well as the scope and detail of the information these variables provide, differ. In the following section, the two datasets are subjected to the same harmonisation process which was described earlier in Section 4.1.

4.3.1 Harmonisation of SHIW and HBS datasets

In this section, the common household characteristics are identified and their names and units are standardised. Afterwards, the information unit is converted from individual to household level for

Table 1: Availability of data

Country	HBS available	SILC available	All variables available	Evaluated	Selected
Austria	—	✓	inapplicable	—	—
Belgium	✓	✓	✓	✓	✓
Bulgaria	✓	✓	✓	✓	—
Cyprus	✓	✓	✓	✓	✓
Czech Republic	✓	✓	✓	✓	✓
Denmark	✓	✓ (2012)	✓	✓	✓
Germany	✓	✓	✓	✓	✓
Estonia	✓	✓	✓	✓	—
Greece	✓	✓	✓	✓	✓
Spain	✓	✓	✓	✓	✓
Finland	✓	✓	✓	✓	✓
France	✓	✓	✓	✓	✓
Croatia	✓	✓ (2012)	—	—	—
Hungary	✓	✓	✓	✓	✓
Ireland	✓	✓	✓	✓	✓
Italy	✓	✓	—	✓	✓
Lithuania	✓	✓	✓	✓	✓
Luxembourg	✓	—	inapplicable	—	—
Latvia	✓	✓	✓	✓	—
Malta	✓	✓	—	—	—
The Netherlands	—	✓	inapplicable	—	—
Poland	✓	✓	✓	✓	✓
Portugal	✓	✓	✓	✓	✓
Romania	✓	✓	✓	✓	✓
Sweden	✓	✓	—	—	—
Slovenia	✓	✓	✓	✓	✓
Slovakia	✓	✓	✓	✓	✓
The United Kingdom	✓	—	inapplicable	—	—

both datasets such that each row contains the collective information of one household. Finally, the overlap of the distributions of the common variables of the SHIW and HBS micro–data are discussed.

4.3.1.1 Harmonisation of the definition of units

The definition of a household used in both datasets are similar enough to assume that they correspond to the same concept.[11] The SHIW and HBS datasets contain both information at the individual (such as age, education level, gender) and household level (such as residency ownership status, total household expenditure, the region the household resides in). Since the disposable income variable which we will later use for the imputation of expenditures from HBS to EU–SILC, is defined at the household level, income will be imputed from SHIW to HBS at the household level as well. For that reason, the information at the individual level will be transformed into information at the household level in a similar way as we did for the harmonisation of HBS and EU–SILC datasets (see Section 4.1.1).

4.3.1.2 Harmonisation of reference period

Taking into account the *caveats* we made concerning the HBS's reference period in Section 4.1.2, reference periods for both datasets can be assumed to be 2010.

4.3.1.3 Adjustment for missing data

The information available on citizenship status, marriage status, work type (part time/full time) and the primary income source of the household could not be utilised due to too many observations with unknown values in either one of the datasets. The activity status (self–employed, employee, employer, unemployed, pensioner, student, public servant, *etc.*) could also not be utilised in full detail for the same reason. However, the common variables in SHIW and HBS datasets which contain no missing values are sufficiently rich for performing the imputation of disposable household income.

4.3.1.4 Harmonisation of existing variables, classifications and derivation of new variables

The common variables in both datasets are the following.

1) Common variables at the household level.
   - The region the households resides in: determined with respect to NUTS1 code.
   - Residence ownership: whether the household owns the residence they live in/lives free of charge.

[11] According to the website of the Bank of Italy `Survey on Household Income and Wealth`, the term household as used in the survey refers to all persons that normally reside in the same dwelling on 31 December of the year to which the survey refers and that contributed at least part of their income to the household. It also includes any members temporarily absent and any non-relatives living permanently in the home at 31 December of the reference year.

  - Total annual household expenditure.

2) Common variables at the individual level.
   - Gender: female, male.
   - Education level (completed). Determined by ISCED1 classification:
     a. less than primary education,
     b. primary education
     c. lower secondary education
     d. upper secondary and post–secondary non–tertiary education,
     e. tertiary education.
   - Current activity status:
     a. employed,
     b. unemployed,
     c. retired,
     d. pupil, student, further training, unpaid work experience,
     e. fulfilling domestic tasks,
     f. permanently disabled,
     g. in compulsory military or community service,
     h. not applicable (legal age to work unfulfilled).
   - Age.

     The SHIW dataset contains the exact age of each individual whereas the HBS dataset contains only the intervals in which the age of the individual falls. These intervals are defined as follows:
     a. between 0 and 14,
     b. between 15 and 29,
     c. between 30 and 44,
     d. between 45 and 59,
     e. 60 or more.

The variables above are harmonised and used for deriving new common variables at the household level. Within this process, the least straightforward step was the harmonisation of total annual household expenditure variables. Therefore, this step is described in detail below.

In SHIW, total annual household expenditure variable is composed of:

- fringe benefits,
- imputed rents,
- the difference between the total values of transport equipments bought and sold annually (such as cars, motorcycles, caravans, motor boats, boats, bicycles),
- total annual expenditure on furniture, household appliances and technological equipment (furniture, furnishings, carpets, lamps, small household appliances, durable household appliances, TV, PC, radio, video–recorder, CD player, mobile phone, fax machine, camera, *etc.*),

- average total monthly expenditure of the household multiplied by 12, *except* for the expenditures on the following items:
  a. valuables (jewellery, ancient or gold coins, works of art, antiques including furniture), transport equipments, maintenance & alimony, allowances, gifts,
  b. extraordinary maintenance of dwelling,
  c. mortgage instalments,
  d. life insurance premiums, and
  e. contributions to supplementary pension schemes.

In contrast, the total annual household expenditure variable in the HBS contains, among others, expenditures on jewellery and extraordinary maintenance of dwellings. Moreover, it does not contain fringe benefits and contains the total expenditure on the purchase of transport equipments rather than subtracting the value of the transport equipments sold. On the other hand,as is the case with the SHIW data, mortgage instalments, life insurance premiums and contributions to supplementary pension schemes are not treated as expenditures in HBS data, similar to the case with the SHIW data `https://ec.europa.eu/eurostat/cache/metadata/Annexes/hbs_esms_an1.pdf`. Overall, the content of these two variables is different and need to be harmonised.

In order to harmonise the definition of these two expenditure variables, the corresponding variable in SHIW is modified in the following way by means of other available variables in the dataset:

- the total value of the annual purchases of valuables is added,
- the total value of the annual expenditure made for the extraordinary maintenance of dwellings is added,
- the total annual income gained by selling the transport equipments which previously belonged to the household is added,
- fringe benefits are subtracted.

The following final set of common variables at the household level is selected to be used for the imputation.

- Household type:
  a. single person: composed of only the reference person,
  b. single parent: composed of the reference person and a dependent children,
  c. couple without children: composed of the reference person and his/her spouse,
  d. couple with children: composed of the reference person and his/her spouse and at least one dependent child,
  e. other: composed of a different combination of household members than the alternatives above,
- The region in which the household resides in.

- Residence ownership: whether the household owns the residence they live in/lives free of charge.
- Annual total expenditure of the household.
- Number of pensioner household members.
- Number of employed household members.
- Number of male household members over the age of 14.
- Number of household members with higher education.
- Number of household members between the age of 0 and 14.
- Number of household members between the age of 15 and 29.
- Number of household members between the age of 30 and 44.
- Number of household members between the age of 45 and 59.
- Number of household members over the age of 60.
- Are there any household member between the age of 0 and 14?
- Are there any household members between the age of 15 and 29?
- Are there any household members between the age of 30 and 44?
- Are there any household members between the age of 45 and 59?
- Are there any household members over the age of 60?

The income variable of the SHIW also needs to be harmonised with respect to the SILC income variable that is used for the matching, i.e. the SILC income variable that is already harmonised with respect to the income definition of HBS. The content of this last variable has already been defined in section 4.1.4. The content of the SHIW income variable can be found in the documentation of the SHIW survey for the year 2010. In accordance with these definitions, the following values are subtracted from the original SHIW income variable for harmonisation of the two income concepts:

- fringe benefits, and
- imputed rents.

4.3.2 Comparing distributions of the harmonised SHIW and HBS datasets

In this section, the overlap between the common variables in the harmonised SHIW and HBS datasets is discussed. Descriptive statistics of these variables are presented in detail in the sheet `descriptive statistics` of the file `IT income imputation summary.xlsx`, that we deliver jointly with this report.

The first significant difference between the two datasets is their sample size. The SHIW sample consists of 7,951 households, while the HBS contains 22,246 households.

The maximum number of household members that possess a certain characteristic (*e.g.* fall within a certain age interval, being employed or pensioner) is almost always higher for the HBS dataset. This stands to reason as it is natural to observe more outlier values when sample size increases. On the other hand, the minimum values of these variables, as well as their values at the 5th, 25th,50th,75th and 95th percentiles, are almost always close to each other for datasets. Moreover, their sample variances and means are very similar. Therefore, we can safely conclude that the overlap between the distributions of these variables is very good.

The two variables with the worst overlap are *number of household members between the age of 30 and 44* and *number of household members over the age of 60*. The difference between sample means is approximately 0.1 for both variables. Moreover, the HBS dataset contains 7 percentage point more observations with a positive number of household members between the age of 30 and 44 and 6 percentage point less observations with a positive number of household members over the age of 60.

Unfortunately, the overlap between the sample distributions of the annual total household expenditure variable is not as good as the remaining common variables. Its mean is approximately 26,670 euros for the SHIW, while it is 28,695 euros for the HBS. Its minimum value is 1,420 euros for the HBS while it is 2,760 euros for the SHIW. The HBS value of the variable exceeds the corresponding SHIW value for the first time around the 25th percentile and remains higher until well beyond the 95th percentile. The maximum value this variable takes in the HBS is approximately 40,000 euros less than the maximum value in SHIW.

The discrepancies between the two distributions might be due to several factors. First, definitions of the variables used for the harmonisation might not be exactly the same for the SHIW and the HBS. Secondly, it might be due to different methodologies that surveys use to construct annual total expenditure values. The SHIW asks households either their annual expenditure or their average monthly expenditure on a certain category. Average monthly expenditure records are later multiplied by 12 and added to other annual expenditure records to obtain annual total household expenditure. Conversely, in the HBS households have to keep detailed records of the daily expenditures over a certain period. The documentation on the 2010 HBS does not specify this period. Moreover, the expenditure variables in the HBS dataset correspond to monthly values. There is no indicator stating whether these values represent household expenditures in a given month or their average monthly expenditures. If the first case is true, then multiplying these values by 12, which is how we calculated annual total household expenditures for the HBS, does not lead to accurate estimations. This might not only lead to certain biases if a disproportionate number of households are observed during a certain time of the year (due to seasonality of consumption patterns on certain products) but also increase the variance of the annual total expenditure variable. Indeed, we observe that the standard deviation of this variable is almost 1,700 euros higher for the HBS as compared to the SHIW. Finally, it is important to note that significant inconsistencies are spotted in the expenditure records of the

HBS data which brings the reliability of these observations into question. This issue is discussed in detail in Section 4.4.1.

4.3.3 Imputation Method

The Predictive Mean Matching (PMM) method, which was described in Section 2.2.3, is used for performing the imputation. A linear regression model is estimated on the SHIW data where the annual net household income is the dependent variable and the common variables of the two harmonised datasets are the covariates. By means of this estimates, income values are fitted for the households in both the SHIW and the HBS. Then, each household in the HBS is matched the household in the SHIW with the closest fitted income value. Finally, the *observed* income of the matched SHIW household serves as the imputed value of income for the corresponding HBS household.

4.3.4 Evaluation of the imputation

The extent to which common variables are capable of estimating net household income is an important factor in determining the quality of imputation. Therefore, we start by examining the regression model used for fitting income values of HBS and SHIW households. This models is summarised in detail in the sheet `regression results` of the file `IT income imputation summary.xlsx`. The coefficients of the following covariates are statistically significantly estimated at the 1% level:

- region: SUD,
- region: ISOLE,
- region: CENTRO (IT),
- total expenditure,
- HH type: single individual,
- number of pensioner household members,
- number of employed household members,
- number of male household members over the age of 14,
- number of household members with higher education degree,
- number of household members between the age of 15 and 29, and
- there is a household member over the age of 60.

Additionally, residency ownership is significant at the 10% level.

Another output we use for evaluation is the correlation difference matrix. This matrix displays the absolute difference between the sample correlations of any two common variables in SHIW and imputed HBS data. It is presented in the sheet `correlation differences` of the file `IT income imputation summary.xlsx`. The correlations between net household income and common variables seem to be preserved fairly well during the imputation.

As a third evaluation output, we construct and compare non-parametric regressions of the observed and imputed income variables on total household expenditure. This plot is presented in the sheet `non-par. reg. of incomes` of the file `IT income imputation summary.xlsx`. It can be seen that the imputed income values tend to be slightly higher than the corresponding observed values for the households with comparatively lower total expenditures. This trend reverses around the annual total expenditure value of 20,000 euros and from then on, imputed income values tend to be slightly higher than the observed ones.

Our final evaluation outputs are the two QQ plots comparing the unconditional distribution of the imputed HBS income to that of the observed SILC and SHIW incomes. Distributions of imputed HBS income and observed SHIW income seem to be highly similar to each other, while the income quantiles of the SILC are higher than those of the imputed HBS income distribution. One possible explanation to this is that income definitions of the two datasets could not be fully harmonised despite our efforts. These QQ plots can be found in the sheet `QQ plots` of the file `IT income imputation summary.xlsx`.

4.4 HBS expenditure data

While preparing the HBS expenditure data for the imputation, several problems and inconsistencies in these variables were detected. In this section we explain in detail the problems we discovered, on how we processed the data in order to make them suitable for being imputed.

4.4.1 Data issues

4.4.1.1 Not every broad expenditure category is disaggregated at the same level

Expenditures in the HBS data are recorded at four different levels of aggregation. The categories at the first level are the most aggregated ones, while the categories at the fourth level are the least aggregated. An example of this structure is provided below:

- Level 1: Food and non–alcoholic beverages (EUR_HE01)
- Level 2: Food (EUR_HE011)
- Level 3: Bread and cereals (EUR_HE0111)
- Level 4: Rice (EUR_HE01111)

For some categories, however, expenditures are not disaggregated until the fourth level. In most of such cases, the dataset contains a single category at the fourth level which carries the same name and value as the expenditure category at the third level of aggregation. The structure of the aggregation then looks *e.g.* as follows:

- Level 1: Transport
- Level 2: Purchase of vehicles
- Level 3: **Motor-cycles**
- Level 4: **Motor-cycles**

However, we observe three exceptions to this structure. For the categories presented below, there are no data available for the aggregation levels denoted with NA.

- Level 1: Health (EUR_HE06)
- Level 2: Hospital Services (EUR_HE063)
- Level 3: **NA**
- Level 4: **NA**

- Level 1: Recreation and culture (EUR_HE09)
- Level 2: Package holidays (EUR_HE096)
- Level 3: **NA**
- Level 4: **NA**

- Level 1: Miscellaneous goods and services (EUR_HE12)
- Level 2: Other services (EUR_HE127)
- Level 3: Other services (EUR_HE1271)
- Level 4: **NA**

This holds for all countries.

4.4.1.2 Inconsistencies between data at different aggregation levels

We started our work on the data assuming that the four levels of aggregation were consistent in the following way. Assume that X1 denotes expenditure on a certain broad category and is divided into two sub-categories, namely X11 and X12. Assume further that X11 is disaggregated into three different sub-categories, namely X111, X112 and X113, while X12 is disaggregated into four, namely, X121, X122, X123 and X124. For a dataset including such variables to be consistent, the equations below should hold:

$$X1 = X11 + X12 \tag{13}$$

$$X11 = X111 + X112 + X113 \tag{14}$$

$$X12 = X121 + X122 + X123 + X124 \tag{15}$$

However, we find numerous examples in each country's HBS dataset for which the logic above does not hold for some households' expenditures on some categories. We observe three types of such inconsistencies.[12]

4.4.1.2.1 The expenditure on a category at the third level with only one sub–category is recorded at either the third or the fourth level, but not at both

This first case is observed in the HBS datasets of Belgium, Germany, France, Poland, and Romania. For instance, for the household with the id *388334* in the Belgian HBS data, the expenditures on 'therapeutic appliances and equipment' at the third and fourth level of aggregation (EUR_HE0613 and EUR_HE06131) are recorded as 0 € and 78.1 € respectively while they should be equal to each other. There also occur cases where the expenditure on the category at the third level of aggregation is positive while the other is zero. For instance, for the household with id *30911* in the German HBS data, the expenditures on 'postal services' at the third and fourth level of aggregation (EUR_HE0811 and EUR_HE08111) are recorded as 52.7 € and 0 € respectively. If there exists such an inconsistency between the third and fourth level of aggregation of a broad category, it is observed in all households' expenditures (except for cases where the recorded expenditures at both levels are zero and hence, trivially consistent.)

The categories that suffer from this problem per each country are the following:

- in the Belgian HBS, it is observed in the category of 'therapeutic appliances and equipment' (EUR_HE0613) under the 'health' aggregate (EUR_HE06), and in the category of 'pets and related products' (EUR_HE0934) under the 'recreation and culture' aggregate (EUR_HE09);
- in the German, French and Polish HBS datasets, it is observed in the category of 'postal services' (EUR_HE0811) under the 'communication' aggregate (EUR_HE08);
- in the Romanian HBS, it is observed in the category of 'small electrical household appliances' (EUR_HE0532) under the 'furnishings, household equipment and routine maintenance of the house' aggregate (EUR_HE05).

[12] There are a few categories for which some of the most detailed level of aggregation variables have been hidden as mentioned in the EUROSTAT HBS manual (*e.g.* expenditures on narcotics and prostitution, respectively under the broad categories 'Alcoholic beverages, tobacco and narcotics' and 'Other goods and services'). However, the aggregation rule just mentioned should not be affected, as the EUROSTAT HBS manual mentions that total expenditures on broader categories have been recalculated in such cases.

4.4.1.2.2 Total expenditure on a broader category is larger than the sum of expenditures on its sub–categories

The next problem can be illustrated by the following example stemming from the Czech data. The recorded expenditures of the household with the id *1760* on the 'restaurants and hotels' category and its sub–categories are presented below (Table 2):

Table 2: Expenditures on EUR_HE 11 Restaurants and hotels

HBS code	HBS label	Value
EUR_HE 11	**Restaurants and hotels**	358.3 €
EUR_HE 111	Catering services	286.9 €
EUR_HE 1111	Restaurants, cafés and the like	262.0 €
EUR_HE 11111	*Restaurants*	278.2 €
EUR_HE 11112	*Cafés, bars and the like*	8.7 €
EUR_HE 1112	Canteens	0.0 €
EUR_HE 11121	*Canteens*	0.0 €
EUR_HE 112	Accommodation services	71.3 €
EUR_HE 1121	Accommodation services	71.3 €
EUR_HE 11211	*Accommodation services*	71.3 €

For this household we observe an inconsistency between the third aggregation level and the rest since

$$\text{EUR\_HE11} > \text{EUR\_HE1111} + \text{EUR\_HE1112} + \text{EUR\_HE1121}, \tag{16}$$

$$\text{EUR\_HE111} > \text{EUR\_HE1111} + \text{EUR\_HE1112}, \text{ and} \tag{17}$$

$$\text{EUR\_HE1111} < \text{EUR\_HE11111} + \text{EUR\_HE11112}, \tag{18}$$

while the expenditures on remaining aggregation levels are consistent with each other.[13]

This problem is observed in the HBS datasets of Cyprus, Czech Republic, Germany, Denmark, Greece, Spain, France, Hungary, Ireland, Italy, Poland and Portugal. The number of households that suffer from this problem per each country and category are presented below.

Cyprus (sample: 2707 households)
- Clothing and footwear: 557 households,
- Food and non–alcoholic beverages: 1689 households,
- Miscellaneous goods and services: 537 households,
- Recreation and culture: 308 households,
- Restaurants and hotels: 671 households.

Czech Republic (sample: 2932 households)
- Restaurants and hotels: 858 households.

Germany (sample: 53998 households)

[13] In this report, minor differences between expenditure sums – *i.e.* less than one euro – are not considered as inconsistencies but as rounding errors.

- Alcoholic beverages & tobacco: 46160 households,
- Clothing and footwear: 33187 households,
- Education: 8182 households,
- Food and non–alcoholic beverages: 53961 households,
- Furnishings, household equipment and routine maintenance: 48270 households,
- Health: 442 households,
- Housing, water, electricity, gas and other fuels: 38565 households,
- Miscellaneous goods and services: 16482 households,
- Recreation and culture: 32219 households,
- Restaurants and hotels: 49143 households,
- Transport: 29934 households.

Denmark (sample: 2484 households)

- Health: 28 households,
- Miscellaneous goods and services: 544 households,
- Recreation and culture: 798 households.

Greece (sample: 3513 households)

- Housing, water, electricity, gas and other fuels: 3082 households,
- Miscellaneous goods and services: 348 households.

Spain (sample: 22203 households)

- Miscellaneous goods and services: 3031 households.

France (sample: 15797 households)

- Alcoholic beverages & tobacco: 115 households,
- Clothing and footwear: 89 households,
- Food and non–alcoholic beverages: 6223 households,
- Furnishings, household equipment and routine maintenance: 113 households,
- Housing, water, electricity, gas and other fuels: 15 households,
- Miscellaneous goods and services: 8830 households,
- Recreation and culture: 5765 households,
- Restaurants and hotels: 2 households,
- Transport: 1 household.

Hungary (sample: 9937 households)

- Communication: 9125 households,
- Education: 2747 households,
- Food and non–alcoholic beverages: 2593 households,
- Housing, water, electricity, gas and other fuels: 8590 households,
- Miscellaneous goods and services: 2140 households,
- Restaurants and hotels: 661 households.

Ireland (sample: 5891 households)

- Restaurants and hotels: 5111 households.

Italy (sample: 22246 households)

- Alcoholic beverages & tobacco: 12097 households,
- Communication: 21775 households,

- Education: 1376 households,
- Food and non–alcoholic beverages: 20561 households,
- Furnishings, household equipment and routine maintenance: 13378 households,
- Housing, water, electricity, gas and other fuels: 3786 households,
- Miscellaneous goods and services: 4326 households,
- Recreation and culture: 247 households,
- Restaurants and hotels: 1173 households.

Poland (sample: 37412 households)

- Food and non–alcoholic beverages: 2 households,
- Housing, water, electricity, gas and other fuels: 36219 households.

Portugal (sample: 9489 households)

- Miscellaneous goods and services: 6 households,
- Recreation and culture: 1311 households.

A household whose expenditures on a certain category are inconsistent, does not necessarily has inconsistent expenditures on other categories where inconsistencies are detected within the same dataset (or vice versa). For instance, in the HBS dataset of Cyprus, we observe that the expenditures on the broad category of 'food and non–alcoholic beverages' (EUR_HE01) of household with id *13460* are inconsistent, while its expenditures on the category of 'clothing and footwear' (EUR_HE03) are consistent. In the meantime, for the household with the id *11555* in the same dataset, we observe that the expenditures on the broad category of 'food and non–alcoholic beverages' (EUR_HE01) are consistent, while the expenditures on the broad category of 'clothing and footwear' (EUR_HE03) are inconsistent.

4.4.1.2.3 Total expenditure on a broad category is smaller than the sum of expenditures on its sub–categories

This problem is observed in the HBS datasets of Czech Republic, Germany, Greece, France, Italy, Poland and Romania. An example where this problem is spotted in the Polish HBS dataset is the expenditure of the household with id *113850511* on the category of 'food and non–alcoholic beverages' (EUR_HE01). The sum of expenditures on this category at each different aggregation level is recorded in the following way:

- Level 1: 3538.8 €
- Level 2: 3538.8 €
- Level 3: 3538.7 €
- Level 4: 3544.5 € .

The number of households that suffer from this problem per each country and category are presented below.

Czech Republic (sample: 2932 households)

- Restaurants and hotels: 744 households.

Germany (sample: 53998 households)
- Furnishings, household equipment and routine maintenance: 334 households,
- Transport: 13944 households.

Greece (sample: 3513 households)
- Miscellaneous goods and services: 348 households.

France (sample: 15797 households)
- Alcoholic beverages & tobacco: 10 households,
- Clothing and footwear: 88 households,
- Food and non–alcoholic beverages: 1153 households,
- Furnishings, household equipment and routine maintenance of the house: 30 households,
- Housing, water, electricity, gas and other fuels: 24 households
- Recreation and culture: 8 households,
- Restaurants and hotels: 1 household,
- Transport: 9 households

Italy (sample: 22246 households)
- Food and nonalcoholic beverages: 18004 households,
- Furnishings, household equipment and routine maintenance of the house: 12920 households,
- Housing, water, electricity, gas and other fuels: 1732 households,
- Miscellaneous goods and services: 1560 households.

Poland (sample: 37412 households)
- Food and non–alcoholic beverages: 4403 households.

Romania (sample: 31336 households)
- Food and non–alcoholic beverages: 293 households.

4.4.1.3 Other known issues with the expenditure data

Apart from the issues discussed above there are three country–specific issues worthy of attention.

- In the HBS dataset of Germany, expenditures of **all** households on the categories at the third and fourth disaggregation levels of the broad categories 'food and non–alcoholic beverages' (EUR_HE01) and 'alcoholic beverages & tobacco' (EUR_HE02) are recorded as zero.
- The same problem is observed for Hungary: For the categories 'communication' (EUR_HE08) and 'education' (EUR_HE10) the expenditure records at the third and fourth disaggregation levels are zero for **all** households.
- There occur records with negative expenditures on some goods or aggregates in the HBS's of Czech Republic, Denmark, Finland, France, and Slovenia.

4.4.1.4 Principles of expenditure data processing

We started by creating additional variables where needed and filling them with the values observed at the lower level of disaggregation. This provided the data a consistent structure and cured the

problem described under Section 4.4.1.1. It involved creating columns EUR_HE0631, EUR_HE06311, EUR_HE0961, EUR_HE09611, and EUR_HE12711 such that:

$$\text{EUR\_HE063} = \text{EUR\_HE0631} = \text{EUR\_HE06311}. \tag{19}$$

$$\text{EUR\_HE096} = \text{EUR\_HE0961} = \text{EUR\_HE09611}. \tag{20}$$

$$\text{EUR\_HE127} = \text{EUR\_HE1271} = \text{EUR\_HE12711}. \tag{21}$$

As our objective was to impute HBS expenditure data at the most detailed, i.e. the fourth, level of disaggregation into SILC, we took the expenditures at this level as our base. The only exception to this are the cases where *all* recorded values for a category at a certain level of disaggregation are zero. In such cases, we chose the highest level of disaggregation that contains non–zero records as our base level. This only affects Germany, Hungary and Ireland (see Section 4.4.1.3 for Germany and Hungary, and Section 4.4.1.2.2 for Ireland where the subdivision of the fourth level of disaggregation of the third level category 'catering services' (EUR_HE1111) contains only zero observations). For Germany this implies that we can only impute expenditures at the second level of disaggregation for the categories 'food and non–alcoholic beverages'(EUR_HE01) and 'alcoholic beverages & tobacco' (EUR_HE02). Similarly, for Hungary we can only impute expenditures at the second level of aggregation for the categories 'communication' (EUR_HE08) and 'education' (EUR_HE10). For Ireland, we can only impute 'catering services' (EUR_HE1111) at the third level.

For the categories at the third aggregation level which have only one subcategory, there are cases where only one of these variables contains non–zero values while the other only contains zeros (see the problem described in Section 4.4.1.2.1). We replace values at both levels with the maximum of both records. This means that if we observe a positive expenditure for that category at either of the disaggregation levels, we take this value for granted, and overrule the zero record at the other level with this value. This treats the problem discussed in Section 4.4.1.2.1.

After these operations, we took the values recorded at the highest level of disaggregation as base. This is always the fourth level of disaggregation, except for 'food and non–alcoholic beverages' (EUR_HE01) and 'alcoholic beverages & tobacco' (EUR_HE02) in Germany, for which we have only data at the highest aggregation level, 'communication' (EUR_HE08) and 'education' (EUR_HE10) for Hungary, where we only have data up to the second level, and 'catering services' (EUR_HE1111) for Ireland, where the third level is the most detailed one. We construct expenditures at a higher level of aggregation lower disaggregation levels as sums of the subcategories at a higher level of disaggregation. This ensures that we get a dataset where values are consistently aggregated, contrary to the current state of the EUROSTAT HBS data.

For example, the recorded expenditures of the household with the id *1760* in the HBS dataset of Czech Republic are updated in the following way:

EUR_HE1111	=	EUR_HE11111 + EUR_HE11112	=	**286.9 €**
EUR_HE1112	=	EUR_HE11121	=	0.0 €
EUR_HE1121	=	EUR_HE11211	=	71.3 €
EUR_HE111	=	EUR_HE1111 + EUR_HE1112	=	286.9 €
EUR_HE112	=	EUR_HE1121	=	71.3 €
EUR_HE11	=	EUR_HE111 + EUR_HE112	=	358.3 € .

So, the value of the variable EUR_HE1111 has been overwritten in this case.

Households with negative expenditures on goods or good aggregates are excluded from the imputation process. This only occurs in the HBS data of Czech Republic, Denmark, Finland, France, and Slovenia. Information on the number of observations concerned is given in cells `3Y:4Z` of the sheet `descriptive statistics XX` of the file `summary XX.xlsx` which we provide for each country `XX` (see Appendix I for the country codes).

Imputation is executed after having performed these operations on the data.

4.4.1.5 Warning on data use

An important finding of the detailed data–work undertaken for this project, is that the 2010 version of the EUROSTAT HBS has not been sufficiently validated with respect to consistency of the reported expenditures at different levels of aggregation. We therefore recommend to use these data cautiously. We are unable to say at this stage how seriously this might affect simulations with the new indirect tax tool of EUROMOD. We also suggest to contact EUROSTAT to ask them to check their data manipulation and validation code which creates the current versions of the HBS rendered available for research purposes.

4.4.2 Descriptive statistics on expenditures

After accomplishing this extensive data processing step, we present descriptive statistics for the resulting income shares and levels of household expenditure in the HBS micro–data for each category in the sheet `descriptive statistics XX` of the file `summary XX.xlsx` which we provide for each country `XX` (see Appendix I for the country codes) and summarises the imputation process.

It is common to observe some unrealistically high income shares of expenditure in the HBS datasets. Countries like Germany, France, Ireland, Italy and many more display examples of that. Sometimes these values can be so unreasonably high that it may drive up the mean income share of total expenditure considerably, as observed in the case of Germany.

Such observations mostly belong to households whose observed disposable income values are unrealistically low. In order to avoid matching SILC households with such HBS households, we decided to

exclude HBS households whose annual disposable income is less than 100 euros from the imputation process.

5 Implementing the imputation

5.1 Definition of the 20 broad categories

After having processed the HBS expenditure data in the way described in Section 4.4, we constructed 20 broad aggregated expenditure categories to which our regression model was applied. Appendix III provides a list of the 20 categories and indicates how they are constructed from the HBS data.

For each country we provide sample summary statistics on expenditures and income shares of expenditures on these broad categories, total expenditure and saving in cells `34R:60AN` of the sheet `descriptive statistics XX` of the file `summary XX.xlsx` with summary tables of the imputation which we provide for each country (`XX` stands for the country code, and the legend of country codes can be found in Appendix I).

5.2 Explanatory variables

Our basic regression model correlates income shares of expenditures on the 20 broad categories defined in the previous section, with household characteristics. Though we stress that we do not give any structural interpretation to the regression model, the selection of covariates is very much inspired by the specification of Engel curves. More specifically, a third degree polynomial in the log of incomes, and a rich set of household composition characteristics were included, containing detailed information on the number of household members in different socio–demographic groups, such as gender, labour market status, and age.

For Belgium, we also tested whether age *per se* had an additional effect by including age dummies. It turned out this was not the case. We therefore did not include the age dummies for other countries either.

A list of all potential covariates can be found in Appendix IV. A covariate is excluded from the regression models for a specific country when the information on the variable is absent in either SILC or HBS datasets, or not relevant for a particular country. Appendix IV contains more detailed information on which variables are excluded for which country. This information can also be found in the sheet `regression results XX` of the file `summary XX.xlsx`).

Note that for all 20 linear OLS regressions for positive expenditures and 20 probits for estimating the probability of positive expenditures, the covariates are the same per country.[14]

[14] In some cases some covariates did not vary across the sub–sample of observations with positive expenditures. In such a case the covariate was dropped from the continuous regression specification automatically. We did not fully document these cases.

5.3 Treatment of outliers and functional forms

The linear part of our regression model (Equation 7) is known to be sensitive to outliers. Such outliers in the dependent variable (income shares of expenditures) might occur especially in case of incomes close to zero, yielding extremely high income shares. In a first step, we removed observations from the source dataset with extremely low incomes (*i.e.* lower than or equal to 100 € per year). Still then, outliers in the income shares occur. In what follows, we treat an observation on an income share as an outlier when it surpasses 500%. In some cases, such outliers account for 90% of the sample variance of that variable. When that is the case, a linear model and estimation method (such as OLS) will try to fit the outlier values well, and will basically use no information of the bulk of the observations for which the model was constructed.

We explored three different ways to deal with the outlier problem. First we excluded them from the regression. Next, we made a log transformation of the dependent variable. In this way, outliers come closer to regular observations, eliminating the outlier problem without having to exclude observations while performing the regressions. However, this transformation itself creates its own outliers: while it brings values higher than one closer to each other, it spreads out the values between 0 and 1, and more intensely so for values close to zero. Thus, we are not guaranteed to get rid of the outliers problem with this transformation. Therefore, a third strategy consisted of taking the logarithm of the share plus one ($\ln(W_{shX}+1)$) as dependent variable. Indeed, in this case, the value inside the log function will always be greater than 1. Thus, none of the shares will be spread out. Our tests on the Belgian HBS data showed that the simple log share transformation performed best from the point of view of the resulting imputation. Therefore, we decided to do the imputation for all countries on the basis of a log share transformation of the shares as a dependent variable in the linear regression equations (7). Our final regression equations thus become:

$$\ln W_{shX} = \boldsymbol{\beta}'_X \mathbf{x}_{sh} + \varepsilon_{hX}, \qquad \text{for } W_{shX} > 0. \tag{22}$$

Notice that the outlier problem not only affects the estimates, but also the R^2 of those regressions. If a variable accounting for much of the variance is fitted well, a high R^2 will result, even though the regressions might result in a bad fit for the bulk of the observations. *Vice versa*, a bad fit for the outlier, might cause a low R^2 for a regression that fits the bulk of the observations quite well. As explained in the fourth step of the description of the imputation methodology in Section 2.3, the pseudo–R^2 of fitted expenditures on broad categories serves as a measure of how well the regression model (the linear part and the probit together) can explain the true expenditure behaviour for a given category. This measure is equally sensitive to the outlier fitting problem just explained for regular R^2's in a linear regression.[15] We therefore calculated and report the pseudo–R^2's excluding outliers.

[15] In case there are no zero observations and there would have been no log transformation of the dependent variable in the continuous linear regression (Equation 7), the R^2 of the continuous regression and the pseudo–R^2 defined in Equation (9) would coincide.

Since we used the log share version, the definition of the pseudo–R^2 has to be slightly adapted. Indeed, the concept of fitted log share is ill defined as it would amount to the estimated probability of positive expenditures times the fitted value of the log share conditional on positive expenditures, plus the estimated probability of a zero share times the log of zero, which is minus infinity. We therefore converted the fitted values of the linear regression (Equation 22), which are logs of shares, back into shares by taking an exponent. That is, $\widehat{W}_{dhX}$, defined in Equation (8), becomes:

$$\widehat{W}_{dhX} = \Phi\left(\widehat{\boldsymbol{\gamma}}'_X \mathbf{x}_{dh}\right) \exp\left(\widehat{\boldsymbol{\beta}}'_X \mathbf{x}_{dh}\right), \qquad \text{for } d = s, r. \tag{23}$$

We do realise that this is a biased estimate of the shares. Yet, we did not attempt to correct for this.

Finally, the threshold for the pseudo–R^2 above which fitted expenditures on a given category will be selected as input into the distance function, needs to be fixed. The trade–off at hand is to use more reliable estimates of expenditure behaviour (*i.e.* more closely connected with the covariates), at the cost of retaining a smaller number of aggregates into the distance function, and therefore potentially loosing information on the correlation structure of expenditures on the broad aggregate categories. We performed tests on the Belgian data, with cut–off values 0.1 (more goods pass, but the estimates are less reliable), 0.3 and 0.45. The best results were obtained with a value of 0.1. We therefore used this threshold for other countries as well. This means that in the distance function we only retained fitted values of aggregates with pseudo–R^2 above or equal to 0.1. In the sheet `regression results` of the files `summary XX.xlsx` with summary information of the imputation for each country, the pseudo–R^2 values for each of the 20 broad categories' regression models are reported. Categories that surpass the 0.1 threshold, and for which the fitted value thus enters the distance function, are highlighted in green. The same sheet also includes more detailed information of the regression results for the 20 probits and linear regressions we ran for each country.

5.4 Saving and expenditures as derived variables

In our data we observe disposable income in both source and recipient data, respectively denoted by y_{sh} for a household h in the source data and y_{rg} for a household g in the recipient data. The source dataset also contains observations on expenditures on specific detailed goods, indexed by $i, j, \ldots \in \mathcal{N}$. These expenditures for household h, denoted by e_{shi}, are non–negative and sum to total expenditures, denoted by E_{sh}. Saving, S_{sh}, is defined as the difference between disposable income and total expenditures. So, the source dataset contains information on:

$$E_{sh} \equiv \sum_{i \in \mathcal{N}} e_{shi}, \tag{24}$$

$$S_{sh} \equiv y_{sh} - E_{sh}. \tag{25}$$

We can then define shares of detailed and total expenditures in disposable income as

$$w_{shi} = \frac{e_{shi}}{y_{sh}}, \tag{26}$$

$$W_{sh} = \frac{\sum_{i \in \mathcal{N}} e_{shi}}{y_{sh}} = \frac{E_{sh}}{y_{sh}}. \tag{27}$$

Consequently, the share of saving in disposable income equals:

$$\theta_{sh} = \frac{S_{sh}}{y_{sh}} = 1 - W_{sh}. \tag{28}$$

Suppose a household g in the recipient data r is matched with household h in the source data s. For imputing expenditures in the recipient data, two alternatives are open. We could simply impute the observations on detailed expenditures in the source, e_{shi}, as the imputed values for these expenditures in the recipient:

$$e_{rgi} = e_{shi}. \tag{29}$$

However, since income is only one of several common variables used in the imputation procedure, a pair of matched households can have different levels of income. *E.g.* a household with a relatively low income in the recipient dataset, can be matched with a household with relatively high income in the source dataset. In that case, using Equation (29) attributes the presumably high expenditure levels of that household with high income to the household with a low income. Since in this case saving is computed as the residual between disposable income and total expenditures, this alternative attributes the whole difference between the disposable incomes of households h and g to imputed saving. To avoid this in our eyes unattractive feature, we opted for a different approach.

The *income shares* of detailed expenditures *and saving* of the household h in the source dataset were used to obtain imputed detailed expenditures and saving for household g in the recipient dataset by multiplying these income shares with the disposable income of household g in the recipient dataset. Then, total expenditure is obtained as the sum of imputed detailed expenditures. Formally,

$$e_{rgi} = y_{rg} w_{shi} = y_{rg} \frac{e_{shi}}{y_{sh}}, \tag{30}$$

$$S_{rg} = y_{rg} \theta_{sh} = y_{rg} \frac{S_{sh}}{y_{sh}}, \tag{31}$$

$$E_{rg} = \sum_{i \in \mathcal{N}} e_{rgi} = y_{rg} \frac{E_{sh}}{y_{sh}}. \tag{32}$$

This method distributes the difference in disposable income between a pair of matched household across all expenditures and saving by rescaling the observed values with the ratio of the disposable income of household g to that of household h as showed in equations (30), (31) and (32). A potential concern is that when a household with low income and negative saving is matched to a household with higher income, the imputed dis–saving level of the latter will be even higher.

6 Imputation results

In this section, we first present, for each country, the list of expenditure categories whose fitted values enter as inputs in the distance function. An expenditure category is selected as input in the distance function if its pseudo–R^2 value is higher than a predefined threshold value (which is 0.1 in our case). The pseudo–R^2 values of these input categories are also presented.

We then continue with an evaluation by means of the tools developed in Section 3. We first discuss the graphs of mean income shares of expenditures per income ventile, then we briefly comment upon the performance of the imputation in terms of safeguarding the correlation structures between variables. Finally, this section also discusses the comparison of estimated total expenditures and tax revenues on the basis of the imputation in SILC with similar estimates from HBS (for expenditures) and National Accounts (for expenditures and tax revenues).

For each country, we provide an Excel file `summary XX.xlsx` (where `XX` stands for the country code defined in Appendix I) which contains all outputs discussed in this report. A detailed overview of the content of these files is given in Appendix II.

6.1 Regression results

For each country, summaries of both, the probit and linear regression models, for each broad expenditure category, can be found in the sheet `regression results XX` of the file `summary XX.xlsx`.

In order to evaluate the goodness of fit of our two–step regression model as a whole, we look at the pseudo–R^2, as defined by Equations (9) in Section 2.3 and (23) in Section 5.3.

Table 3 provides a list of each broad expenditure category, the corresponding number of countries for which the fitted values of the category enter the distance function, and the average pseudo–R^2 value of its regression models conditional on exceeding the 0.1 threshold. Good categories are ordered in terms of the number of times they were used in the distance function, and then in terms of this average pseudo–R^2 value. Table 4 shows for each country the pseudo–R^2 values for the categories whose fitted values are used as input for the distance function.

The regression models for particularly three expenditure categories, namely *food and non–alcoholic beverages*, *utilities*, and *communications* seem to perform significantly better in comparison to the rest of the categories. Expenditure categories that are expected to suffer from neither the infrequent expenditures problem nor the zero expenditures problem are both much more likely to enter distance functions and have much higher average pseudo–R^2 values conditional on exceeding the 0.1 threshold. A noteworthy exception to that rule is the category *housing and rental*, which despite many zero observations (usually owner–occupiers) in all countries, enters the distance function eleven times, and exhibits overall relatively large pseudo–R^2's.

The pseudo–R^2 values for the categories that are expected to suffer the strongest from the infrequent

expenditures problem, namely *travel and holidays*, *house durables*, and *vehicles*, do not exceed the 0.1 threshold in any of the cases. This also holds for the category of *other* which gathers all expenditure variables that fall under none of the other categories. This result is once again in line with our expectations as it is not plausible to expect a strong relation between the expenditures made on such a diverse collection of goods and services, and household characteristics.

Table 3: Number of times category was selected and average pseudo–R^2 values

Categories	**Number of times selected**	**Average pseudo–R^2**
Food and non–alcoholic beverages	17	0.45
Utilities	15	0.41
Communications	14	0.36
Housing and rental	11	0.24
Personal care	8	0.22
Culture and recreation	8	0.21
Public transportation	6	0.24
Tobacco	6	0.16
Alcoholic beverages	5	0.29
Restaurants	5	0.19
Insurance	4	0.43
Private transportation	4	0.29
Health and care	4	0.13
House goods and services	3	0.16
Education	2	0.17
Clothing and personal items	1	0.15
House durables	0	–
Other	0	–
Travel and holidays	0	–
Vehicles	0	–

Table 4: Pseudo–R^2 values of good categories surpassing the 0.1 threshold (per country)

Category	BE	CY	CZ	DE	DK	EL	ES	FI	FR	HU	IE	IT	LT	PL	PT	RO	SI	SK
Food and non–alcoholic beverages	0.63	0.41	0.38	0.52	0.32	0.40	0.31	0.31		0.34	0.54	0.50	0.49	0.61	0.25	0.64	0.64	0.33
Utilities	0.56	0.58	0.35	0.17	0.29	0.52	0.31		0.51	0.60	0.57	0.31	0.16		0.35		0.63	0.27
Communications	0.40	0.23		0.24	0.29			0.17	0.43		0.47	0.42	0.32	0.27	0.21	0.45	0.51	0.64
Housing and rental	0.32	0.16	0.12	0.37	0.40	0.34		0.36			0.27				0.10	0.11		0.12
Personal care		0.69		0.19		0.23				0.13			0.15			0.14	0.10	0.14
Culture	0.19	0.32			0.12			0.15			0.16					0.14	0.51	0.10
Public transportation	0.27	0.29			0.23	0.26		0.23			0.15							
Tobacco						0.12				0.14	0.11		0.29		0.19	0.10		
Alcoholic beverages		0.27				0.20		0.45								0.20	0.36	
Restaurants	0.14					0.28			0.13		0.26							0.14
Insurance		0.57				0.20					0.56						0.40	
Private transportation		0.37				0.49					0.15					0.15		
Health and care										0.15			0.11	0.10			0.15	
House goods and services										0.12			0.12				0.24	
Education		0.22													0.11			
Clothing and personal items		0.15																
House durables																		
Other																		
Travelling																		
Vehicles																		

6.2 Ventile graphs

In this section, we evaluate the ventile graphs of mean income shares of expenditures against income for each country in the frame of the three performance aspects described in Section 3.1, *i.e.* similarity of patterns, similarity of expectations conditional on income, and unbiasedness of the imputations. To recall, these graphs plot mean imputed (SILC) and observed (HBS) income shares of expenditures on the 20 broad categories against mean disposable income per ventile. An example of such a graph is given in Figure 2.

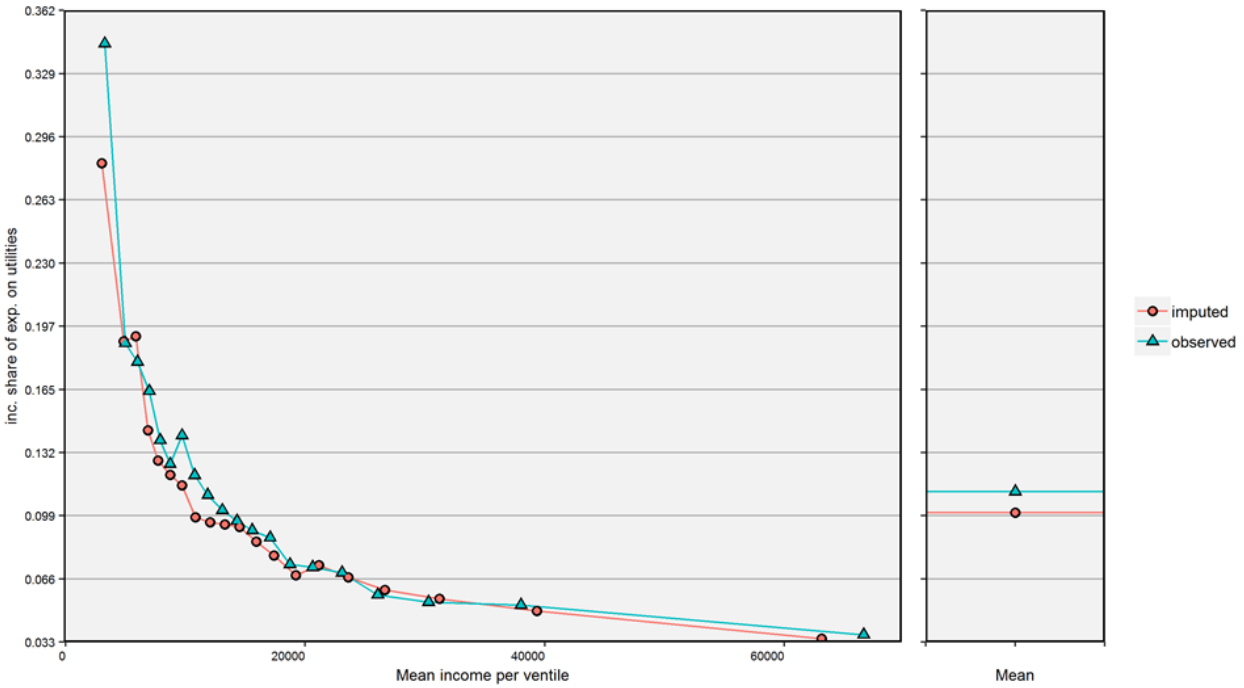

Figure 2: Example: mean imputed (SILC) and observed (HBS) income share of *utilities* against mean income per ventile – Portugal

We examine the imputation results for all broad categories, mentioning the categories that perform the best and the worst in terms of all three aforementioned aspects. We also discuss the derived imputations of total expenditure, defined as the sum of income shares of all expenditures, and saving (one minus the income share of total expenditure).

As ventile graphs demonstrate, the first income ventile might contain substantial outliers where households combine extremely low income levels with regular expenditure levels. This remains the case even when we exclude households with disposable incomes lower than 100 euro per year. These outliers may have a serious impact on ventile means, and so might easily lead to large difference in the imputed and observed means for a given ventile, depending on whether or not, or how frequently, SILC households are matched with such outliers. Therefore, while comparing imputation performances for different categories, we will not put emphasis on the first ventile. The category

other will not be taken into consideration as this variable stands for the total expenditure on a very diverse collection of goods and services.

Our general finding is that the imputations for the categories of *food and non–alcoholic beverages* and *utilities* perform by far the best across all countries. This can be explained by the fact that these expenditures do not contain many records with zero expenditures.

6.2.1 Belgium

Imputations for all categories perform well in terms of preserving the observed relation between net household income and income share of expenditures.

Imputations seem to be unbiased for most categories, however there are some exceptions. There seems to be a general upward bias in the imputations of expenditures on *tobacco* and *vehicles*. Moreover, there seems to be a downward bias in the imputation of expenditures on *education*.

The similarity between the average observed and imputed income shares of expenditure conditional on income is quite high for approximately half of the categories.

The categories with best performing imputations are *food and non–alcoholic beverages* and *utilities*. The categories with worst performing imputations are *education* and *vehicles*.

The imputation of income shares of total expenditure and saving seem to be unbiased and perform well in terms of preserving the observed relation between net household income and income share of expenditures. The similarity between the average observed and imputed income shares of total expenditure conditional on income is fairly good.

6.2.2 Cyprus

Imputations for all categories perform well in terms of preserving the observed relation between net household income and income share of expenditures.

Imputations seem to be unbiased for most categories. However, the imputations for the categories *housing and rental* and *education* seem to be downward biased in general. Additionally, the imputations for the categories *utilities* and *public transportation* seem to be downward biased for the second half of the income distribution, while that is the case with the category *communications* for the first half. Finally, the imputations for the *insurance* category seem to be downward biased in the lower end of the income distribution and upward biased in the upper end.

The similarity between the average of observed and imputed income shares of expenditure conditional on income is fairly good for approximately a quarter of the categories. The similarity is neither extremely good nor extremely bad for almost all categories.

The categories with best performing imputations are *food and non–alcoholic beverages* and *personal care*. The categories with worst performing imputations are *housing and rental*, *education* and *public transportation*.

The imputation of income shares of total expenditure and saving seem to be unbiased and perform well in terms of preserving the observed relation between net household income and income share of expenditures. The similarity between the average observed and imputed income shares of total expenditure conditional on income is fairly good.

6.2.3 Czech Republic

Imputations for almost all categories perform fairly well in terms of preserving the observed relation between net household income and income share of expenditures with the exception of the category *education* for which imputed expenditure shares for the higher end of the income distribution can not replicate the positive correlation between income and observed expenditure shares. Imputations seems to be unbiased for the majority of categories, however there are several categories for which imputations seem to be biased either generally or for a certain part of the distribution. Examples include categories like *food and non–alcoholic beverages* and *utilities* which tend to be among the best performing categories in all criteria for most of the other countries.

For most of the categories, the overlap between the average of observed and imputed income shares of expenditure conditional on income is fairly good. Yet, it is not possible to see extremely good overlaps even for highly regular expenditures such as *food and non–alcoholic beverages* or *utilities*.

The categories with best performing imputations are *food and non–alcoholic beverages* and *utilities*. The categories with worst performing imputations are *private transportation*, *vehicles*, *house durables* and *travelling and holiday*.

The imputation of income shares of total expenditure and saving seem to be unbiased and perform well in terms of preserving the observed relation between net household income and income share of expenditures. The similarity between the average of observed and imputed income shares of total expenditure conditional on income is fairly good.

6.2.4 Germany

Imputations for all categories perform extremely well in terms of preserving the observed relation between net household income and income share of expenditures. Imputations seems to be unbiased for most categories with the exceptions of *alcoholic beverages*, *restaurants*, *health and care* and *private transportation* which seem to be imputed upwards biased in general.

For the majority of the categories, the overlap between the average of observed and imputed income shares of expenditure conditional on income is extremely well.

The categories with best performing imputations are *food and non–alcoholic beverages*, *communications*, *utilities*, and *housing and rental* The categories with worst performing imputations are *vehicles* and *public transportation*.

The imputation of income shares of total expenditure and saving perform extremely well in terms of all three evaluation criteria.

6.2.5 Denmark

Imputations for all categories perform extremely well in terms of preserving the observed relation between net household income and income share of expenditures.

Imputations seem to be unbiased for the majority of categories. However, there seems to be an upward bias in general in the imputations of the expenditures on *insurance*, *house goods and services* and *public transportation*. Moreover, there seems to be an upward bias in the middle of the income distribution in the imputation of the income shares of expenditure on *food and non–alcoholic beverages*. Finally, there seems to be a downward bias around the middle of the income distribution in the imputation of the income shares of expenditure on *housing and rental*.

The overlap between observed and imputed values is mediocre for the majority of the categories. The categories with best performing imputations are *utilities*, *culture and recreation*, and *insurance*. The categories with worst performing imputations are *education*, *vehicles* and *travelling and holiday*.

The imputation of income shares of total expenditure and saving seem to be unbiased and perform well in terms of preserving the observed relation between net household income and income share of expenditures. The similarity between the average of observed and imputed income shares of total expenditure conditional on income is fairly good.

6.2.6 Greece

Imputations for all categories perform extremely well in terms of preserving the observed relation between net household income and income share of expenditures.

Imputations seem to be unbiased for almost all categories with three minor exceptions: a downward bias in the imputation of expenditures on *vehicles* for the lower end of the income distribution, an upward bias in the imputation of expenditures on *alcoholic beverages* in the first half of the income distribution, and an upward bias in the imputation of expenditures on *house goods and services* for the higher end of the income distribution.

Except for a few categories with generally worse imputation performances across countries, such as *vehicles*, the overlaps between the average of observed and imputed income shares of expenditure per ventile are generally neither extremely good nor extremely bad. One exception to this is *utilities* for which the overlap is extremely good. In general, the overlap is the best for categories with generally

better imputation performances across countries such as *communications* in comparison to categories with generally worse imputation performances across countries such as *travelling and holiday*.

The categories with best performing imputations are *utilities*, *communications* and *food and non–alcoholic beverages*. The categories with worst performing imputations are *vehicles*, *education* and *travelling and holiday*.

The imputation of income shares of total expenditure and saving seem to be unbiased and perform well in terms of preserving the observed relation between net household income and income share of expenditures. The similarity between the average of observed and imputed income shares of expenditure conditional on income is fairly good.

6.2.7 Spain

Imputations for most categories perform fairly well in terms of preserving the observed relation between net household income and income share of expenditures with the notable exceptions of the categories *culture and recreation* and *education*. While there seems to be no relation between income and income share of expenditure on *culture and recreation* according to observed income shares, imputed values exhibit a slight inverse relationship with income. As for the category of *education*, while observed income shares of expenditure indicate positive correlation between the two variables, the relation between imputed shares and income is less outspoken.

Imputations seem to be unbiased for the majority of the categories. However, there seems to be a downward bias in the imputation of incomes shares of expenditures on *housing and rental* for the first half of the income distribution, on *insurance* and *private transportation* for the second half of the income distribution, and on *tobacco* and *public transportation* in the middle of the income distribution, and, lastly, an upward imputation bias for the first half of the distribution and a downward imputation bias for the second half of the distribution (except for the last two ventiles) for the category *health and care*.

The overlaps between the average of observed and imputed income shares of expenditure per ventile are extremely well for the categories of *food and non–alcoholic beverages* and *utilities*. Moreover, there are some other categories which also display quite good overlaps such as *personal care* and *communications*. Some categories with generally worse imputation performances across countries such as *vehicles*, *house durables*, and *travel and holidays* perform considerably better in terms of the similarity between average observed and imputed incomes shares per ventile, when compared to their average performances across the 18 countries.

The categories with best performing imputations are *food and non–alcoholic beverages* and *utilities*. The categories with worst performing imputations are *education* and *vehicles*.

The imputation of income shares of total expenditure and saving are unbiased and perform well in terms of preserving the observed relation between net household income and income share of

expenditures. The similarity between the average of observed and imputed income shares of total expenditure conditional on income is fairly good.

6.2.8 Finland

Imputations for all categories perform extremely well in terms of preserving the observed relation between net household income and income share of expenditures.

Imputations seem to be unbiased for most categories. The imputations for the categories *insurance* and *travelling and holiday* are slightly upward biased in general. The imputation for the category *communications* is slightly upward biased in the second half of the income distribution.

The overlap between the average observed and imputed income shares of expenditure per ventile is quite well for many categories which on average do not perform very well in terms of such similarities for the 18 countries, such as *housing and rental*, *insurance*, *culture and recreation*, and even *house durables* Moreover, even the categories with the worst level of overlap, *i.e. travelling and holidays*, *education*, and *vehicles* perform much better compared to their average performance across the 18 countries.

The categories with best performing imputations are *food and non–alcoholic beverages* and *communications*. The category with worst performing imputation is *vehicles*.

The imputation of income shares of total expenditure and saving seem to be unbiased and perform well in terms of preserving the observed relation between net household income and income share of expenditures. The similarity between the average of observed and imputed income shares of total expenditure conditional on income is fairly good.

6.2.9 France

Imputations for all categories perform well in terms of preserving the observed relation between net household income and income share of expenditures.

The overlap between the average observed and imputed income shares of expenditure per ventile is quite good and imputations seem to be unbiased for most categories. However, there are some exceptions. There is a downward bias in the imputations of income shares of expenditures on *tobacco* and *health and care*, and an upward bias for *private transportation*. The imputations for lower income ventiles exhibit a downward bias for *house goods and services*, and *health and care*, and an upward bias for *vehicles*. Another area where imputations are upward biased is higher income ventiles for the categories *education* and *clothing and personal items*.

The categories with best performing imputations are *food and non–alcoholic beverages*, *utilities*, *communication*, and *insurance*. The category with worst performing imputation is *health and care*.

The imputation of income shares of total expenditure and saving perform extremely well in terms of all three evaluation criteria.

6.2.10 Hungary

Imputations for all categories perform extremely well in terms of preserving the observed relation between net household income and income share of expenditures.

Imputations seem to be unbiased for most categories. However, there is some general tendency of imputation being downward biased for *education*. Similarly, the imputations for the categories *communications*, *insurance*, and *clothing and personal items* are slightly downward biased in the lower part of the income distribution.

The overlap between the average observed and imputed income shares of expenditure per ventile is quite good for many categories including *health and care*, *clothing and personal items*, *alcoholic beverages*, and *tobacco*. Moreover, even the comparatively worst performing categories do not exhibit an extremely bad overlap in the case of Hungary.

The categories with best performing imputations are *food and non–alcoholic beverages* and *communications*. The category with worst performing imputation is *education*.

The imputation of income shares of total expenditure and saving perform extremely well in terms of all three evaluation criteria.

6.2.11 Ireland

Imputations for all categories perform fairly well in terms of preserving the observed relation between net household income and income share of expenditures.

Imputations seem to be unbiased for most categories. However, the imputations for the categories *public transportation* and *housing and rental* seem to be downward biased in general, while the imputations for the categories *house durables* and *education* seem to be downward biased for the lower end of the income distribution.

The overlap between the average observed and imputed income shares of expenditure per ventile is quite poor for almost half of the categories. The categories with best performing imputations are *food and non–alcoholic beverages* and *utilities*. The categories with worst performing imputations are *house durables*, *education* and *travelling and holiday*.

The imputation of income shares of total expenditure and saving seem to be unbiased and perform well in terms of preserving the observed relation between net household income and income share of expenditures. The similarity between the average observed and imputed income shares of total expenditure conditional on income is fairly good.

6.2.12 Italy

Imputations for most categories perform extremely well in terms of preserving the observed relation between net household income and income share of expenditures. Two exceptions to this are the categories of *education* and *house durables*. For the first category, while both imputed and observed values indicate a positive correlation between the income share of expenditure and household disposable income, the former one indicates a less steep increase in the income share of expenditure when income rises. Conversely, for the second category, while both imputed and observed values indicate a positive correlation between the income share of expenditure and household disposable income, the former one indicates a steeper increase in the income share of expenditure when income rises.

Imputations seem to be unbiased for the majority of the categories. However, the imputations for the categories *insurance* and *alcoholic beverages* seem to be slightly downward biased in general, while the imputation for the category of *tobacco* seem to be downward biased for the first half of the income distribution. Lastly, the imputations for the category *health and care* seem to be upward biased in general.

The overlap between the average observed and imputed income shares of expenditure per ventile is extremely good for approximately half of the categories including many categories which are expected to suffer from infrequent or zero expenditures issues such as *alcoholic beverages*, *tobacco*, *private transportation*, *clothing and personal items* and *restaurants*.

The categories with best performing imputations are *food and non–alcoholic beverages* and *utilities*. The categories with worst performing imputations are *vehicles*, *education*, and *house durables*.

The imputation of income shares of total expenditure and saving perform extremely well in terms of all three evaluation criteria.

6.2.13 Lithuania

Imputations for most categories perform fairly well in terms of preserving the observed relation between net household income and income share of expenditures with the exceptions of the categories *housing and rental* and *education*.

Imputations seem to be unbiased for the majority of the categories. However, the imputations for the categories *education* and *housing and rental* seem to be downward biased in general while the same goes for the categories *vehicles*, *utilities* and *house goods and services* – but only for the first half of the income distribution – as well as the category of *restaurants* – but only for the second half. Additionally, the imputations for the category of *house durables* seem to be biased upwards in general.

The overlap between the average of observed and imputed income shares of expenditure per ventile

is extremely poor for only few categories including *housing and rental*, *education* and *vehicles*. The overlap is fairly good for approximately half and extremely good for some categories. Categories where overlap is extremely good includes *communication*, *public transportation* and *travelling*. The overlap for *clothing and personal items* category, which is expected to suffer from infrequent or zero expenditures issues, is also extremely good.

The categories with best performing imputations are *food and non–alcoholic beverages*, *utilities*, *communications* and *alcoholic beverages*. The categories with worst performing imputations are *vehicles*, *travelling and holiday* and *housing and rental*.

The imputation of income shares of total expenditure and saving seem to be unbiased and perform well in terms of preserving the observed relation between net household income and income share of expenditures. The similarity between the average of observed and imputed income shares of total expenditure conditional on income is fairly good.

6.2.14 Poland

Imputations for most categories perform fairly well in terms of preserving the observed relation between net household income and income share of expenditures with the notable exception of *housing and rental* category. For this category, while the observed values indicate a negative correlation between the income share of expenditure and household disposable income, the imputed values indicate the opposite, *i.e.* a positive correlation which is also in contrast with our expectations.

Imputations seem to be unbiased for the majority of the categories. However, the imputations for the categories *communication*, *culture and recreation*, *personal care* and *public transportation* seem to be downward biased for the first half of the income distribution. On the other hand, the imputations for the *tobacco* category seem to be upward biased for the second half of the income distribution.

The overlap between the average of observed and imputed income shares of expenditure per ventile is very good for only two categories, for *food and non–alcoholic beverages* and for *vehicles* despite the poor imputation performance for the latter. The categories where the overlap is extremely bad are *culture*, *private transportation*, *travelling and holiday* and *education*.

The categories with best performing imputations are *food and non–alcoholic beverages*, *communication* and *health and care*. The categories with worst performing imputations are *housing and rental* and *vehicles*.

The imputation of income shares of total expenditure and saving perform well in terms of preserving the observed relation between net household income and income share of expenditures. However the imputations seem to be biased upward (downward for saving) for the second half of the income distribution. The similarity between the average of observed and imputed income shares of total expenditure conditional on income is fairly good.

6.2.15 Portugal

Imputations for most categories perform fairly well in terms of preserving the observed relation between net household income and income share of expenditures.

Imputations for half of the categories perform well in terms of preserving the observed relation between net household income and income share of expenditures. The categories where the imputations perform relatively poorly are *housing and rentals*, *house goods and services*, *culture and recreation*, *personal care*, *alcoholic beverages*, *tobacco*, *public transportation*, *travelling and holiday*, *education*, and *household durables*.

The overlap between the average of observed and imputed income shares of expenditure per ventile is very good for approximately half of the categories, while only few categories, namely *housing and rental* and *vehicles*, perform badly.

The imputations for the categories of *food and non–alcoholic beverages* and *utilities* perform best, while *culture and recreation*, *education*, *vehicles* and *household durables*.

The imputations of total expenditure and saving perform well in terms of preserving the observed relationship between net household income and income share of total expenditure. However, total expenditure (saving) seems to be under(over)-–imputed in all ventiles except one.

6.2.16 Romania

Imputed expenditures are able to preserve the observed relationship between net household income and income share of expenditure for all categories.

Most categories seem to be imputed in an unbiased manner except for the categories of *housing and rentals* and *personal care* which are under–imputed for the vast majority of income ventiles.

The overlap between the average of observed and imputed income shares of expenditure per ventile is good for approximately half of the categories, while only few categories, namely *housing and rental* and *education*, perform badly. Extremely good overlaps are observed in *food and non–alcoholic beverages*, *house goods and services*, *alcoholic beverages*, *tobacco* and *public transportation*.

The categories with best performing imputations are *food and non–alcoholic beverages* and *utilities*. The categories with worst performing imputations are *housing and rentals*, *insurance* and *vehicles*.

The imputation of income shares of total expenditure and saving seem to be unbiased and perform well in terms of preserving the observed relation between net household income and income share of expenditures. The similarity between the average observed and imputed income shares of total expenditure conditional on income is fairly good.

6.2.17 Slovenia

Imputed expenditures are able to preserve the observed relationship between net household income and income share of expenditure for all categories.

Most categories seem to be imputed in an unbiased manner. That being said, there is a slight upward bias in higher income ventiles for *public transportation* and *travelling and holiday*. We also observe a minor downward bias in lower ventile imputations for *utilities*.

For Slovenia, the overlap between the average of observed and imputed income shares of expenditure per ventile is relatively poor compared to other countries. This is mostly driven by the high expenditure shares in the lowest income ventile which our imputation method was not able to capture. Despite that, there is no category where the overlap is extremely bad and for *public transportation*, *travelling and holiday* and *restaurants* categories the overlap is still relatively good.

The categories with best performing imputations are *food and non–alcoholic beverages*, *house goods and services* and *communication*. The categories with worst performing imputations are *housing and rental*, *tobacco*, *public transportation* and *travelling and holiday*.

The imputation of income shares of total expenditure and saving seem to be unbiased and perform well in terms of preserving the observed relation between net household income and income share of expenditures. The similarity between the average observed and imputed income shares of total expenditure conditional on income is fairly good.

6.2.18 Slovakia

Imputations for most categories perform fairly well in terms of preserving the observed relation between net household income and income share of expenditures.

Most categories seem to be imputed in an unbiased manner. Few exceptions are *private transportation* for lower income ventiles (upward bias) and *education* for higher income ventiles (downward bias). *Housing and rental* is another category where we observe a relatively consistent downward bias in imputations though this bias is small in magnitude.

The overlap between the average of observed and imputed income shares of expenditure per ventile is extremely good for almost all categories. There is only one category where the overlap is extremely bad (*education*) and only two categories where the overlap could be considered as bad (*insurance* and *vehicles*).

The imputations of total expenditure and saving perform fairly well in terms of preserving the observed relationship between net household income and income share of total expenditure. However, total expenditure (saving) seems to be under(over)–imputed in the vast majority of ventiles.

6.3 Difference in correlation structure

In the sheet `correlation differences XX` of the `summary XX.xlsx` files, we report the differences in the correlation matrices as discussed in Section 3.2. We explained there how we could summarise the matrix by three figures: the mean absolute difference in correlations among socio–demographic characteristics (`within covariates`), the mean absolute difference in correlations among income shares of expenditures (`within expenditures`), and the mean absolute difference in correlations between socio–demographic characteristics and income shares of expenditures (`between covariates and expenditures`).

In the Figures 3–4, we plot the mean absolute difference of the last two, `within expenditures` and `between covariates and expenditures`, against the first one (`within covariates`). Recall that larger differences in `within covariates` might lead to larger values for the other two figures which are not to be ascribed to the imputation method *per se*, but might stem from the datasets not being representative for the same population. Therefore we say that an imputation has relatively lower quality if it exhibits relatively large values for `within expenditures` or `between covariates and expenditures` for given values of `within covariates`.

We see that this is the case for `within expenditures` against `within covariates` for Cyprus, Ireland, and to a lesser extent for Belgium, Slovenia, and Greece. They exhibit similar correlation statistics in HBS and SILC for the socio–demographic characteristics as for example Germany (for Greece), Finland (for Slovenia), and Poland (for Belgium), but are doing much worse than those countries in safeguarding the correlation structure among income shares of expenditures in the imputation.

The correlation structure between socio–demographic characteristics and income shares of expenditures seems to be much less affected by `within covariates`. Even then, Spain exhibits a larger difference in the correlation pattern between covariates and income shares of expenditures of HBS and SILC than for example Cyprus, which has similar correlation deviations between the covariates in SILC and HBS as compared to Spain.

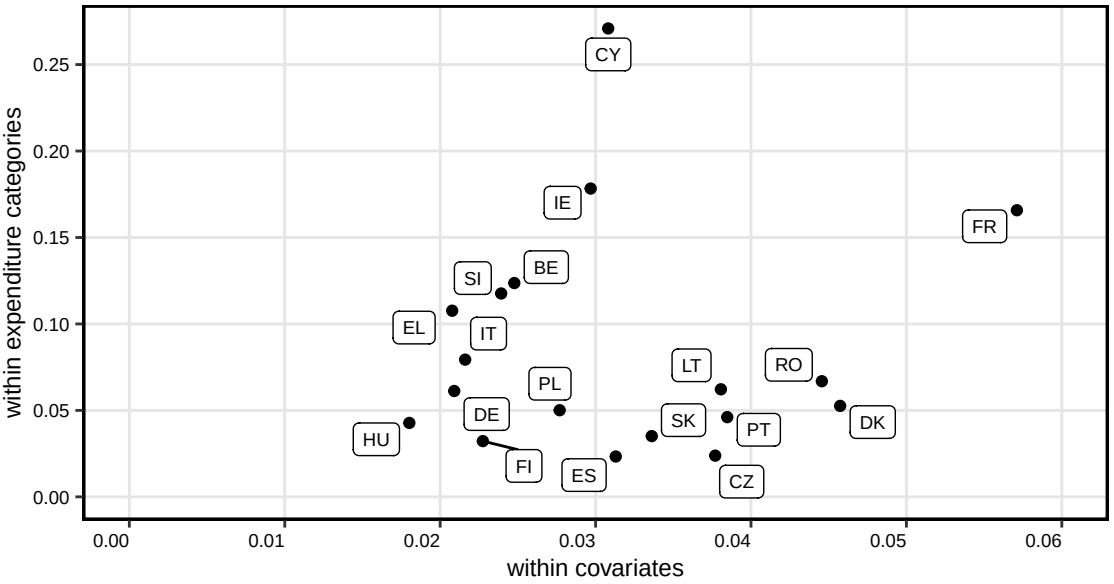

Figure 3: Scatterplot of mean diff. in correlation 'within cov' *versus* 'between cov and exp'

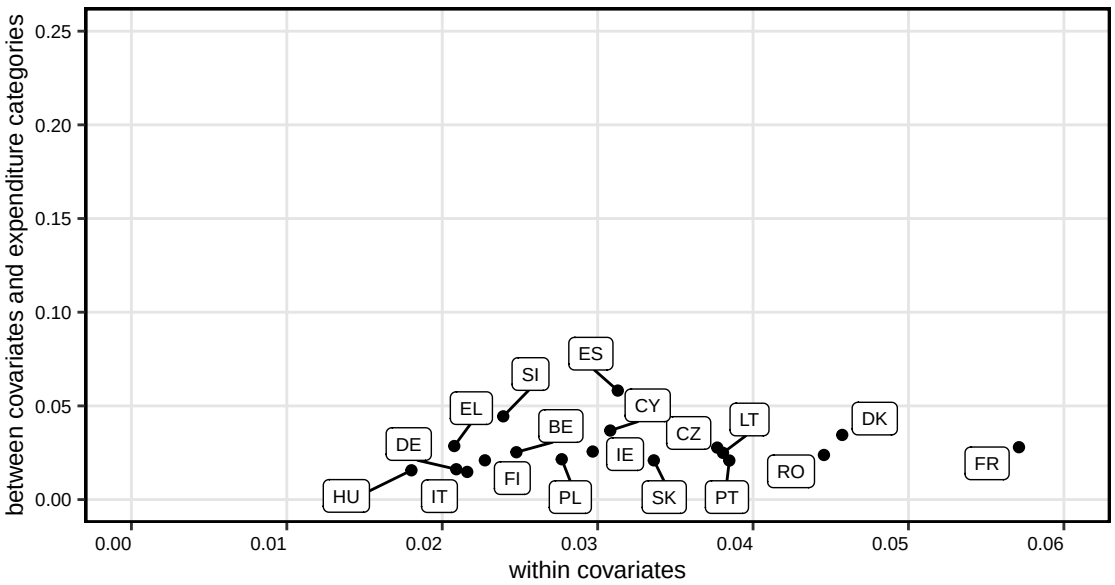

Figure 4: Scatterplot of mean diff. in correlation 'within cov' *versus* 'within exp'

6.4 Macro-validation

As a final point of evaluation, we also compared the imputed values with available macro figures on both, expenditures and indirect tax revenues. The results are summarised in the Excel file `Macro-validation ITTv3.xlsx`. The sheet `Expenditures` contains the results for the expenditures, and the sheet `Indirect taxes` contains the part on indirect tax revenues.

With respect to **expenditures**, it should be noted that we concentrate here on expenditure *levels*, contrary to what was the case for the evaluation of the imputation in previous subsections, which was done on the basis of income shares of expenditures. Simulated expenditure levels in SILC are obtained by multiplying the imputed income shares with disposable incomes as simulated by EUROMOD for the policy year 2010.[16] We then estimated total expenditures on the 12 most aggregated COICOP categories (the first level of aggregation). We compare these values with the estimates stemming from the observed expenditures in the HBS of 2010. Next, we collected 2010 National Accounts (NA) figures on household expenditures for the same categories.[17] For the category *Housing, water, electricity, gas and other fuels*, the reported statistics and simulated values are excluding virtual rents for owners–occupiers.

We then calculated the coverage of the HBS and NA expenditure levels by the simulated SILC expenditures (that is 100*SILC Expenditures/HBS expenditures and 100*SILC Expenditures/NA expenditures).

The sheet `Expenditures` is organised as follows:

- cells `1A:15T`: simulated total household expenditures in SILC on 12 COICOP categories;
- cells `17A:31T`: estimated total household expenditures on 12 COICOP categories based on the HBS;
- cells `33A:47T`: estimated total household expenditures on 12 COICOP categories based on the NA;
- cells `49A:65U`: coverage of HBS expenditure level by simulated SILC expenditures;
- cells `67A:83U`: coverage of NA expenditure levels by simulated SILC expenditures.

The entries of the coverage tables are shaded: darker colors mean higher deviations between simulated SILC figures and observed HBS, respectively NA figures, blue for underestimation, red for overestimation. These tables are augmented with a row and column, indicating the number of COICOP aggregates (per country) and the number of countries (per COICOP aggregate) respectively for which the deviation of simulated and observed values are within the 10% range.

[16] These simulated disposable incomes by EUROMOD might differ from the by the respondents reported disposable incomes in SILC (Decoster *et al.*, 2014) that we used for fitting expenditure shares in the SILC–data during the imputation procedure (see step 4 of the imputation method in Section 2.3.1).

[17] `https://ec.europa.eu/eurostat/web/products-datasets/-/nama_10_co3_p3`.

HBS–coverage is mainly affected by differences in EUROMOD simulated disposable incomes and observed HBS incomes, as the SILC values are based on imputed income shares from the HBS. So, we expect to see small differences only, as ideally, both datasets are representative for the same population in that respect. This holds largely true.

- Simulated total expenditures do not deviate more than 10% from the HBS values for 16 out the 18 countries. The largest deviation is 12%.
- Six out of the 18 countries score very well (10 or more COICOP aggregates are within the 10% bound): Cyprus, Czech Republic, Germany, Finland, Poland, and Slovakia.
- For another seven countries, 8 to 9 COICOP aggregates are within the 10% bound: Belgium, Greece, Spain, Ireland, Italy, Romania, and Slovenia.
- For Denmark, France, and Portugal, 7 categories fall within the 10% bound.
- Simulated expenditure on *education* is performing extremely bad in Germany (coverage of 27%) and France (coverage of 161%), even though these countries score well for most other commodity groups. It may be explained by the fact that the share of *education* in total expenditures is rather small. This can also be seen from the additional column `U` in the table, which shows that *education* is the most problematic group with only five countries for which simulated expenditures are in the 10% bounds of the HBS expenditures.
- Hungary and Lithuania perform worst (only 5 categories within the 10% bound), but no large outliers are found for the two countries (the deviation is at most 24%).

When comparing simulated SILC expenditures with national accounts, we observe an overall underestimation of expenditures.

- The overall average coverage is highest in Denmark (86%), Finland (86%) and Belgium (82%).
- The lowest scoring countries are Romania (44%), Lithuania (54%), Hungary (54%), Slovakia (55%), and Poland (57%).
- The best coverage is observed among categories with few zero expenditure observations. *Food & non–alcoholic beverages*, *Clothing & footwear*, and *Communications* are covered best.
- Under–coverage is largest for *Alcoholic beverages and tobacco*, with only Denmark and France covering more than half of national account figures (respectively 70% and 51%).

It should be noted that the under–coverage of simulated SILC expenditures with respect to national account figures is to a large extent a reflection of a similar under–coverage of the national account figures by HBS.

A similar exercise was performed for the countries for which the 2019 national accounts figures are already available: the ITTv3 simulated expenditures for policy year 2019 are compared with the

2019 NA figures.[18] The results can be found in cells `1W:15AO` (simulated expenditures with ITTv3 for policy year 2019 for all 18 countries), cells `33W:47AO` (NA figures on expenditures for Czech Republic, Denmark, France, Ireland, Italy, Romania, and Slovakia), and cells `67W:83AP` (coverage of NA figures by simulated expenditures). The underlying population of the SILC (2010) is in this case different from that to which the NA apply (2019). This can therefore not be considered as a true validation exercise, and we therefore do not further discuss the results here.

For the **indirect tax** validation we collected figures from EUROSTAT for the year 2010 on the VAT component of indirect tax revenues and from the European Commission on the excise components.[19] The collected tax revenues from VAT are reported on row 29 of the tab `Indirect taxes` of Excel file `Macro-validation ITTv3.xlsx`, and for excises on rows 36 to 56 of the same sheet.

Simulated indirect taxes by ITTv3 for policy year 2010 can be found on cells `A1:23T` of the same sheet.

Aggregate coverage rates of these statistics by ITTv3 simulated indirect taxes can be found in cells `55A:78T` of that sheet.

- For the VAT–component the coverage ranges between 86% (Spain) and 49% (Cyprus). Part of the gap is explained by the under–coverage of total expenditures (with respect to National Accounts). It should however be stressed that EUROMOD ITTv3 only covers the private household sector. Net VAT–payments of the corporate sectors are not included, and this may explain another part of the gap.
- This remark is even more important when considering the coverage of excises on *energy products*, comprising among other things motor fuels and heating fuels excises, which are also levied on companies (and in the case of motor fuels, on non–resident households filling up their gas tank while driving through). The revenue statistics do not distinguish by economic sector or by residency of the consumer. Excises from *Fuels and lubricants* make up the brunt of energy excise revenues, and therefore drive the coverage statistics.
- The non–covering of the corporate sector is much less of an argument for the under–coverage of excises on tobacco. But we saw already that under–coverage was highest for *alcoholic beverages and tobacco*, which certainly helps explain the results.

All in all, the results on indirect tax revenue coverage are in line with those of previous versions of ITT.

[18] Simulating expenditures for another policy year than that of the underlying SILC dataset with imputed income shares of expenditures (2010) is done by uprating the various components of gross incomes in SILC 2010 to 2019, using the uprating tools standard available in EUROMOD. Other household characteristics are left unchanged.

[19] VAT figures stem from `https://ec.europa.eu/eurostat/cache/metadata/en/gov_10a_taxag_esms.htm`.
Excises were collected from:
`https://ec.europa.eu/taxation_customs/.../excise_duties_alcohol_en.pdf` for alcoholic beverages;
`https://ec.europa.eu/taxation_customs/.../excise_duties_tobacco_en.pdf` for tobacco;
`https://ec.europa.eu/taxation_customs/.../excise_duties_energy_products_en.pdf` for energy products.

7 Integration of the Indirect Tax Tool into EUROMOD

Now that the underlying datasets in EUROMOD have been enriched with detailed expenditure data, it is possible to extend the EUROMOD calculator with routines to compute indirect taxes (VAT and excises). We recall that the underlying database of EUROMOD is a household database. Therefore the taxable base of VAT and excises is limited to the household sector. VAT and excises levied on the non–household sector, *e.g.* on expenditures by corporations, remain outside the scope of EUROMOD.

In Section 7.1 we lay out the foundations for the indirect tax simulations and the key assumptions facing these simulations. Section 7.2 reports on the implementation in EUROMOD.

7.1 Indirect tax simulations

In this subsection we lay out the general principles and definitions of VAT, and specific and *ad valorem* excises. We then turn to the way a baseline and indirect tax reforms are simulated in EUROMOD. Finally, we explain the implications of relaxing the basic assumption of full pass–through.

7.1.1 Indirect tax instruments and liabilities

The principles of calculating indirect taxes in ITTv3 are similar to ITTv2 (see De Agostini *et al.* 2017), but in ITTv3 we do the calculations at the most detailed level of aggregation available (in most cases COICOP level 4). A good at this level of aggregation is indexed by k. In the sequel of this subsection we act as if the statutory rates are uniquely defined at this level of aggregation. We will explain in Section 7.2 how we resolve cases where statutory tax rates are not unique at this level of aggregation.

Table 5 summarises the notation for prices and taxes on individual commodities. We explain the different elements of the table and their mutual relations below.

Table 5: Notation for prices and taxes for individual commodity k

Consumer (unit) price	q_k	per (quantity) unit
Producer (unit) price	p_k	per (quantity) unit
VAT rate	t_k	% of producer price plus excises
Specific excise	a_k	per (quantity) unit
Ad valorem excise	v_k	% of consumer price
Implicit tax rate	τ_k	% of producer price

The total tax liability payable on a good k by a household h, denoted by T_k^h is the difference between expenditure on good k by the household, e_k^h, and seller's revenues obtained from this expenditure. Defining household expenditure on commodity k as quantity bought, x_k^h, times consumer price, that is

$$e_k^h = q_k x_k^h, \tag{33}$$

and seller's revenues by producer price times that quantity, $p_k x_k^h$, we obtain the following definition T_k^h:

$$T_k^h = (q_k - p_k)\, x_k^h. \tag{34}$$

The wedge between consumer price and producer price originates from different indirect tax instruments, to wit:

- a specific excise tax a_k, levied on the quantity x_k,
- an *ad valorem* excise rate, levied on the consumer price q_k, and
- a VAT rate t_k, levied on the producer price p_k, augmented with excises $a_k + v_k q_k$ for goods on which excises are payable.

This gives the following relation between consumer price q_k and producer price p_k:

$$q_k = (1 + t_k)(p_k + a_k + v_k q_k). \tag{35}$$

To summarise the different indirect tax instruments, we define one implicit tax rate, on good k, τ_k, as:

$$q_k \equiv (1 + \tau_k)\, p_k. \tag{36}$$

Using Equation (35), an explicit formula for the implicit indirect tax rate can be obtained:

$$\tau_k = \frac{q_k}{p_k} - 1 = (1 + t_k)\, \frac{1 + \frac{a_k}{p_k}}{1 - (1 + t_k)\, v_k} - 1. \tag{37}$$

We use Equation (37) to report the total implicit indirect tax rates on the commodities in EUROMOD. Notice that for goods on which no specific excises are levied ($a_k = 0$), the implicit rate can be calculated solely from the statutory rates, and no information on prices is needed. We explain below how the implicit indirect tax rate is calculated for goods on which also specific excises are levied.

Equations (33) and (36) allow to rewrite the total indirect tax liability of Equation (34) as:

$$\begin{array}{rcccl} T_k^h & = & \frac{q_k - p_k}{q_k} q_k x_k^h & = & \frac{(1+\tau_k)p_k - p_k}{(1+\tau_k)p_k} e^h \\ & & & = & \frac{\tau_k}{1+\tau_k} e_k^h. \end{array} \tag{38}$$

We now discuss how to decompose the total indirect tax liability into the three components, VAT, *ad valorem* and specific excises. We start with the VAT component:

$$\begin{array}{rcll} T_{t_k}^h & = & t_k\, (p_k + a_k + v_k q_k)\, x_k^h & \\ & = & \frac{t_k}{1+t_k}\, (1 + t_k)\, (p_k + a_k + v_k q_k)\, x_k^h & \text{(divide and multiply by } 1 + t_k) \\ & = & \frac{t_k}{1+t_k} e_k^h & \text{(using Equations 33 and 35).} \end{array} \tag{39}$$

The *ad valorem* component equals

$$T^h_{v_k} \;=\; v_k q_k x^h_k \;=\; v_k e^h_k \quad \text{(the last equality follows from Equation 33).} \tag{40}$$

The specific excise is equal to:

$$T^h_{a_k} \;=\; a_k x^h_k \;=\; \frac{a_k}{q_k} e^h_k \quad \text{(the last equality follows from Equation 33).} \tag{41}$$

One can now verify that

$$\begin{aligned} T^h_{t_k} + T^h_{v_k} + T_{a_k} &= \left(\tfrac{t_k}{1+t_k} + v_k + \tfrac{a_k}{q_k}\right) e^h_k \\ &= \left(q_k \tfrac{t_k}{1+t_k} + q_k v_k + a_k\right) x^h_k && \text{(using Equation 33)} \\ &= \left(\tfrac{(1+t_k)(p_k+a_k+v_k q_k)+t_k\ q_k-(1+t_k)p_k}{1+t_k}\right) x^h_k \\ &= (q_k - p_k)\, x^h_k && \text{(using Equation 35)} \\ &= T^h_k && \text{(using Equation 34)} \end{aligned} \tag{42}$$

Equations (39)–(40) show that one can calculate indirect tax liabilities for commodities on which no specific excises are levied solely on the basis of information on expenditures and statutory rates. For goods on which specific excises are levied we need in addition to calculate quantities measured (in units in which the rate a_k is specified). Using Equation (33), these quantities can be calculated as follows:

$$x^h_k = \frac{e^h_k}{q_k}. \tag{43}$$

So, only for goods on which specific taxes are levied, we need consumer price information (in units in which the specific excise is specified) to calculate the indirect taxes. Once such information on the consumer price is available, one can derive from it also the associated producer price (in the same units) using Equation (35):

$$p_k = q_k \left(\frac{1}{(1+t_k)} - v_k\right) - a_k. \tag{44}$$

This producer price can then be used to calculate the implicit tax rate for the goods with specific excises in Equation (37). It will also prove useful when implementing the fixed producer prices assumption below.

For the purpose of welfare analysis we do use a measure of quantities for *all* commodities, but these quantities are measured in monetary terms at producer prices. To construct this measure we only need expenditures on good k, e^h_k and the implicit tax rate, τ_k. Indeed, let quantities measured at producer prices be denoted by $\widetilde{x}^h_k$. Then,

$$\widetilde{x}^h_k \equiv p_k x^h_k = \frac{q_k}{1+\tau_k} x^h_k = \frac{e^h_k}{1+\tau_k}, \tag{45}$$

where the first equality follows from Equation (36) and the last from Equation (33). Notice that when one measures quantities at producer prices, the producer prices themselves are normalised to one and consumer prices to $1+\tau_k$. Indeed, it follows from Equation (45) that $(1+\tau_k)\cdot 1\cdot \widetilde{x}^h = e_k^h \equiv \widetilde{q}_k \widetilde{x}_k^h$, where $\widetilde{q}_k$ are the renormalised consumer prices. They are defined as

$$\widetilde{q}_k \equiv 1+\tau_k. \tag{46}$$

These are the consumer prices as reported in EUROMOD. Once again, it should be stressed that they are calculated from the implicit tax rate, and need not be collected separately.

Finally, let us denote household disposable income by y^h, and total expenditures by E^h. The latter is defined as the sum of expenditures on all goods:

$$E^h \equiv \sum_k e_k^h. \tag{47}$$

Household h's saving, denoted by S^h, is then derived as

$$S^h = y^h - E^h. \tag{48}$$

Household h's (disposable) income share of expenditure on good k, denoted by w_k^h, is defined as

$$w_k^h = \frac{e_k^h}{y^h}. \tag{49}$$

Similarly, the share of expenditure on k in total expenditures of household h, denoted by ω_k^h, is equal to:

$$\omega_k^h = \frac{e_k^h}{E^h}. \tag{50}$$

We summarise the variables explained in this section in Table 6.

Table 6: Variables at household h and commodity k level

Expenditures	$e_k^h = q_k x_k^h$	
Quantity (needed for goods with specific excises only)	$x_k^h = \frac{e_k^h}{q_h^k}$	
Quantity at producer prices (for welfare analysis)	$\widetilde{x}_k^h = \frac{e_k^h}{1+\tau_k}$	
Income share of expenditures on k	$w_k^h = \frac{e_k^h}{y^h}$	
Expenditure share of expenditures on k	$\omega_k^h = \frac{e_k^h}{E^h}$	
VAT tax liability	$T_{t_k}^h = \frac{t_k}{1+t_k} e_k^h$	(see Equation 39)
Ad valorem excise liabilities	$T_{t_k}^h = \nu_k e_k^h$	(see Equation 40)
Specific excise liabilities	$T_{a_k}^h = \frac{a_k}{q_k} e_k^h$	(see Equation 41)

7.1.2 Simulation and behavioural assumptions

In this section we lay out how ITTv3 effectively simulates indirect taxes, and which assumptions are made in the process. We distinguish between a baseline run and the simulation of a(n indirect) tax reform or other price or income shocks.

The (disposable) income shares of expenditures have been imputed from the HBS's of 2010 into the EUROMOD input data of 2010 (2012 for Denmark). To run a baseline for a policy system of year xxxx, one needs the direct and indirect tax parameters of that year and the consumer prices of the goods on which specific excises are levied. At this moment two policy systems are available for the 18 countries on which ITTv3 can be run: 2010 and 2019. In case the policy year is different from the income collection year of the input data on which the income shares of expenditures are imputed, incomes of the input dataset are uprated according the uprating rules currently implemented in EUROMOD (see EUROMOD Modelling Conventions, 2015). An alternative that we did not explore during the present project, but may be worthwhile to evaluate, is to impute within EUROMOD the income shares on the EUROMOD input dataset with income collection year closest to the policy year.

For simulating tax reforms or other price or income shocks, three assumptions on expenditure behaviour are available: either (1) constant income shares, (2) constant quantities, or (3) constant expenditure shares. All three available options use the calculation of producer prices for the goods with specific excises in the baseline. The parameters necessary to calculate them (indirect tax parameters and consumer prices of goods with specific excises) are included in the EUROMOD policy parameter set. Simulating with the constant quantities assumption needs in addition the baseline quantities. Simulating with the constant expenditure shares needs in addition the baseline disposable income and saving level. Note that none of these baseline variables are in the input dataset, and thus the last two behavioural assumptions require two EUROMOD runs, where the output of the first run will be used as an input in the second.

In the sequel, parameters and variables that are tied to the baseline simulation are subscripted with letter b.

7.1.2.1 Baseline

The from the HBS imputed income shares of expenditures on commodity k, with which the EUROMOD input data are augmented (see Section 7.2.1) are denoted by $w_{k,b}$. We can currently simulate a baseline for policy years 2010 and 2019.

When simulating a baseline, EUROMOD calculates household disposable income y_b^h. Expenditures on commodity k are then simulated as:

$$e_{k,b}^h = w_{k,b} y_b^h \tag{51}$$

These expenditures are then used in Equations (39)–(41) to calculate VAT, and *ad valorem* and specific excises, using the baseline rates $t_{k,b}$, $a_{k,b}$, and $v_{k,b}$. As can be seen from Equation (41) consumer prices in units of the specific excises, are needed for the calculation of the specific excises. These are collected for the policy years 2010 and 2019 and included in the EUROMOD policy parameter set (see Section 7.2.3.1).

The baseline consumer prices allow by Equation (44) to calculate baseline producer prices for these commodities too:

$$p_{k,b} = q_{k,b}(\frac{1}{(1+t_{k,b})} - v_{k,b}) - a_{k,b}. \tag{52}$$

Using Equation (37), one can calculate the baseline implicit rate as:

$$\tau_{k,b} = (1+t_{k,b})\,\frac{1+\frac{a_{k,b}}{p_{k,b}}}{1-(1+t_{k,b})\,v_{k,b}} - 1, \tag{53}$$

and baseline quantities measured at producer prices, are then:

$$\widetilde{x}^h_{k,b} = \frac{e^h_{k,b}}{(1+\tau_{k,b})}. \tag{54}$$

7.1.2.2 Reforms

The tax liabilities for VAT, *ad valorem* and specific excises (39)–(41) are functions of expenditures on each commodity, e^h_k, and the post–reform tax rates. The behavioural assumptions determine how a tax reform or income shock translates into these new expenditure levels e^h_k. This is further explained in Section 7.1.2.2.2.

Once these new expenditure levels are simulated, VAT liabilities and *ad valorem* excises are calculated using (39)–(40) with the post–reform tax rates. The method of calculation does not depend therefore on the behavioural assumptions. The same holds true for the simulation method of the new specific excises and the new implicit rate τ_k. But in order to be able to effectively calculate them, an additional assumption is needed, namely that of constant producer prices. Next, we explain how this is accomplished.

7.1.2.2.1 Constant producer prices and specific excises

We impose the assumption that producer prices do not change due to a tax reform. Therefore, we can equate post–reform producer prices, p_k, with baseline prices, $p_{k,b}$: $p_k = p_{k,b}$. These baseline producer prices are deduced from the baseline consumer prices by Equation (52). So, in order to simulate specific excises, we require not only the relevant tax reform parameters, but also those of the baseline, as well as the baseline consumer prices of goods on which specific excises are levied. These prices and baseline parameters are added to the policy parameters (see Section 7.2.3.1).

Given that $p_{k,b}$ can be calculated again from the policy parameters, and equals p_k by the assumption of constant producer prices, we can calculate the new consumer prices for goods with specific excises

under the reform policy characterised by the new rates, t_k, v_k, and a_k, by inverting Equation (44), to obtain:

$$q_k = \frac{p_{k,b}+a_k}{\left(\frac{1}{1+t_k} - v_k\right)}. \tag{55}$$

These post–reform consumer prices can then be used to calculate the new quantities on which the new specific excises apply:

$$x_k = \frac{e_k^h}{q_k} = e_k^h \frac{\left(\frac{1}{1+t_k} - v_k\right)}{p_{k,b}+a_k}. \tag{56}$$

Equation (41) for specific excise revenue then becomes:

$$\begin{aligned} T_{a,k}^h &= a_k x_k^h &&= a_k \frac{e_k^h}{q_k} \\ &&&= a_k e_k^h \frac{\left(\frac{1}{1+t_k} - v_k\right)}{p_{k,b}+a_k}. \end{aligned} \tag{57}$$

The new implicit rate τ_k can be calculated as:

$$\tau_k = (1+t_k)\frac{1+\frac{a_k}{p_{k,b}}}{1-(1+t_k)v_k} - 1. \tag{58}$$

Again, it should be stressed that for goods without specific excises, no calculation of baseline producer prices is necessary, and the implicit rate, τ_k, and normalised consumer prices, $(1+\tau_k)$, are determined solely by the post–reform rates t_k and v_k.

7.1.2.2.2 Behavioural assumptions

The tax liability formulae for VAT and *ad valorem* excises (39)–(40) are functions of expenditures on each commodity k, e_k^h, and the post–reform tax rates. For the specific excise tax component, we explained in Section 7.1.2.2.1 how we can exploit the assumption of constant producer prices to obtain the post–reform quantities on which these taxes levied. This is not affected by the behavioural assumptions. The same holds true for the calculation of the post–reform implicit tax rate τ_k (see Equation 58).

We can thus limit our further discussion of the three behavioural assumptions the EUROMOD user can choose from, to the way household expenditures are determined by each of them.

1. Constant income shares of expenditures

   Under the first behavioural assumption, households always spend a constant share of their disposable income on each commodity k: $w_k = w_{k,b}$, where $w_{k,b}$ are the income shares that are imputed from the HBS dataset, and with which the EUROMOD input data are augmented:

   $$e_k^h = w_{k,b} y^h \tag{59}$$

   As these income shares $w_{k,b}$ have been imputed from the HBS data and given that disposable household income y^h is simulated within EUROMOD, we have sufficient information to calculate all expenditures, and hence, all indirect taxes.

Under this assumption, own price elasticities of all commodities and saving are all equal to -1, cross price elasticities are zero, and income elasticities are all equal to one. Compensated price elasticities are household specific.

2. Constant quantities

The second behavioural assumption holds that households always consume the same quantity of commodities, $\widetilde{x}_k = \widetilde{x}_{k,b} = e^h_{k,b}/\left(1+\tau_{k,b}\right)$ (see Equation 54), no matter the change in their income, and irrespective of relative price changes. One then derives the expenditures from the product of quantities and the (new normalised) consumer prices, $1+\tau_k$ (which are derived from Equation 58).

$$\begin{aligned} e^h_k &= \widetilde{x}_{k,b}(1+\tau_k) \\ &= \widetilde{x}_{k,b}(1+t_k)\frac{1+\frac{a_k}{p_{k,b}}}{1-(1+t_k)v_k}. \end{aligned} \tag{60}$$

It is important to note that this assumption requires two EUROMOD runs:

- Step 1: run the baseline. From the baseline expenditures $e^h_{k,b}$, baseline quantities at producer prices $\widetilde{x}^h_{k,b} = \frac{e^h_{k,b}}{1+\tau_{k,b}}$ can be derived.[20]
- Step 2: add these baseline quantities $\widetilde{x}^h_{k,b}$ to the EUROMOD input data.
- Step 3: run the reform, and switch on the function calculating expenditures under the *constant quantities* assumption, in which new expenditures are calculated as follows: $e^h_k = \widetilde{x}^h_{k,b}\left(1+\tau_k\right)$.

Under the constant quantities assumption all price elasticities and income elasticities of commodities equal zero. Saving is adapted to close the household budget.

3. Constant expenditure shares

According to the third behavioural assumption, nominal saving is kept constant at the baseline level. This amount of saving determines post–reform total expenditures by subtracting it from the post–reform disposable income level. While keeping saving constant, a change in disposable income translates one–to–one into a change in total expenditures. Total expenditures is allocated to the different goods by applying constant expenditure shares, that is, using the same expenditures shares as in the base line. Formally,

$$\begin{aligned} S^h &= S^h_b, \\ E^h &= y^h - S^h_b, \\ e^h_k &= \omega^h_{k,b} E^h. \end{aligned} \tag{61}$$

As is the case with the *constant quantities* assumption, the *constant expenditure shares* assumption requires two EUROMOD runs:

[20] The reader can verify that $\widetilde{x}^h_{k,b} = x^h_{k,b}p_{k,b} = x^h_k p_{k,b}$, or alternatively, that $x^h_{k,b} = x^h_k$. Of course, this can only be checked empirically for goods with specific excise taxes, as we do not have, neither need non–normalised consumer prices of the other goods.

- Step 1: run the baseline. From the simulated expenditures, household saving S_b^h is derived.
- Step 2: add baseline saving S_b^h and baseline income y_b^h to the EUROMOD input data.
- Step 3: run the reform, and switch on the function calculating expenditures under the *constant expenditure shares* assumption. This function simulates total expenditures $E^h = y^h - S_b^h$, and then calculates the expenditures as $e_k^h = \omega_{k,b}^h E^h$, with baseline expenditure shares $\omega_{k,b}^h = w_{k,b}^h \frac{y_b^h}{E_b^h} = w_{k,b}^h \frac{y_b^h}{y_b^h - S_b^h}$.

As with constant income shares, own price elasticities of all commodities are all equal to -1, cross price elasticities are zero. Compensated price elasticities are household specific. Income elasticities are greater than one if saving is positive, and smaller than one if there is dissaving. Income and price elasticities of saving are zero.

7.1.3 Introducing a tax incidence parameter

As explained in Section 7.1.2.2.2, ITTv3 implements three scenarios on consumers' behavioural reactions to changes in commodity prices: constant income or expenditure shares (*i.e.* assuming no cross price effects and own price elasticities equal to minus one) and constant quantities (assuming both own and cross price elasticities equal to zero).[21] Currently, those scenarios assume producer prices to be fixed (see Section 7.1.2.2.1), which implies a proportional change in indirect tax rate being fully reflected in consumer prices. In the literature, this is referred to as full pass–through. This section investigates the implications of relaxing the full pass–through assumption for the constant income share scenario.

7.1.3.1 Tax incidence or pass–through

The present section introduces the notion of a tax incidence or pass–through parameter for any commodity k . We drop therefore the subscript k in the notation. The relation between producer and consumer prices is:

$$q = p\,(1+\tau)\,. \tag{62}$$

So far producer prices were assumed to be fixed. This means, among other things, that they are not affected by changes in the tax rate. Consequently, the proportional change in the consumer price due to a change in the tax rate $\mathrm{d}\,\tau$ is equal to:

$$\frac{\mathrm{d}\,q}{q} = \frac{\mathrm{d}\,\tau}{1+\tau}. \tag{63}$$

When this holds, we say there is full pass–through (or that the pass–through is equal to one or 100%).

[21] The distinction between constant income and constant expenditure shares is irrelevant for the present analysis. We illustrate things further only for the constant income shares case.

Incomplete pass–through occurs when

$$\frac{\mathrm{d}\,q}{q} = \theta \frac{\mathrm{d}\,\tau}{1+\tau}, \tag{64}$$

with $\theta \neq 1$. The parameter θ is the pass–through parameter. Full pass–through occurs when $\theta = 1$. In order to allow for incomplete pass–through, it is necessary to drop the fixed producer price assumption. More specifically, the producer price p becomes a function of the tax rate:

$$p = p\left(\tau\right), \tag{65}$$

and thus,

$$q = p\left(\tau\right)\left(1+\tau\right). \tag{66}$$

In the present approach we will define the functional form of $p(\tau)$ implicitly by means of a partial equilibrium framework. This framework assumes an equilibrium between market demand (X^D) and supply (X^S) of a commodity, which are respectively function of the consumer and producer price:

$$X^D\left(p(\tau)\left(1+\tau\right)\right) = X^S\left(p(\tau)\right). \tag{67}$$

Differentiating with respect to τ and solving for $\frac{\partial \ln p(\tau)}{\partial \tau}$ gives:

$$\frac{\partial \ln p(\tau)}{\partial \tau} = \frac{\varepsilon^D}{\varepsilon^S - \varepsilon^D}\frac{1}{1+\tau}, \tag{68}$$

where ε^D and ε^S are respectively price elasticities of market demand and supply, i.e. $\varepsilon^D = \frac{\partial \ln X^D}{\partial \ln q}$ and $\varepsilon^S = \frac{\partial \ln X^S}{\partial \ln p}$.

Using the definition of q in Equation (66), one can derive that:

$$\frac{\partial \ln q}{\partial \tau} = \frac{\varepsilon^S}{\varepsilon^S - \varepsilon^D}\frac{1}{1+\tau}. \tag{69}$$

So,

$$\frac{\mathrm{d}\,q}{q} = \frac{\partial \ln q}{\partial \tau}\mathrm{d}\,\tau = \frac{\varepsilon^S}{\varepsilon^S - \varepsilon^D}\frac{\mathrm{d}\,\tau}{1+\tau}.\ ^{22} \tag{70}$$

Consequently, in this partial equilibrium framework the pass–through parameter equals:

$$\theta = \frac{\varepsilon^S}{\varepsilon^S - \varepsilon^D}. \tag{71}$$

Table 7 specifies the value of the pass–through parameter for some assumptions on the price elasticity of market demand and supply.

Notice that for the constant quantity demand approach, all price elasticities of individual demand, and hence price elasticity of market demand, are zero. Consequently in the constant quantities approach, the pass–through parameter is fixed to one, and there is always full pass–through. Furthermore, fixed producer prices can be rationalised by the assumption of an infinite supply elasticity, which also amounts to full pass–through.

[22] In fact, Equation (70) only holds approximately, but here we will neglect these approximation errors due to linearisation.

Table 7: Pass–through parameter for different values of elasticities

Elasticity	Pass–through parameter θ
$\varepsilon^S \to \infty$	$\theta = 1$
$\varepsilon^D \to -\infty$	$\theta = 0$
$\varepsilon^S = 0$	$\theta = 0$
$\varepsilon^D = 0$	$\theta = 1$

7.1.3.2 Implication of introducing a pass–through parameter for constant shares

With the ITTv3 option of constant income shares, all individual own price elasticities, and hence price elasticity of market demand, are equal to -1. From Equation (71), the pass–through parameter in this case becomes:

$$\theta = \frac{\varepsilon^S}{1 + \varepsilon^S} \tag{72}$$

There is therefore scope for incomplete pass–through within the currently implemented demand model of constant shares in ITTv3. However, with constant income shares, expenditures by a household h on a commodity k, e_k^h, are fixed and equal to:

$$e_k^h = w_k^h y^h, \tag{73}$$

where w_k^h is independent of the price in a constant income share demand system. Consequently, the VAT–bill and the *ad valorem* tax bill, which are calculated as follows:[23]

$$\begin{aligned} T_{t_k}^h &= \frac{t_k}{1+t_k} e_k^h, \\ T_{\nu_k}^h &= v_k e_k^h, \end{aligned} \tag{74}$$

are independent of the pass–through parameter, and remain unaltered as compared to the case of full pass–through currently implemented. This is an important insight for further discussions: since we have, besides the constant quantity assumption discussed above, only the constant shares option built in in ITTv3 version, the implementation of a pass–through parameter different from 1, does not influence the tax liabilities for VAT and *ad valorem* taxes.

However, this does of course not imply that everything remains the same under incomplete pass–through. First, producer prices do change, and this affects profits. But the analysis of this effect – relevant of course in a general equilibrium approach – lies beyond the scope of the current EUROMOD microsimulation model. Second, consumer prices change differently depending on the pass–through parameter (see Equation 70). It implies that alternative assumptions on the pass–through will have different implications for consumer welfare, the latter being determined by the quantities consumed (instead of expenditures). For assessing this effect quantitatively, we only need the impact of a tax change on the consumer price. This requires assessing the effect of a change in the VAT or *ad valorem*

[23] See Section 7.1.1 Equation (39) and (40) for the derivation of these formula's.

rate on the implicit tax rate. Indeed, Equation (70) expresses the effect on consumer price in terms of the change in the implicit tax rate τ and the latter is a function of the statutory VAT rates and the *ad valorem* and specific excises.[24] For goods without specific excises, we can use Equation (37) with $a_k = 0$, to obtain the change in implicit tax rate as:

$$\begin{array}{lcll} \tau_{k,b} & = & \frac{1+t_{k,b}}{1-\left(1+t_{k,b}\right)\, v_{k,b}} - 1 & \text{(baseline implicit rate)}, \\ \tau_k & = & \frac{1+t_k}{1-(1+t_k)\, v_k} - 1 & \text{(post– reform rate)}, \\ \mathrm{d}\,\tau_k & = & \tau_k - \tau_{k,b}. & \end{array} \tag{75}$$

Using Equation (70), we obtain the new consumer price as:

$$\widetilde{q}_k = 1 + \tau_{k,b} + \theta\left(\tau_k - \tau_{k,b}\right). \tag{76}$$

Notice that when $\theta = 1$, this reduces to the already familiar expression $\widetilde{q}_k = 1 + \tau_k$. Post–reform quantities are then:

$$\widetilde{x}_k = \frac{e_k^h}{\widetilde{q}_k} = \frac{e_k^h}{1 + \tau_{k,b} + \theta\left(\tau_k - \tau_{k,b}\right)}. \tag{77}$$

As expressed here, these post–reform quantities are measured in the same units as pre–reform quantities, that is in monetary terms at pre–reform producer prices.

Things become more involved for goods on which also a specific excise is levied. The implicit tax rate now becomes a function of the producer price (see Equation 37), while the latter depends on τ_k by Equation (68). We get the following set of two equations in two unknowns (τ_k and p_k):

$$\begin{array}{lcl} \tau_k & = & (1 + t_k)\, \frac{1+\frac{a_k}{p_k}}{1-(1+t_k)\, v_k} - 1, \\ \frac{\mathrm{d}\, p_k}{p_{k,b}} & = & (\theta - 1)\, \frac{\mathrm{d}\,\tau_k}{1+\tau_{k,b}}, \end{array} \tag{78}$$

where $d\, p_k = p_k - p_{k,b}$, and $\mathrm{d}\,\tau_k = \tau_k - \tau_{k,b}$, and baseline variables are known (or can be calculated).

There is no obvious explicit solution for this set of two equations, and the current EUROMOD architecture does not allow to implement iterative procedures to arrive at a solution. It should be stressed that when producer prices depend on the tax rate, one should take care to keep the units in which quantities are measured fixed (*e.g.* at the baseline producer prices). For goods on which a specific excise is specified, one should also calculate the new price in units in which the statutory specific excise is expressed, to calculate the new implicit rate.

Notice finally that the solution to Equations (78) is required even if the specific excise tariff remained unchanged and only the VAT or *ad valorem* rate are changed. That is because the post–tax–reform quantity on which the specific excise is levied depends on an explicit solution for the new implicit rate and price (in units in which the specific tariff is specified):

$$x_k = \frac{e_k^h}{q_k} = \frac{e_k^h}{p_k\left(1 + \tau_{k,b} + \theta\left(\tau_k - \tau_{k,b}\right)\right)}. \tag{79}$$

[24] See Equation (37) in Section 7.1.1 for an explicit definition of the implicit rate τ.

7.1.3.3 Conclusions on tax incidence

In this section we investigated the consequences of introducing a pass–through parameter within the framework of the constant shares demand model currently implemented in ITT v3. Our main result is that expenditures, and consequently the VAT–bill and the *ad valorem* excises remain unaffected by the value of the pass–through parameter. It does affect quantities consumed however. In principle there is no problem for simulating changes in consumption (quantities) for goods on which no specific excises are levied. This is important for welfare analysis. Since ITTv3 currently does not contain a welfare analysis module, we did not implement that option currently.

Moreover, things become more complicated for goods on which a specific excise is levied. As there is no explicit solution for producer prices and implicit tax rates, an iterative (numerical) procedure should be invoked. As the current EUROMOD architecture does not allow this, an incomplete pass–through parameter applicable for all goods, could currently not be implemented in ITTv3.

7.2 Implementation in EUROMOD

Contrary to the ITTv2, the ITTv3 does not rely on a EUROMOD add–on. All the operations have been modelled inside of the EUROMOD spine, in a policy sheet called `tco`. The full integration aims at increasing the model's transparency and ease of use, as it follows the EUROMOD modelling conventions and no additional knowledge is required to run a policy simulation including an indirect tax reform. Moreover the new modelling improves greatly on automating simulations, allowing for simultaneous runs of systems, and running indirect tax simulations from the command line. The new results regarding expenditures and indirect taxes are also included in the `Statistics Presenter`. The ITTv3 modifies EUROMOD on three levels: it expands the input dataset (Section 7.2.1), it adds a new policy–sheet to the spine (Section 7.2.3), and it expands the Statistics Presenter (Section 7.2.4). Currently, the ITT is implemented for the 18 countries with imputed income shares of expenditures (see Table 1 for the list of these countries), for the policy years 2010 and 2019.

7.2.1 The EUROMOD input data

Simulating taxes requires that the usual input data are extended with income shares of expenditures at COICOP level 4 (with some exceptions for countries where the HBS contains less detail for some goods, e.g. the Germany HBS only includes pooled expenditures on alcoholic beverages and tobacco products). The user is required to inform EUROMOD that the input data indeed contain these additional variables by ticking the box `Read Expenditure-related Variables` in the `Configure Databases` Country Tool.

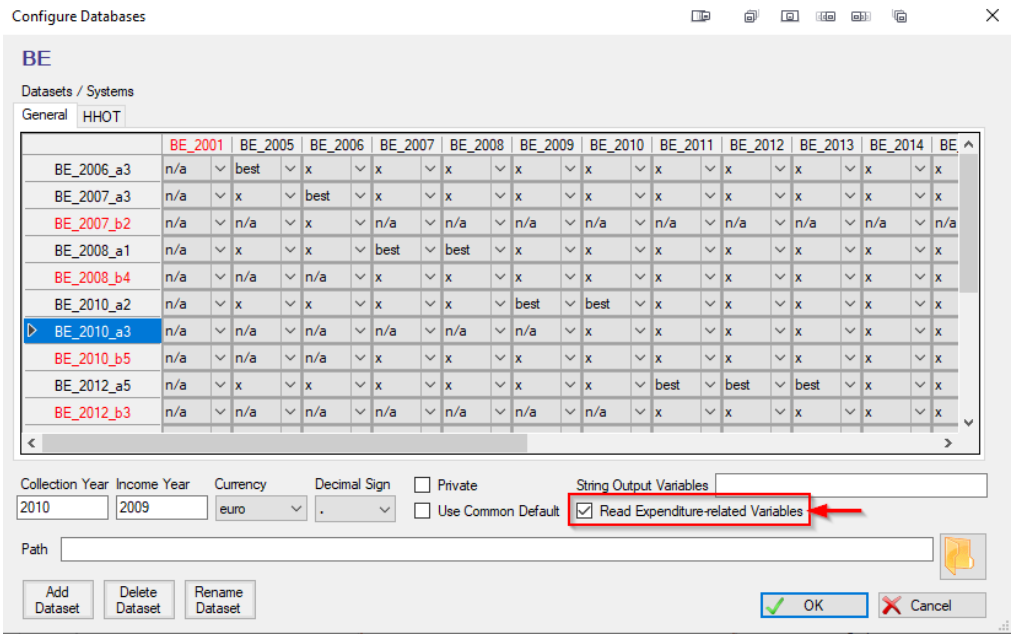

Figure 5: Tick box indicating the presence of expenditure related variables

7.2.2 The tax parameters

This subsection documents how the tax parameters that feature in the actual tax simulations were gathered, which sources we used, and which assumptions were made for their selection or manipulation. We discuss VAT rates, specific and *ad valorem* excises, and consumer prices for specific excise goods. All parameters that are included in ITTv3 are presented with their source in the Excel file `ITTv3 parameters.xlsx`.

7.2.2.1 VAT rates

The ITTv3 requires a VAT rate for each commodity in the data. These rates have been collected using the *Taxes in Europe Database v3*. For 7 of the 12 COICOP level 2 categories, linking the available commodities at the fourth COICOP level to the statutory VAT rates is rather straightforward as all commodities belonging to the group bear the same tax rate, with few exceptions.[25] These categories are *01 Food and non–alcoholic beverages*, *02 Alcoholic beverages and tobacco*, *03 Clothing and footwear*, *04 Housing, water, electricity, gas and other fuels*, *05 Furnishings, household equipment and routine maintenance of the house*, *08 Communication*, and *10 Education*.

The other 5 COICOP aggregates contain more exceptions. It requires some discretionary decisions as the commodity specification for the VAT rate does not necessarily align with the COICOP grouping in the HBS (and thus with the ITTv3 input data).

[25] The interested reader can find these exceptions by inspecting the policy sheet of the country of interest or the excel file `ITTv3 parameters.xlsx`.

- *06 Health*: the product level detail provided on the *Taxes in Europe Database v3* does not match well the COICOP categories.
- *07 Transport* has the difficulty that international public transport is often taxed at a lower rate than domestic travel, or even is exempt from VAT. However, tehre is no division between expenditures on domestic and international transport.
- *09 Recreation and culture*: this group requires substantial judgement calls on which activities have the largest weight, as often the COICOP aggregation has activities that are taxed at different rates. A common example is the different taxation of newspapers and periodicals. Same for books and e–books, or the category *Recreational and sporting services*, whereby only the latter often receives a preferential VAT rate, whereas the recreational services do not.
- *11 Restaurants and hotels*: a number of countries tax takeaways differently from regular restaurant visits (e.g. Greece, Belgium). As we cannot differentiate takeaway–meals, we only use the rate that applies to restaurant visits. Moreover, drinks and meals are often also taxed at different rates. Again, if that is the case, we maintain the rate that applies to the food.
- *12 Miscellaneous goods and services*: this is a residual group that contains goods and services of a very different nature, which are taxed very differently. Social services *social protection services* and *daycare*, *financial services* and *insuarance* are usually VAT exempt, services like *hairdressing* are often taxed at a reduced rate, and *electrical appliances for personal care*, and *jewellery, clocks and watches* are taxed at the regular VAT rate.

7.2.2.2 Excises

The specific and *ad valorem* excises are gathered using the *Excise duty tables*, published by the European Commission.[26] These tables provide yearly information on excises for alcoholic beverages, tobacco products, and energy products.[27] The EUROSTAT HBS data contain generally 13 commodities that we consider *excise goods*. Excise goods that are not well defined by these 13 commodities cannot be simulated. The provided level of detail on the statutory excises always equals or exceeds the level of detail of the commodities in the HBS and ITTv3 input data. Table 8 presents an overview of these excise goods, and summarises which specific excises are retained from the EC tables. The only *ad valorem* excises that apply are for COICOP categories *02211 Cigarettes*, *02212 Cigars*, and *02213 Other tobacco*.

- The group of ***Alcoholic beverages*** contains four commodities. For these commodities, we only retain the standard excise rates, hence discarding the reduced rates that may apply, as we cannot identify the share of the commodity that is taxed at a reduced rate (*e.g.* for spirits produced by small distilleries). As we assume this share rather small in total consumption, we are convinced the omission does not generate an important bias in the simulations. If consumption shares were known for beverages that are taxed at standard and at reduced rate,

[26] The tables can be consulted on the DG TAXUD website: http://ec.europa.eu/taxation_customs/index_en.htm#.

[27] Note that it is possible that some countries have other excsises, outside of the sphere of alcoholic beverages, tobacco or energy products, such as excises on soda's or on coffee. These fall outside of the scope of this project.

Table 8: Specific excise parameters

Commodity	**Selection of excise type**	**Unit of excise**
02111 Spirits and liqueurs	Ethyl alcohol	100 L of pure alcohol
02121 Wine from grapes or other fruit	Wine, still	100 L
02122 Wine, Other	Wine, sparkling	100 L
02131 Beer	Beer (standard rate)	100 L/°Plato or %alc.
02211 Cigarettes	Cigarettes	1000 pieces
02212 Cigars	Cigars and cigarillos	1000 pieces
02213 Other tobacco	Fine cut smoking Tobacco	per kg
04511 Electricity	Electricty (non–business)	MWh
04521 Town gas and natural gas	Natural gas (heating, non–business)	GJ
04522 Liquefied hydrocarbons	Butane, propane	1000 kg
04531 Liquid fuels	Gasoil (heating)	1000 L
04541 Solid fuels	Coal and Coke (heating, non–business)	GJ
07221 Fuels and lubricants	Petrol, Gasoil	1000 L

one might consider calculating a weighted average instead.

  - *02111 Spirits and liqueurs* uses the *Ethyl alcohol* excise rate, expressed in euros per 100 litres of pure alcohol.
  - *02121 Wine from grapes or other fruit* uses the *Wine (still)* excise rate, expressed in euros per 100 litres of product.
  - *02122 Wine, Other* uses the *Wine (sparkling)* excise rate, expressed in euros per 100 litres of product.
  - *02131 Beer* uses the *Beer* excise rate, expressed in euros either per 100 litres per °Plato, or per 100 litres per % of alcohol. To calculate these specific excises we therefore require information on either °Plato or the % of alcohol in the beer. As this information is lacking, we assume all beer consumption to be of a lager type, with 5% of alcohol, and 12° of Plato. These assumptions can be adjusted in the `tco` policy sheet.

- The group of ***Tobacco products*** contains three commodities. For most countries both a specific and an *ad valorem* excise apply. One category from the excise tables cannot be tied to a HBS commodity: *other smoking tobaccos*, *e.g.* chewing tobacco. As information on this group is absent, we discard it.
  - *02211 Cigarettes* uses the *Cigarette* excise rate, expressed in euros per 1000 pieces.
  - *02212 Cigars* uses the *Cigars and cigarillos* excise rate, expressed in euros per 1000 pieces.
  - *02213 Other tobacco* uses the *Fine Cut Smoking Tobacco (intended for the rolling of cigarettes)* excise rate, expressed in euros per kg.
- The group of ***Energy products*** contains six commodities and is the group whereby the excise tables and the HBS commodities are least aligned. The excise tables contain considerably more detail than the HBS does. For some of these goods, described hereunder, we selected one excise tariff from the more detailed EC excise tables for each of the broader commodities available in the HBS.

- For *04511 Electricity* we use the *Electricity — non–business use* excise rates, expressed in euros per MWh.
- For *04521 Town gas and natural gas* we use the *Natural gas — heating non–business use* excise rate, expressed in euros per GJ.
- For *04522 Liquefied hydrocarbons* we use the *Liquid Petroleum Gas — heating non–business use* excise rate, expressed in euros per 1000 kg.
- For *04531 Liquid fuels* we use the *Gas oil — heating non–business use* excise rates, expressed in euros per 1000 litres.
- For *04541 Solid fuels* we use the *Coal and coke — heating non–business use* excise rates, expressed in euros per GJ.

07221 Fuels and lubricants is a commodity level that – unfortunately – encompasses several important categories that are indistinguishable from one another, as it does not differentiate between gas oil and petrol. The ITTv3 therefore contains the excise rates on the most important fuel types, including *Gas oil — propellant* and *Petrol — unleaded petrol*, expressed in euros per 1000 litres. As we briefly discuss in Section 7.2.3.1, the ITTv3 then calculates the average, whereby the user can choose to calculate a simple average of the excise rates, or feed the ITT with information on the overall expenditure shares of each fuel type, and thus create a weighted average instead.

A number of excise rates could not be tied to a HBS commodity, or were discarded as they were deemed to be too insignificant. *LPG* was for example not included in *07221 Fuels and lubricants*. *Kerosene* and *Heavy fuel oil* have not been taken up either.

7.2.2.3 Consumer prices

In order to calculate specific excises the consumer price of each good on which specific excises are levied, is needed, in the same unit as the unit in which the excise tariff is specified, or convertible in these units. These prices are collected for the policy years 2010 and 2019 and included in the `Parameters baseline` function (*cf. infra* Section 7.2.3.1). We use a mixture of consumer prices that are observed and consumer prices that are calculated by means of aggregate statistics on total sales in euro and in quantities (See Table 9).

- The consumer prices for the four commodities within ***Alcoholic beverages*** are generally calculated as total revenues for a given year in a given country divided by the total volume of sales. The data are provided by the online database *Statista*.
  - For *02111 Spirits and liqueurs* we calculate a weighted average of the prices for 1 L of Whisky, Vodka, Rum, Gin, Brandy and Liqueurs, whereby the price of each spirit is weighted by the sales in euro.
  - The consumer price for *02121 Wine from grapes or other fruit* is the price for 1 L wine, consumed at home, measured by total sales in euro divided by total sales in litres.
  - The consumer price for *02122 Wine, Other* is the price for 1 L sparkling wine, consumed

Table 9: Excise goods and consumer prices

Commodity		**Product type**	**Unit**	**Source**
02111	Spirits and liqueurs	Total revenue from sales of spirits (home consumption) / total volume	1 L	Statista
02121	Wine from grapes or other fruit	Total revenue from sales of wine / total volume (home consumption)	1 L	Statista
02122	Wine, other	Total revenue from sales of sparkling wine / total volume (home consumption)	1 L	Statista
02131	Beer	Total revenue from sales of beer / total volume (home consumption)	1 L	Statista
02211	Cigarettes	Total revenue from sales of cigarettes / total volume	1000 pieces	Statista
02212	Cigars	Total revenue from sales of cigars / total volume	1000 pieces	Statista
02213	Other tobacco	Total revenue from sales of smoking tobacco / total volume	1 kg	Statista
04511	Electricity	Electricity: 2500 kWh < Consumption < 5000 MWh	MWh	Eurostat
04521	Town gas and natural gas	Natural gas: 20 GJ < Consumption < 200 GJ	GJ	Eurostat
04522	Liquefied hydrocarbons	Butane (or propane)	1000 kg	National sources
04531	Liquid fuels	Heating oil	1000 L	EC–DG Energy
04541	Solid fuels	Coal — anthracite	1000 kg	National sources
07221	Fuels and lubricants	Petrol (Euro Super-95), Diesel (Gas–Oil)	1000 L	EC–DG Energy

at home, measured by total sales in euro divided by total sales in litres.
  - The consumer price for *02131 Beer* is the average price for 1 L of beer, consumed at home, measured by total sales in euro divided by total sales in litres.

  For a number of countries these data were supplemented with reported consumer prices (for a comprehensive overview and sources see the Excel file `ITTv3 parameters.xlsx`).
- Similarly, the consumer price for ***Tobacco products*** are generally calculated as total sales in euro for a given year in a given country divided by the total sales in units, using aggregate statistics from *Statista*.
  - The consumer price for *02211 Cigarettes* is expressed per 1000 cigarettes.
  - The consumer price for *02212 Cigars* is expressed per 1000 cigars or cigarillos.
  - The consumer price for *02213 Other tobacco* is expressed per kg of rolling tobacco.

  For a number of countries these data were supplemented with reported consumer prices (for a comprehensive overview and sources see the Excel file `ITTv3 parameters.xlsx`).
- The consumer price for ***Energy products*** uses mainly reported consumer prices.
  - EUROSTAT reports consumer prices for *04511 Electricity*. We use the price per MWh for consumption between 2500 kWh and 2000 MWh.
  - EUROSTAT also reports consumer prices for natural gas, which we use for *04521 Town gas and natural gas*. We use the price for consumption between 20 GJ and 200 GJ.
  - Consumer prices for *04522 Liquefied hydrocarbons* are more difficult to find, as we did not find a centralised database reporting either historical price evolutions, or information on total sales and total volumes sold. So instead we searched the prices ourselves (first quarter 2020), and assume the prices to be constant over time.
  - The Directorate General of Energy of the European Commission reports on the consumer prices of *Heating oil*, which is our proxy for the category *04531 Liquid fuels*.
  - We chose coal as a proxy for *04541 Solid fuels*. Unfortunately, just like for *04522 Liquefied hydrocarbons*, we did not find a database that documents these prices. Instead, we use first quarter of 2020 prices, and deflate to 2010.
  - The consumer price for *07221 Fuels and lubricants* is derived by averaging the consumer prices for petrol and gas oil. Both prices are provided by the Directorate General of Energy of the European Commission.

  As explained in Paragraph 7.2.3.1 these prices are, where necessary, converted into the same unit as the specific excises.

7.2.2.4 Aggregating parameters

For two countries, the aggregation level of goods with specific excises, is higher than COICOP level 4.

- *Germany*: German HBS data has information on only 121 commodities. The COICOP categories *011 Food*, *012 Non–alcoholic beverages*, *021 Alcoholic beverages*, and *022 Tobacco products* are groups where further detail is not presented. Especially, the absence of detailed

data on alcoholic beverages and on tobacco products is problematic for the simulation of indirect taxes. As a first step we derive the aggregate expenditures, using the national account statistics, the aggregate excise revenues at COICOP level 4 and the collected parameters (for reference see tab `Weights DE & IT` in the Excel file `ITTv3 parameters.xlsx`). We then apply the within–group expenditure weights to construct weighted averages of the parameter at the required COICOP level. For food and non–alcoholic beverages there is no problem as there is only one VAT–rate for each of these categories, and no excises are modelled for goods belonging to this categories.

– *Italy*: The Italian HBS does not make a distinction between wine types (COICOP codes 02121 and 02122). Hence we calculate specific excise rates and prices for the aggregate as a weighted average of the specific excises and prices of the more detailed goods. Statista reports that sparkling wine accounts for 10% of the Italian wine sales (in euro) in 2010, thus leaving 90% for regular wine. We use these ratio's as weights to calculate the average excises and prices for the aggregated commodity, *0212 Wine*. Given that the products are so similar, and that the excises are expressed identically, we don't have to harmonise the excises first. A similar approach is taken to aggregate excises and prices for tobacco products, which are grouped into one high–level group *022 tobacco products*. Now we use the respective parameters and weights for *02211 cigarettes*, *02212 cigars*, and *02213 rolling tobacco*. As the units are more dissimilar here, one cigar is quite different from one cigarette, we also need to create a 'common product'. For the calculation of the specific excises, we chose to calculate the volume of pure tobacco sold (expressed in tons of pure tobacco), rather than the number of cigarettes and cigars, as it allows us to attach a weight to otherwise dissimilar products.[28]

7.2.3 The TCO policy sheet

The tco policy sheet `tco_cc` is situated in the spine just before defining the EUROMOD output `output_std_cc`, and thus after disposable incomes have been determined. It contains all the indirect tax parameters and functions calculating the expenditures, indirect taxes, and other variables of interest. Below we provide an overview of the structure of this policy sheet.

7.2.3.1 Part 1: Parameter definitions

The first set of functions define the indirect tax and (selected) price parameters. Each commodity in the input dataset must be assigned a VAT rate, and –if applicable– a specific excise rate and an *ad valorem* excise rate. For the specific excise goods consumer prices for the policy year(s) are required as well. For ease of modelling, another set of functions group the parameters into several lists.

– The function `Parameters` contains the parameters for each policy scenario:

[28] Naturally, other approaches can be taken as well.

a. standard and reduced VAT rates `$tco_t_xxx`,
b. the specific excises `$tco_a_[xxxxx]`,
c. *ad valorem* excises `$tco_v_[xxxxx]`.

Note that the construction of EUROMOD functions requires the parameter values for the excise goods to be non–missing. Hence, if an excise good does not have an *ad valorem* or specific excise, that parameter value must be set to 0.

- The function `Parameters (calculations)` contains the necessary mutations, such as taking (weighted) averages of the statutory parameters. It is common that the level of detail on the various commodities in the input data is more limited than the information on statutory excises (e.g. there is only one variable for *Car fuel*, while most countries tax gas oil and petrol differently.)
- The function `Parameters (manipulation & aggregation due to restricted detail of alcoholic beverages and tobacco products)` is only used for two countries (Germany and Italy), which do not have income shares of expenditures at the fourth level of detail. We must therefore create weighted averages of the excises, using country–level aggregate consumption statistics, as discussed in Section 7.2.2.4.
- The functions `Parameters baseline` and `Parameters baseline (calculations)` contain the baseline indirect tax parameters and the parameter manipulations (exclusively for the specific excise goods in the baseline), plus the baseline consumer prices `$tco_base_q_[xxxxx]`. Notice that the placeholders contain `_base`, to differentiate them from the reform parameters. When simulating an indirect tax change, the user must change only the content of the `Parameters` function, the `Parameters (baseline)` function must remain unchanged.
- The function `Parameters vat rates` assigns the various VAT–rates defined in the function `Parameters`–function to each of the commodities.
- The functions labelled as `Parameter list:...` group the parameters for the specific excise goods by parameter type (*e.g.* baseline specific excises, reform specific excises, reform *ad valorem* excises, ...). Having the parameters of each group in a list allows the ITTv3 to use the lists as vectors in operations. Note that all commodities must be included in the VAT-list, and that all specific excise goods must be included in each of the other parameter lists (even if the parameter value is zero).
- The function `Total expenditures COICOP aggregates` contains the simulated and the national account aggregates for the policy year. The simulated baseline and the national account aggregate are used to adjust the simulation results to national account statistics (see Section 7.2.3.7).
- The two functions `Commodity list:...` define two variable lists:
  a. `il_xs`: list containing the income shares of all commodities,
  b. `il_xs_exc`: list containing the income shares of only the excise goods.

7.2.3.2 Part 2: Producer and consumer prices

In this part the producer and consumer prices are calculated for each commodity on which specific excises are levied.

– The function[29] `Producer prices (baseline)` generates the producer prices `$tco_base_p_[xxxxx]` for the specific excise goods, following equation (52), or, using ITTv3 notation,[30]

$$\texttt{\$tco\_base\_p} = \texttt{\$tco\_base\_q}\left(\frac{1}{1+\texttt{\$tco\_base\_t}} - \texttt{\$tco\_base\_v}\right) - \texttt{\$tco\_base\_a}. \quad (80)$$

– The functions `Consumer prices (step 1 - all commodities)` and `Consumer prices (step 2 - excise goods, overwrites prices step 1)` generate the normalised consumer prices.

a. Step 1: all commodities, only taking into account VAT:

$$\texttt{\$tco\_q} = 1 + \texttt{\$tco\_t}. \quad (81)$$

b. Step 2: only for (specific) excise goods, whereby specific and ad valorem excises are applied as well.

$$\texttt{\$tco\_q} = (1+\texttt{\$tco\_t})\,\frac{1+\frac{\texttt{\$tco\_a}}{\texttt{\$tco\_base\_p}}}{1-(1+\texttt{\$tco\_t})\,\texttt{\$tco\_v}}. \quad (82)$$

Note that the output of this function overwrites the results from step 1 for the specific excise commodities.

– The function `Join the xq's in a new income list` groups all consumer prices into a new list.

7.2.3.3 Part 3: Expenditure levels

This part calculates the actual expenditures of a household. As explained earlier, there are three behavioural assumptions for calculating post–reform expenditures: constant income shares of expenditures, constant quantities, and constant expenditure shares. Constant income shares is the default. By adjusting the switches in the `Run EUROMOD` menu one of the other assumptions can be chosen (see Figure 6).

[29] This is the first time the new function type, `IlArithOp`, features. It was developed within this project in order to conduct vector calculations. Without the new function we would require much more repetition in the code, given the nature of the expenditure data and almost 200 commodities.

[30] The COICOP commodity suffix `_[xxxxx]` is dropped for ease of notation.

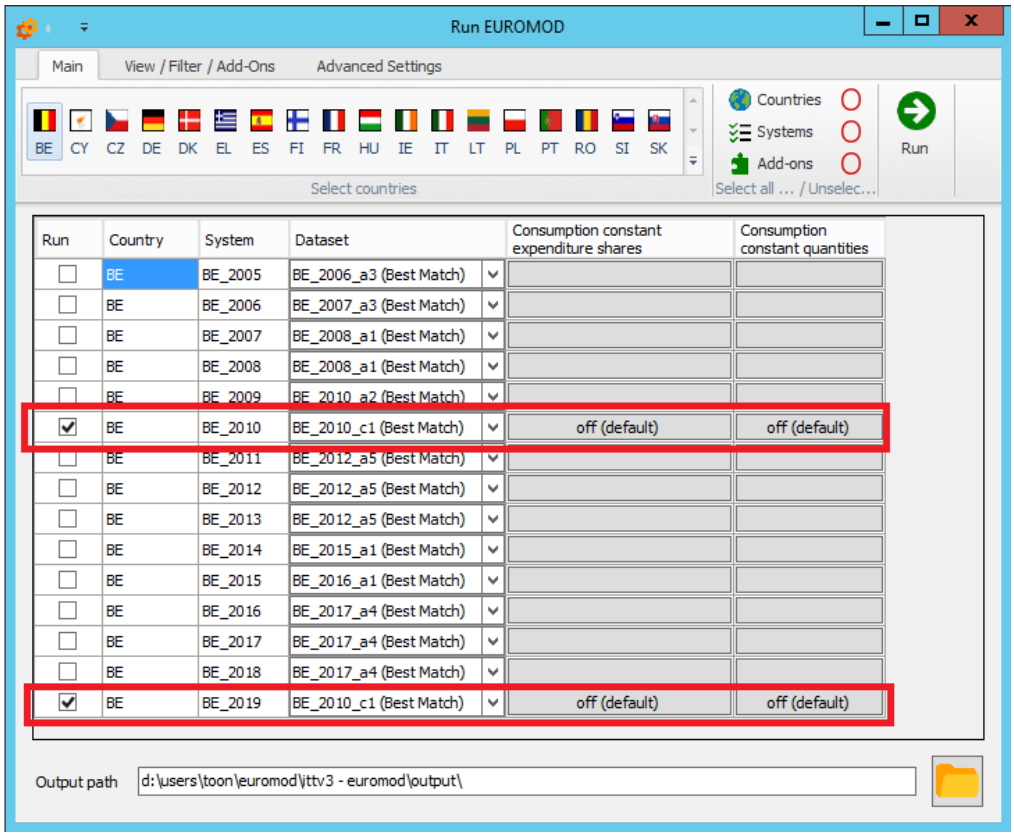

Figure 6: Choosing between the behavioural assumptions

– Three functions, starting with `Define new vars` initialise the required variables.

– The function `Define expenditure levels [CIS]` applies to the constant income share case, and calculates new expenditure levels using the simulated disposable household income, `dispy_hh`, and the imputed income shares from the input data as follows:

$$\texttt{x[xxxxx]} = \texttt{dispy\_hh} \cdot \texttt{xs[xxxxx]}. \tag{83}$$

– The four consequent functions denoted by `[CQ]` simulate expenditures through the constant quantities assumption. These functions must be switched on by the user in the `Run EUROMOD` menu. The ITT merges the data with a support file containing all the quantities consumed in the baseline, under the variable name `xx[xxxxx]` (see 7.2.3.5).

Expenditures are then calculated as follows:

$$\texttt{x[xxxxx]} = \texttt{\$tco\_q\_[xxxxx]} \cdot \texttt{xx[xxxxx]}. \tag{84}$$

– Then eight functions indicated by `[CES]` prepare and execute the calculation of expenditures using constant expenditure shares. Similarly as under the constant quantities assumption, the ITT merges the data with a support file generated by a baseline run. The file contains

the baseline total expenditures and saving, which allow to reconstruct baseline disposable household income as baseline saving plus baseline total expenditures (see 7.2.3.5). In case baseline total expenditures differ from zero, the expenditure shares are derived from the income shares and the baseline income and expenditures. The formula for expenditures thus becomes:

$$\texttt{x[xxxxx]} = \texttt{xs[xxxxx]} \frac{\texttt{baseline income}}{\texttt{baseline expenditures}} \texttt{total expenditures}. \tag{85}$$

In the case where baseline expenditures are zero, the expenditures are calculated using constant income shares instead:

$$\texttt{x[xxxxx]} = \texttt{xs[xxxxx]} \cdot \texttt{total expenditures}. \tag{86}$$

Note that a reform with constant quantities or with constant expenditure shares always requires a baseline run. From this baseline run the baseline quantities (for constant quantities), saving and total expenditures (for constant expenditure shares) are stored for use in the calculations of expenditures after reform.

7.2.3.4 Part 4: Indirect tax liabilities

In this part the code simulates the indirect taxes due per commodity.

- The function `Compute VAT` simulates VAT liabilities as follows:

$$\texttt{tva[xxxxx]} = \texttt{x[xxxxx]} \frac{\texttt{\$tco\_t\_[xxxxx]}}{(1 + \texttt{\$tco\_t\_[xxxxx]})}. \tag{87}$$

- The function `Calculate ad valorem excise` simulates *ad valorem* excise liabilities as follows:

$$\texttt{txv[xxxxx]} = \texttt{x[xxxxx]} \cdot \texttt{\$tco\_v\_[xxxxx]}. \tag{88}$$

  Excise liabilities are only simulated for the selection of excise commodities.

- The function `Calculate specific excise` simulates specific excise liabilities as follows:

$$\texttt{txa[xxxxx]} = \texttt{x[xxxxx]} \frac{\left(\frac{1}{1+\texttt{\$tco\_t\_[xxxxx]}} - \texttt{\$tco\_v\_[xxxxx]}\right)}{(\texttt{\$tco\_base\_p\_[xxxxx]} + \texttt{\$tco\_a\_[xxxxx]})} \texttt{\$tco\_a\_[xxxxx]}, \tag{89}$$

  where the middle term of the product on the right hand side is the inverse of the consumer price (in the same units as the statutory specific excise).
  Excise liabilities are only simulated for the selection of excise commodities.

- The function `IL - Total excises` generates total excise liabilities per excise good, summing up specific and *ad valorem* excises, $\texttt{tx[xxxxx]} = \texttt{txa[xxxxx]} + \texttt{txv[xxxxx]}$.

7.2.3.5 Part 5: Input files to run constant quantities and constant expenditure shares

In this section the ITTv3 produces two support txt-files, required to run simulations using constant quantities and constant expenditure shares. These files can be renamed, as long as the names match with the text values for `File` in the `DefInput` functions.

- The function `Define quantities` simulates quantities as follows:

$$\texttt{xx[xxxxx]} = \frac{\texttt{x[xxxxx]}}{\texttt{\$tco\_q\_[xxxxx]}}. \tag{90}$$

- Next, the function `Create support output file [CQ]` creates a file `cc_yyyy_tco_cq.txt`, which contains the simulated quantities `x[xxxxx]`.
- Next, saving is determined as the residual of total household disposable income and total expenditures.
- The function `Create support output file [CES]` creates the second support file `cc_yyyy_tco_ces.txt`, for simulations using constant expenditure shares. The support file contains household saving and total expenditure (the sum of which equals household disposable income).

7.2.3.6 Part 6: Stone price index

In order to analyse the impact of pure price changes on consumer welfare (neglecting labour supply effects), household disposable income divided by a Stone price index can be used as a welfare measure of consumption (see Section 8.4). The ITT computes the Stone price index, $P_w(\mathbf{q})$ (`xpi` in ITT language), using the income shares of expenditures, w_k:

$$P_w(\mathbf{q}) = \prod_k (q_k)^{w_k}. \tag{91}$$

Recall that these income shares are household specific, and thus, so is the Stone price index.

This is implemented in two steps, as EUROMOD cannot calculate the product of multiple variables.

- Step 1. Log transformation of Equation (91):

$$\texttt{il\_price\_index\_tmp} = \sum \texttt{xs[xxxxx]} \cdot \ln(\texttt{\$tco\_t\_[xxxxx]}). \tag{92}$$

- Step 2. Taking the exponent of Equation (92):

$$\texttt{xpi} = \exp(\texttt{il\_price\_index\_tmp}). \tag{93}$$

7.2.3.7 Part 7: National accounts adjustment

The final part of the `tco` policy sheet consists of functions to perform the national account calibration. In this calibration phase, household expenditures and indirect tax are multiplied with a factor, such that in the baseline the new adjusted expenditures sum up to the national account total at the population level.

The goal is to provide aggregate statistics that adjust for the discrepancies that occur between the simulated expenditure levels and the corresponding national accounts levels of the simulated policy year. The calibration is done by multiplying the expenditure and indirect tax variables with the ratio of the national accounts aggregate and the corresponding simulated levels. The resulting variables are indicated with suffix `_na`. In the current set–up the level of aggregation that is used, is the first level of COICOP aggregation (12 commodity groups). However, the user can choose a level of aggregation that is more detailed. This would come at the cost of having to add considerably more code.

For each simulated expenditure and corresponding tax liability variable, a new NA–adjusted variable is created as follows (using expenditure as an example):

$$\texttt{x[xxxxx]\_na} = \texttt{x[xxxxx]} \frac{\texttt{national account COICOP aggregate}}{\texttt{simulated COICOP aggregate}}. \tag{94}$$

These outcomes are also stored at the level of the 12 COICOP commodities. For example not–adjusted and adjusted expenditures on *food and non–alcoholic beverages* are stored in the variable (and income list) `il_x01` and `il_x01_na`.

7.2.4 Updates to the Statistics Presenter

The Statistics Presenter (SP) has been extended with a module that provides a standard analysis of the ITTv3 output. It provides both budgetary and distributional statistics, and statistics by major COICOP category. The SP does not calculate baseline–reform output statistics, contrary to what it does for direct tax analyses.

To start, the user needs to select `Indirect Taxes Analysis` in the SP menu (see Figure 7).

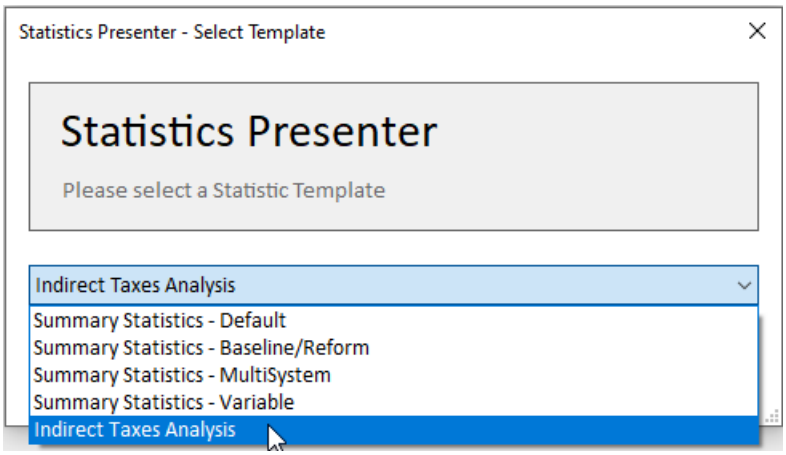

Figure 7: New Statistics Presenter module

There are eight output tables, with a sole focus on the output generated by the `tco` policy sheet.

– `Tabl 1 - Budget / totals` returns the aggregated monthly and yearly results for expenditures and indirect taxes (Figure 8). Both, simulated figures and figures after adjustment to national account aggregates (see Section 7.2.3.7) are reported

Table 1 - Budget / totals
in million

	Monthly	Annual	Monthly (macro-adjusted)	Annual (macro-adjusted)
Expenditures	13,295.30	159,543.55	16,172.35	194,068.18
Indirect taxes	1,898.23	22,778.76	2,345.96	28,151.50
VAT	1,523.05	18,276.54	1,834.23	22,010.74
Specific excises	336.50	4,038.06	423.23	5,078.79
Ad valorem excises	38.68	464.16	88.50	1,061.98

Figure 8: SP - Table 1

– `Table 2 - Budget / totals` returns a more detailed table with the aggregated monthly figures for expenditures and indirect taxes, by aggregate COICOP category. It also reports the national account adjusted figures (Figure 9).

Table 2 - Budgetary totals per COICOP aggregate category
annual, in million

	Expenditures	Indirect taxes	VAT	Specific excises	Ad valorem excises	Expenditures (macro-adjusted)	Indirect taxes (macro-adjusted)	VAT (macro-adjusted)	Specific excises (macro-adjusted)	Ad valorem excises (macro-adjusted)
01 Food, non-alcoholic beverages	24,833.17	1,414.17	1,414.17	0.00	0.00	27,847.07	1,585.81	1,585.81	0.00	0.00
02 Alcoholic beverages and tobacco	3,855.43	1,853.67	669.12	720.38	464.16	8,821.13	4,241.13	1,530.94	1,648.21	1,061.98
03 Clothing and footwear	8,029.64	1,363.80	1,363.80	0.00	0.00	10,049.20	1,706.82	1,706.82	0.00	0.00
04 Housing and utilities	25,724.47	3,260.30	2,602.50	657.80	0.00	26,266.73	3,329.03	2,657.36	671.67	0.00
05 Furnishings and household equipment	11,259.72	1,954.16	1,954.16	0.00	0.00	13,884.55	2,409.72	2,409.72	0.00	0.00
06 Health	8,666.60	184.97	184.97	0.00	0.00	13,816.42	294.89	294.89	0.00	0.00
07 Transport	25,689.19	6,836.71	4,176.84	2,659.88	0.00	26,645.68	7,091.27	4,332.36	2,758.91	0.00
08 Communication	4,818.86	807.53	807.53	0.00	0.00	5,278.94	884.63	884.63	0.00	0.00
09 Recreation and culture	15,165.88	2,157.63	2,157.63	0.00	0.00	19,511.97	2,775.95	2,775.95	0.00	0.00
10 Education	644.95	0.00	0.00	0.00	0.00	849.41	0.00	0.00	0.00	0.00
11 Restaurants and hotels	11,204.83	1,299.91	1,299.91	0.00	0.00	12,093.57	1,403.02	1,403.02	0.00	0.00
12 Miscellaneous goods and services	19,650.81	1,645.90	1,645.90	0.00	0.00	29,003.50	2,429.25	2,429.25	0.00	0.00
All	**159,543.55**	**22,778.76**	**18,276.54**	**4,038.06**	**464.16**	**194,068.18**	**28,151.50**	**22,010.74**	**5,078.79**	**1,061.98**

Figure 9: SP - Table 2

– `Tabl 3A` and `Table 3B` present the average expenditures and indirect tax liabilities by decile, where the deciles are based on equivalent disposable household income, and by equivalent household expenditures respectively, using modified OECD equivalence scales (Figure 10 and Figure 11).

Table 3.A - Average household expenditures & indirect taxes per income decile
monthly

	Interval eq.income deciles	Income	Expenditures	VAT	Specific excises	Ad valorem excises
1	1,132.50	1,267.87	1,727.12	174.56	48.26	8.54
2	1,357.71	1,720.85	1,784.46	183.29	46.17	6.91
3	1,550.61	2,119.76	2,159.47	232.65	55.32	8.27
4	1,748.09	2,577.06	2,659.48	299.46	71.03	8.66
5	1,931.91	2,957.08	2,910.35	334.22	72.15	9.38
6	2,122.56	3,331.04	3,094.41	359.82	81.79	11.36
7	2,338.11	3,809.43	3,481.60	410.71	91.01	9.02
8	2,598.34	4,162.02	3,547.66	435.55	85.55	8.22
9	3,029.92	4,548.88	3,644.33	432.61	89.10	8.07
10		6,338.77	4,172.77	495.62	96.82	4.75
All		**3,173.77**	**2,845.72**	**325.99**	**72.03**	**8.28**

Figure 10: SP - Table 3A

Table 3.B - Average household expenditures & indirect taxes per expenditure decile
monthly

	Interval eq.expenditure deciles	Income	Expenditures	VAT	Specific excises	Ad valorem excises
1	884.91	2,009.55	1,065.88	103.19	29.93	4.26
2	1,086.32	2,540.92	1,577.01	158.86	46.94	5.55
3	1,251.65	2,708.57	1,813.52	189.40	54.50	6.27
4	1,397.88	2,945.65	2,119.68	219.61	61.04	7.57
5	1,567.56	3,258.56	2,359.97	253.48	71.74	9.49
6	1,739.06	3,230.79	2,604.54	282.83	75.22	9.82
7	1,946.12	3,349.28	2,858.72	311.97	74.48	10.98
8	2,255.95	3,633.72	3,289.13	372.40	95.49	10.34
9	2,882.94	3,804.37	3,925.60	460.29	95.85	9.91
10		4,178.57	6,587.58	871.50	111.94	8.49
All		**3,173.77**	**2,845.72**	**325.99**	**72.03**	**8.28**

Figure 11: SP - Table 3B

– `Table 4A` and `Table 4B` present the average monthly expenditures by aggregate COICOP category, and by equivalised disposable household income decile and by equivalised disposable expendtiures (Figure 12 and Figure 13).

Table 4.A - Average household expenditures per COICOP aggregate category per income decile
monthly

	01 Food, non-alcoholic beverages	02 Alcoholic beverages and tobacco	03 Clothing and footwear	04 Housing and utilities	05 Furnishings and household equipment	06 Health	07 Transport	08 Communication	09 Recreation and culture	10 Education	11 Restaurants and hotels	12 Miscellaneous goods and services	Total expenditures
1	320.84	55.07	65.24	475.00	104.23	83.80	166.97	59.75	125.53	2.60	82.15	185.96	1,727.12
2	333.89	49.19	67.44	459.70	104.35	100.50	143.71	67.61	154.02	6.05	98.48	199.51	1,784.46
3	382.03	62.51	83.97	473.99	127.35	127.95	251.33	72.47	184.98	4.37	118.18	270.35	2,159.47
4	440.72	66.07	122.03	454.85	155.65	166.18	456.41	85.83	211.29	9.38	162.73	328.34	2,659.48
5	457.17	66.14	154.81	437.33	215.90	163.85	491.67	86.66	261.07	15.89	193.01	366.86	2,910.35
6	473.42	79.27	162.04	450.70	204.78	164.05	574.56	95.34	294.48	10.01	224.41	361.37	3,094.41
7	502.98	81.20	176.67	464.98	267.54	194.66	620.94	99.24	341.58	27.13	266.10	438.57	3,481.60
8	501.56	72.65	218.67	439.65	267.28	164.58	659.47	101.07	374.11	11.14	261.53	475.95	3,547.66
9	505.70	79.85	192.40	465.76	262.79	192.23	706.84	97.69	393.02	13.82	295.04	439.16	3,644.33
10	582.49	86.49	243.50	460.81	368.80	222.77	711.96	107.92	454.72	20.21	373.03	540.08	4,172.77
All	**442.94**	**68.77**	**143.22**	**458.84**	**200.84**	**154.58**	**458.21**	**85.95**	**270.51**	**11.50**	**199.86**	**350.50**	**2,845.72**

Figure 12: SP - Table 4A

Table 4.B - Average household expenditures per COICOP aggregate category per expenditure decile
monthly

	01 Food, non-alcoholic beverages	02 Alcoholic beverages and tobacco	03 Clothing and footwear	04 Housing and utilities	05 Furnishings and household equipment	06 Health	07 Transport	08 Communication	09 Recreation and culture	10 Education	11 Restaurants and hotels	12 Miscellaneous goods and services	Total expenditures
1	258.88	27.98	31.75	269.32	42.50	53.97	65.84	50.34	75.87	2.30	58.40	128.73	1,065.88
2	345.85	40.05	64.95	366.64	67.54	89.65	115.61	65.37	115.27	6.71	84.97	214.40	1,577.01
3	372.62	45.45	84.54	375.77	90.28	103.28	152.87	73.91	154.75	6.81	101.67	251.58	1,813.52
4	412.01	51.52	117.61	421.25	117.78	135.38	164.40	75.96	197.87	11.30	137.36	277.23	2,119.68
5	437.64	68.64	138.62	422.16	153.11	141.63	213.56	81.00	209.10	9.12	166.14	319.25	2,359.97
6	474.57	74.09	144.41	466.19	168.25	152.80	231.63	87.92	251.62	13.90	189.87	349.31	2,604.54
7	489.78	84.72	162.05	501.73	176.08	165.60	297.45	95.73	283.79	12.92	199.07	389.79	2,858.72
8	513.85	90.26	183.80	511.66	211.68	177.81	419.39	99.44	362.85	15.41	261.20	441.79	3,289.13
9	545.78	97.77	212.66	574.38	344.45	220.80	543.95	113.25	446.04	15.12	347.06	464.35	3,925.60
10	567.22	103.79	281.40	663.31	608.90	294.81	2,269.45	113.81	583.24	20.80	434.52	646.32	6,587.58
All	**442.94**	**68.77**	**143.22**	**458.84**	**200.84**	**154.58**	**458.21**	**85.95**	**270.51**	**11.50**	**199.86**	**350.50**	**2,845.72**

Figure 13: SP - Table 4B

– `Table 5A` and `Table 5B` present the average indirect tax liability per month by decile and by aggregate COICOP category (Figure 14 and Figure 15).

Table 5.A - Average indirect taxes paid by households per COICOP aggregate category per income decile
monthly

	01 Food, non-alcoholic beverages	02 Alcoholic beverages and tobacco	03 Clothing and footwear	04 Housing and utilities	05 Furnishings and household equipment	06 Health	07 Transport	08 Communication	09 Recreation and culture	10 Education	11 Restaurants and hotels	12 Miscellaneous goods and services	Total indirect taxes
1	18.28	28.48	10.96	46.39	18.09	2.22	53.82	9.95	17.56	0.00	10.22	15.39	231.36
2	19.06	25.31	11.31	51.33	18.11	2.43	48.58	11.38	21.90	0.00	12.15	14.81	236.37
3	21.78	31.09	14.21	56.06	22.10	2.92	72.89	12.18	27.04	0.00	13.95	22.01	296.23
4	25.10	32.99	20.78	59.54	27.01	3.60	121.18	14.40	29.12	0.00	19.24	26.19	379.16
5	26.03	33.29	26.35	55.25	37.47	3.31	128.89	14.60	36.48	0.00	22.95	31.12	415.76
6	26.95	39.86	27.44	54.35	35.54	3.23	151.11	15.92	42.39	0.00	25.91	30.28	452.98
7	28.63	38.10	30.12	65.21	46.43	3.86	166.00	16.65	48.82	0.00	30.60	36.32	510.75
8	28.54	33.93	37.29	66.08	46.39	3.24	167.81	17.07	54.45	0.00	30.65	43.88	529.33
9	28.77	36.23	32.88	63.68	45.61	3.86	175.72	16.32	55.97	0.00	33.43	37.31	529.78
10	33.14	35.25	41.25	68.74	64.01	4.85	181.18	17.93	63.97	0.00	41.15	45.72	597.18
All	**25.22**	**33.06**	**24.33**	**58.15**	**34.86**	**3.30**	**121.94**	**14.40**	**38.48**	**0.00**	**23.19**	**29.36**	**406.30**

Figure 14: SP - Table 5A

Table 5.B - Average indirect taxes paid by households per COICOP aggregate category per expenditure decile
monthly

	01 Food, non-alcoholic beverages	02 Alcoholic beverages and tobacco	03 Clothing and footwear	04 Housing and utilities	05 Furnishings and household equipment	06 Health	07 Transport	08 Communication	09 Recreation and culture	10 Education	11 Restaurants and hotels	12 Miscellaneous goods and services	Total indirect taxes
1	14.76	14.50	5.28	32.80	7.38	1.15	27.40	8.44	10.72	0.00	6.89	8.06	137.37
2	19.71	20.21	10.98	44.80	11.72	2.32	47.99	10.99	16.03	0.00	10.51	16.07	211.34
3	21.23	23.47	14.27	47.11	15.67	2.62	60.03	12.50	21.23	0.00	12.44	19.61	250.17
4	23.47	26.36	19.93	52.30	20.44	2.66	65.53	12.79	27.40	0.00	16.81	20.53	288.22
5	24.93	34.29	23.65	53.06	26.57	3.04	81.03	13.68	29.16	0.00	20.22	25.06	334.71
6	27.04	36.71	24.72	60.16	29.20	3.34	84.97	14.79	35.36	0.00	22.45	29.13	367.87
7	27.89	41.18	27.67	60.62	30.56	3.37	93.60	16.14	41.10	0.00	23.32	31.97	397.42
8	29.26	42.46	31.21	64.73	36.74	3.94	133.12	16.68	51.91	0.00	30.41	37.79	478.23
9	31.05	45.21	35.84	76.27	59.78	4.45	149.81	18.85	62.99	0.00	39.46	42.33	566.05
10	32.26	44.95	47.93	87.40	105.68	5.92	455.70	18.73	85.27	0.00	47.45	60.63	991.92
All	**25.22**	**33.06**	**24.33**	**58.15**	**34.86**	**3.30**	**121.94**	**14.40**	**38.48**	**0.00**	**23.19**	**29.36**	**406.30**

Figure 15: SP - Table 5B

All tables can also be saved in an excel-file.

8 Example policy reform simulation

This section presents a hypothetical policy simulation using ITTv3, intending to explain how the tool functions. We analyse how a negative, heterogeneous income shock affects Belgian households, and to what extent lowering VAT would alleviate some of its negative effects.[31] We focus here on how different behavioural assumptions may lead to a different assessment of the global impact of price and income changes.

Using the to 2019 uprated Belgian EU–SILC 2010 data, which were enriched with income shares of expenditures, we simulate a baseline and the effect of two changes in the environment, the first being an economic shock, and the second the same shock in combination with a government intervention. The baseline is the 2019 EUROMOD system for Belgium, containing the actual direct and indirect tax and benefit rules for 2019. In the first scenario of a changing environment, a shock in labour income is simulated. This is done through adjusting the labour incomes in the EUROMOD input dataset (SILC enriched with income shares of expenditures). In the second scenario, the same income shock is applied, but now the government reduces the VAT rates. We explain in detail how we set up EUROMOD to simulate this reform, and how the behavioural assumptions play a role in the simulation of expenditures and indirect taxes.

8.1 Baseline simulation

The baseline is the default EUROMOD policy system for 2019. Since we use the Belgian 2010 EU–SILC, containing 2009 incomes. All incomes are uprated to reflect the observed average increases in market incomes and social benefits between 2009 and 2019. Other household characteristics are kept unchanged. The EUROMOD direct tax calculator then produces 2019 household disposable incomes.[32] As explained in Section 7.1.2.1, baseline expenditure levels for each commodity are constructed by multiplying the simulated household disposable incomes with the imputed income shares. We retain the commodity level of detail from the EUROSTAT Household Budget Survey, which is generally at the 4^{th} COICOP level. Next, ITTv3 calculates the tax liabilities, using the 2019 statutory tax parameters at the level of each commodity, and the 2019 consumer prices for the goods subject to specific excises, as explained in detail in Section 7.1 (Equations 39–41).

The model also computes the implicit indirect tax rate for each commodity k, τ_k (see Section 7.1, Equation 37 and 44), and uses these rates to derive quantities consumed (measured in monetary terms at producer prices) of each commodity k for every household h, by dividing the household expenditures on good k by $1 + \tau_k$ (See Section 7.1, Equation 45). Total household expenditures are

[31] The backdrop of the COVID 19 pandemic served as an inspiration for the simulation. However, it is by no means our intention to simulate the actual socio–economic shock that occurred in Belgium in 2020.

[32] The uprating of gross incomes and the simulation of household disposable incomes are explained in more detail in the EUROMOD Country Report for Belgium 2016–2019 (`https://www.euromod.ac.uk/sites/default/files/country-reports/year10/Y10_CR_BE_Final_.pdf`).

calculated by summing expenditures on all individual commodities, and saving is the residual of disposable income and total expenditures (sees Section 7.1, Equations 47 and 48). Next, ITTv3 stores the quantities, total expenditures and saving, as the tool will require these data to run the reform simulations under the behavioural assumptions of constant quantities and constant expenditure shares (see Section 7.1.2). We will discuss this below.

The ITTv3 baseline run produces a new household level EUROMOD output file, that has identical output as that of a standard EUROMOD run, but expanded with a large number of ITTv3–related variables. More specifically, the output file contains for each household and each commodity, the expenditures, VAT and excise liabilities, implicit indirect tax rates, quantities, and if one chooses to, the tax parameters.

Aggregating the outcomes using the population weights yields a number of macro–statistics. For our baseline run we present them in Table 10.[33] Annual gross earned incomes amount to 184.5 billion euros. Direct taxes sum up to 48.9 billion euros, whereas social security contributions (exclusive of employer contributions) totalled 24.2 billion euros. All social benefits, including pensions, amounted to 66.4 billion euros. Disposable income and total household expenditures totalled 177.8 billion and 159.4 billion euros respectively. Saving amounts to 18.4 billion euros. Indirect tax revenues are 22.8 billion euros (14.3% of total expenditures) of which 80.1% levied through VAT, 17.8% through specific excises, and 2% through *ad valorem* excises.

Table 10: Macro indicators at the baseline (Annual, in million Euros)

Gross earned income	184 450.6
Direct taxes	48 866.3
Social security contributions	24 166.7
Benefits	66 405.8
Disposable household income	177 823.3
Total expenditures	159 445.7
Saving	18 377.6
VAT revenues	18 264.0
Ad valorem Excises	464.2
Specific excises	4 065.0

8.2 Income shock

8.2.1 Description of the simulated income shock

In our first scenario, we introduce a labour income shock in order to show how a change in disposable income affects expenditures and indirect tax revenues. We also discuss how the income shock will

[33] The figures in the table are grossed up at population level, but not calibrated to national account statistics (see Section 7.2.3.7 for the latter option).

induce different responses depending on the assumptions regarding the demand responses.

The current economic setting of the Covid–19 pandemic inspired our design of the income shock. Starting March 2020, the Belgian government implemented a wide range of measures to limit the spread of the Coronavirus. These containment and mitigation measures, together with heightened caution by households, have led to many businesses experiencing very large income drops, or being shut down altogether. In this exercise, we simulate such a negative income shock. Our shock has two components:

1. a loss of the entire labour income that affects a small subgroup of people who are not covered by the social security system, and thus don't receive a replacement income or other type of special Covid–19 related government support. If eligible, these persons may become entitled to the social assistance scheme;
2. a 20% decrease in labour income for those who temporarily lose their job and are eligible to unemployment benefits or some other type of special Covid–19 related government support. Note that this is a strong simplification of how the labour income has been compensated through either unemployment benefits or other compensation schemes.

The people in the first group work in non–agricultural sectors, under a contract that only grants them limited social security rights and are not entitled to unemployment benefits nor to government subsidies implemented to compensate Covid–19 related job losses. We estimate this group of people to account for 2.2% of the workforce (approximately 1% of the total Belgian population).[34]

The second component of the income shock hits a randomly selected sub–sample of people working in each employment sector. Covid–19 related employment and revenue losses vary, however, substantially among sectors. To account for this asymmetry, the second component of the income shock has a sector specific nature. The size of the group receiving the shock depends on the job loss estimates presented in Table 11 (based on Decoster *et al.*, 2020). In Belgium, employment in *hotels and restaurants* got hit the hardest: Decoster *et al.* (2020) report that 83% of the people employed in that sector were temporarily unemployed. The second hardest hit sector is *wholesale and retail* (42%), followed by *transport and communication* (37%). However, not every sector experienced temporary unemployment. Employment levels did not change in *agriculture*, *public administration and defence*, *education*, and *health and social work*.

Following this information, we randomly draw a number of people working in each sector in accordance to the statistics presented in Decoster *et al.* (2020). We consequently reduce the labour income of these persons by 20%, leaving other characteristics unchanged. Overall, 957 000 people (approximately 9% of the total Belgian population) were subjected to such a negative income shock, accounting for almost 21% of the workforce. We assume the shock to last an entire year.

[34] By workforce we mean the number of people declaring positive earned income, *i.e.* an estimate of the number of people working.

Table 11: Temporary unemployment by sector (% of workforce in sector)

1	Agriculture and fishing	0
2	Mining, manufacturing and utilities	21
3	Construction	26
4	Wholesale and retail	42
5	Hotels and restaurants	83
6	Transport and communication	37
7	Financial intermediation	3
8	Real estate and business	21
9	Public administration and defence	0
10	Education	0
11	Health and social work	0
12	Other	27

Source: Decoster *et al.* (2020) and own calculations.

8.2.2 Distributional pattern of the income shock

In this section we identify the characteristics of the people and households who are hit by the income shock. In the first and fourth column of Table 12 we show how the working population is distributed over deciles of equivalised disposable household incomes (summing up to 100%). *E.g.* 3.15% of the people working (declaring positive earned income) belong to the poorest 10% of the population in terms of equivalised disposable income, while 15.73% belong to the richest 10%, and 4.45% of the households in which at least one person is working (earning positive income) belong to the poorest 10% of the households in terms of equivalised disposable income. Next, the table shows how the affected individuals and affected households are distributed across the deciles (again, summing up to 100%). Compared to where we find working people in the overall income distribution deciles (first columns Table 12), the people who lose all of their labour income are predominantly found in the bottom three deciles. Making the same comparison for those who lose 20% of their income, we see a similar over–representation in the six lowest deciles. From the seventh decile on there is under–representation. Such a monotonous pattern was not found for those who lose all their labour income.

8.2.3 Simulating the effects of the income shock with ITTv3

The loss of labour income leads to a decrease in disposable income of on average 1.8%. This income loss will affect expenditures. How it affects expenditures depends on behavioural assumptions. ITTv3 implements three possibilities (Section 7.1.2.2.2): constant income shares (CS_Y), which implies unit income elasticities for all households, constant quantities (CQ), which implies zero income elasticities, and constant expenditure shares (CS_E), with income elasticities greater (smaller) than one for households that save (dissave).

Table 13 summarises the theoretical predictions each of these behavioural assumptions generates at

Table 12: Share of people in each decile

	Individuals			Households		
Decile	**Work-force**	**100% Income loss**	**20% Income loss**	**Work-force**	**100% Income loss**	**20% Income loss**
1	3.15	8.94	5.69	4.45	9.17	6.08
2	3.84	9.60	5.84	4.93	10.82	6.27
3	5.43	8.34	6.29	6.91	9.4	6.57
4	8.12	7.56	9.91	9.16	8.52	9.26
5	9.86	8.74	10.56	10.84	9.42	9.93
6	11.44	14.80	13.42	11.89	10.4	13.38
7	12.76	13.22	11.47	12.01	14.37	11.62
8	14.25	8.86	12.49	12.87	9.33	12.69
9	15.42	3.25	12.00	13.59	3.1	11.93
10	15.73	16.67	12.33	13.35	15.47	12.26
All	100.00	100.00	100.00	100.00	100.0	100.00
All	4 512 362	103 352	957 312	2 939 855	91 742	869 200
% of the total population	42.28	0.97	8.97	62.92	1.96	18.6

Note: Deciles are based on equivalised household disposable income.

the individual level as a response to a negative income shock. Not everybody has been subjected to such a shock, but negative income shock is the only change that occurs in this simulation. So, the theoretical predictions can serve as a guideline for cross–checking the simulation results via the aggregate statistics presented in Table 14.

Table 13: Summary of theoretically determined effects of an income shock at the individual level

Change in	**Negative income shock** $(-\Delta y^h)$		
Variable	CS_Y	CQ	CS_E
Quantity $(\widetilde{x}_k^h)$	$\searrow$	0	$\searrow$
Expenditures (e_k^h, E^h)	$\searrow$	0	$\searrow$
Saving (S^h)	$\searrow$ if $S^h > 0$ $\nearrow$ if $S^h < 0$	$\searrow$	0
VAT liability $(T_{l_k}^h)$	$\searrow$	0	$\searrow$
Ad valorem excise liability $(T_{v_k}^h)$	$\searrow$	0	$\searrow$
Specific excise liability $(T_{a_k}^h)$	$\searrow$	0	$\searrow$

8.2.3.1 The assumption of constant income shares

Since CS_Y exhibits unit income elasticities, a proportional change in disposable income is reflected in an equiproportional change of expenditure levels and saving at the individual household level. A downward adjustment in expenditures at the aggregate level takes place once disposable income is reduced by the negative income shock (see the total expenditures line in columns 1 and 2 of Table 14).

Table 14: Simulated aggregates in response to income shock (Annual, in million euros)

	Baseline	**Income shock**		
		CS_Y	CQ	CS_E
Gross earned income	184 450.6	177 653.3	177 653.3	177 653.3
Direct taxes	48 866.3	46 444.9	46 444.9	46 444.9
Social security contributions	24 166.7	22 942.9	22 942.9	22 942.9
Benefits	66 405.8	66 487.0	66 487.0	66 487.0
Disposable household income	177 823.3	174 752.4	174 752.4	174 752.4
Total expenditures	159 445.7	156 789.2	159 445.7	156 374.9
Saving	18 377.6	17 963.2	15 306.7	18 377.6
VAT revenues	18 264.0	17 962.4	18 264.0	17 922.2
Ad valorem Excises	464.2	456.1	464.2	455.3
Specific excises	4 065.0	3 996.5	4 065.0	3 980.6

As a result of the decrease in expenditures, indirect tax liabilities are lowered as well.

Aggregate saving turns out to decrease under CS_Y. This is an empirical result, as the individual effect of a negative income shock on saving depends on whether baseline saving is negative (dissaving decreases, so saving increases) or positive (saving decreases) (see Table 13).

The CS_Y assumption is the ITTv3's default simulation option, and does not require any action by the user other than changing the reform parameters, just like one would do in a standard reform simulation in EUROMOD.

8.2.3.2 The assumption of constant quantities

Under the CQ assumption, changes in disposable income will lead to a one–to–one change in household saving. Since, at this stage, prices are still unaffected, total expenditures remain constant. The decrease in disposable income thus results in lower saving, in order to keep consumption quantities at their initial levels, and has no impact on the expenditure (compare the saving and total expenditures lines of columns 1 and 3 in Table 14). Indirect tax liabilities do not change either.

To run simulations of reforms under the CQ assumption, users should switch on the `constant quantities`–switch in the `Run` menu. This will activate a set of functions in ITTv3 that will import and merge the quantities for all commodities from the baseline run. ITTv3 will consequently apply the new indirect tax rules of the reform scenario to the constant quantities and calculate new total expenditures (see item 2 in Section 7.1.2.2.2). After this step, ITTv3 uses the standard functions to calculate the indirect tax liabilities (see Equations 39–41 in Section 7.1.1), using the new expenditure levels and new tax parameters. Since in the present case, expenditures and also indirect tax rules have not changed, it stands to reason that indirect tax liabilities remain unaffected by the income shock, as reported in Table 14 (compare the tax revenue lines of columns 1 and 3).

8.2.3.3 The assumption of constant expenditure shares

Under CS_E, nominal saving remains constant at the baseline level. A change in disposable income is fully reflected in expenditure levels: the difference between disposable income in the first and fourth column of Table 14 is equal to the difference in total expenditures. Saving remains unchanged with respect to the baseline.

To run simulations of reforms under the CS_E assumption, users should switch on the `constant expenditures shares`–switch in the `Run` menu. This will import and merge household saving and total expenditures from the baseline run. Keeping saving constant, the ITTv3 calculates new total household expenditures by subtracting the baseline saving from the new disposable incomes. Then it adjusts the income shares in the input data as explained in item 3 of Section 7.1.2.2.2, to retrieve the expenditure shares, which in turn are multiplied with the newly simulated total expenditures to arrive at our new expenditures for each commodity. After this step, ITTv3 uses the standard set of functions to calculate the indirect tax liabilities (see Equations 39–41 in Section 7.1.1), using the new expenditure levels and new tax parameters. These are reported at the bottom lines of the most right column of Table 14.

8.3 VAT rate reduction

As a next step, we introduce an indirect tax policy change to study the effects of price changes. More in particular, we simulate a cut in VAT rates. The reform we simulate is motivated by a recent policy change implemented in Germany. In June 2020, Germany reduced the standard VAT rate from 19 to 16 percent between July 1st and December 31st. Furthermore, the reduced rate, which applies to many food articles and everyday goods, is dropped from 7 to 5 percent. The tax rate reduction is expected to be fully reflected in consumer prices aiming to increase purchasing power of low–income earners and spur consumption. To imitate the German policy change in the Belgian context, we decrease the standard VAT rate from 21% to 19% and the reduced rate from 6% to 4%, and we keep the 12% rate unchanged.

In order to asses the effects of the VAT rate reduction separately from that of the income shock, we did two simulations. The first one takes the after shock incomes (that is the figure reported on line 1 in columns 2–4 of Table 14) as a baseline and applies the lower VAT rates as a reform under the three behavioural hypotheses. The theoretical predictions of lowering VAT–rates at the individual level are summarised in Table 15. Notice that predictions for CS_Y and CS_E are the same. This is because both assumptions exhibit the same price elasticities (own price elasticities are *minus one* and cross price elasticities are all zero) and because expenditure levels remain unaffected by price changes under both assumptions. Therefore, the predictions under CS_Y and CS_E are not only qualitatively the same, but also quantitatively.

Notice that the reduced VAT rate policy applies to everyone, irrespective of whether one is hit by the income shock or not. Therefore, in the case where we isolate the price effect from the income effect,

Table 15: Summary of theoretically determined effects of a VAT rate reduction at the individual level

Change in	**VAT rate reduction** $(-\Delta t_k)$		
Variable	CS_Y	CQ	CS_E
Quantity ($\widetilde{x}_k^h$)	↗	0	↗
Expenditures (e_k^h, E^h)	0	↘	0
Saving (S^h)	0	↗	0
VAT liability ($T_{t_k}^h$)	↘	↘	↘
Ad valorem excise liability ($T_{v_k}^h$)	0	↘	0
Specific excise liability ($T_{a_k}^h$)	↗	0	↗

we would expect no difference between CS_Y and CS_E. This is confirmed by the figures reported in columns 2 and 4 of Table 16. Under CQ, a reduction of VAT rates will reduce expenditures as the constant quantities become cheaper. Therefore, not only VAT liabilities, but also *ad valorem* excise liabilities will drop. As compared to the baseline, specific excises are never affected under constant quantities.

Table 16: Simulated aggregates in response to the VAT rate reduction (Annual, in million euros)

	Baseline is income after shock	**Reduced VAT with incomes after shock as baseline**		
		CS_Y	CQ	CS_E
Gross earned income	177 653.3	177 653.3	177 653.3	177 653.3
Direct taxes	46 444.9	46 444.9	46 444.9	46 444.9
Social security contributions	22 942.9	22 942.9	22 942.9	22 942.9
Benefits	66 487.0	66 487.0	66 487.0	66 487.0
Disposable household income	174 752.4	174 752.4	174 752.4	174 752.4
Total expenditures	156 789.2	156 789.2	154 632.3	156 789.2
Saving	17 963.2	17 963.2	20 120.1	17 963.2
VAT revenues	17 962.4	16 080.7	15 819.4	16 080.7
Ad valorem Excises	456.1	456.1	442.2	456.1
Specific excises	3 996.5	4 066.1	3 996.5	4 066.1

We now examine the combined effect of a VAT rate lowering and a negative income shock. Theoretical predictions presented in Table 17 are derived from combining the results Tables 13 and 15: if changes point in the same direction or combine a zero change with a non–zero change, the joint effect can be signed. Recall that not everybody is subject to the income shock in our micro–simulation, and thus, for some individuals only the effect of lowering VAT rates apply. The aggregate effects are presented in Table 18.

Table 17: Summary of theoretically determined combined effects of income shock and VAT rate reduction at the individual level

Change in	**Combination of negative income shock and VAT rate reduction**		
Variable	CS_Y	CQ	CS_E
Quantity ($\tilde{x}_k^h$)	?	0	?
Expenditures (e_k^h, E^h)	↘	↘	↘
Saving (S^h)	↘ if $S^h > 0$ ↗ if $S^h < 0$	?	0
VAT liability ($T_{l_k}^h$)	↘	↘	↘
Ad valorem excise liability ($T_{v_k}^h$)	↘	↘	↘
Specific excise liability ($T_{a_k}^h$)	?	0	?

Table 18: Simulated aggregates in response to income shock and VAT rate reduction (Annual, in million euros)

		Income shock			**Reduced VAT**		
	Baseline	CS_Y	CQ	CS_E	CS_Y	CQ	CS_E
Gross earned income	184 450.6	177 653.3	177 653.3	177 653.3	177 653.3	177 653.3	177 653.3
Direct taxes	48 866.3	46 444.9	46 444.9	46 444.9	46 444.9	46 444.9	46 444.9
Social security contributions	24 166.7	22 942.9	22 942.9	22 942.9	22 942.9	22 942.9	22 942.9
Benefits	66 405.8	66 487.0	66 487.0	66 487.0	66 487.0	66 487.0	66 487.0
Disposable household income	177 823.3	174 752.4	174 752.4	174 752.4	174 752.4	174 752.4	174 752.4
Total expenditures	159 445.7	156 789.2	159 445.7	156 374.9	156 789.2	157 252.4	156 374.9
Saving	18 377.6	17 963.2	15 306.7	18 377.6	17 963.2	17 500.0	18 377.6
VAT revenues	18 264.0	17 962.4	18 264.0	17 922.2	16 080.7	16 084.9	16 045.5
Ad valorem Excises	464.2	456.1	464.2	455.3	456.1	449.9	455.3
Specific excises	4 065.0	3 996.5	4 065.0	3 980.6	4 066.1	4 065.0	4 050.0

8.3.1 The assumption of constant income shares

Under the constant income shares assumption, own price elasticities are -1 and cross price elasticities are zero. Consequently, expenditures will not change because of the VAT reduction *per se*, that is they will be identical to the expenditure levels after the income shock alone (total expenditure in the 5$^{\text{th}}$ column of Table 18 is equal to that of the 2$^{\text{nd}}$ column). Instead, since commodities have become cheaper, households will buy more of them. Compared to the CS_Y outcome for the pure income shock–scenario, VAT revenues decrease because of the rate changes. Indeed, increased consumption due to lower rates cannot compensate as expenditures remain the same as in the scenario with only an income shock. For the same reason, *ad valorem* excise revenues remain unchanged as compared to the case where there is only an income shock. Specific excise revenues increase because of the increased quantities. They slightly surpass the baseline level.

It is worthwhile to note that the effect on total tax liabilities is not *a priori* determined under CS_Y

though it turns out to be negative under the present simulation. In order to become positive, however, the income share of expenditures on goods with specific excises should be rather high, *quod non*.

8.3.2 The assumption of constant quantities

The CQ assumption implies zero price elasticities. The decrease in prices implies less resources are needed to maintain the same consumption. Hence, lower prices immediately translate into lower expenditure levels under this assumption (6^{th} column of Table 18). However, they still remain higher than in the CS_Y case. Because of the combination of a decrease in expenditures and lower rates, VAT revenues will drop with respect to the baseline. But they remain slightly above their level under the CS_Y–assumption. Lower expenditures also result in a drop in *ad valorem* excises. Contrary to what was the case for VAT revenues, the CQ–assumption yield an even larger drop in *ad valorem* excises than the CS_Y–assumption. Naturally, specific excise revenues still remain unchanged with respect to the baseline level.

8.3.3 The assumption of constant expenditure shares

Price elasticities under the CS_E assumption are equal to those under the CS_Y assumption. Hence, changes in expenditures and indirect tax revenues due to the VAT rate reduction *per se* must be identical to those of the CS_Y–case. However, given that saving in this case is kept constant at the baseline level, expenditures are lower than they were under the CS_Y assumption, as aggregate saving decreased under the CS_Y assumption. Therefore, all indirect tax revenues are lower than in the CS_Y–case too. As was the case for CS_Y, *ad valorem* excises exceed their level in the CQ case. They remain, however, constant as compared to the scenario with only an income shock. Indeed, expenditures did not change with respect to that scenario, neither did the *ad valorem* tariff.

8.4 Distributional and welfare effects

The main purpose of building a microsimulation tool remains of course to do distributional analysis. The purpose of the present section is not to give a full distributional analysis, but to give just an example of the type of output ITTv3 is able to produce. In Table 19, we calculate the changes with respect to the 2019 pre–shock baseline in euros for the combination of the income shock and lower VAT rates. We present in the table the same indicators as we calculated for the aggregate analysis.

Aggregate analysis already signalled that there might be some heterogeneity. Table 17 indicates that the effect of a combined change in disposable income and VAT rates on specific excises is ambiguous under CS_Y and CS_E. This is confirmed in Table 19. Persons with lower disposable income see their specific excise bill increase due to the change in the environment, while the reverse falls to the share of richer people. Similarly, the effect on saving under CS_Y depends on whether a person is initially a borrower or a saver, and it is ambiguous under CQ. As it turns out, in both cases, poorer

persons in terms of equivalised disposable income turn out to increase their saving level (*i.e.* decrease borrowing, as poorer people are more often net borrowers than savers) while saving is predominately diminished by richer persons.

Of course the ultimate question concerns the welfare implications of this change in economic environment. Regarding this question, the presented figures provide only superficial indications. Indeed, when poor people pay more specific excises after the reform, this is because they consume more of these goods. And if welfare can be equated with consumption, this is more a virtue than a vice (neglecting for a while also potential negative externalities of the consumption of excise goods).

If one concentrates on consumption as a welfare indicator, the CQ–hypothesis seems rather unattractive. It imposes that welfare never changes, whatever happens to the economic environment people live in, or the tax policies they are subjected to. This seems rather counter intuitive. Table 17 indicates that the effect on quantities is ambiguous under CS_Y and CS_E, and there is room for heterogeneity in the effects. It is unclear whether both assumptions would yield different simulation results in that respect. The final welfare evaluation might also depend on the type of quantity index one will use.

Table 19: Changes in household budget composition (Monthly, in Euros)

Equivalised disp. inc. deciles	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**All**
Baseline											
Household disposable income	1 270.7	1 720.2	2 108.6	2 584.2	2 954.4	3 321.6	3 791.6	4 162.5	4 543.5	6 345.3	3 280.3
Total expenditure	1 732.9	1 797.1	2 142.5	2 661.1	2 910.5	3 079.1	3 475.3	3 540.9	3 655.6	4 174.2	2 916.9
Saving	-462.2	-76.9	-33.9	-77.0	43.9	242.6	316.3	621.7	887.8	2 171.1	363.3
VAT	174.7	185.0	231.0	298.7	333.7	358.2	410.2	434.6	434.5	495.4	335.6
Ad valorem excises	8.5	7.0	8.1	8.8	9.5	11.1	9.2	8.3	7.8	5.1	8.3
Specific excises	49.6	46.9	55.9	70.7	73.5	82.1	91.0	85.5	89.6	97.2	74.2
Income shock & VAT reduction											
Constant income shares											
Household disposable income	-16.7	-17.9	-25.4	-40.5	-45.7	-64.7	-70.5	-84.3	-71.1	-137.9	-57.5
Total expenditure	-25.7	-18.0	-26.1	-45.1	-44.9	-63.8	-60.0	-66.7	-55.0	-87.4	-49.3
Saving	9.1	0.2	0.8	4.6	-0.8	-0.9	-10.5	-17.6	-16.1	-50.6	-8.2
VAT	-22.0	-22.3	-27.4	-36.8	-39.5	-44.2	-47.8	-50.6	-49.9	-60.2	-40.1
Ad valorem excises	-0.1	0.0	-0.2	-0.2	-0.1	-0.3	-0.1	-0.2	-0.1	-0.2	-0.1
Specific excises	0.3	0.4	0.4	-0.3	0.1	-0.3	-0.1	-0.1	0.0	-0.4	0.0
Constant quantities											
Household disposable income	-16.7	-17.9	-25.4	-40.5	-45.7	-64.7	-70.5	-84.3	-71.1	-137.9	-57.5
Total expenditure	-22.7	-23.6	-28.8	-36.6	-40.0	-43.0	-48.1	-50.2	-50.7	-58.2	-40.2
Saving	6.0	5.7	3.4	-3.9	-5.7	-21.8	-22.3	-34.1	-20.4	-79.7	-17.3
VAT	-22.4	-23.4	-28.5	-36.3	-39.7	-42.6	-47.9	-50.0	-50.5	-58.1	-39.9
Ad valorem excises	-0.3	-0.2	-0.2	-0.3	-0.3	-0.3	-0.3	-0.3	-0.2	-0.2	-0.3
Specific excises	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Constant expenditure shares											
Household disposable income	-16.7	-17.9	-25.4	-40.5	-45.7	-64.7	-70.5	-84.3	-71.1	-137.9	-57.5
Total expenditure	-16.7	-17.9	-25.4	-40.5	-45.7	-64.7	-70.5	-84.3	-71.1	-137.9	-57.5
Saving	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
VAT	-21.3	-22.3	-27.3	-36.0	-39.5	-44.1	-48.7	-52.0	-51.3	-65.4	-40.8
Ad valorem excises	-0.1	0.0	-0.2	-0.2	-0.1	-0.3	-0.1	-0.2	-0.1	-0.3	-0.2
Specific excises	0.4	0.4	0.3	-0.2	0.1	-0.6	-0.4	-0.7	-0.6	-1.9	-0.3

A particular quantity index which can be derived from a constant income share demand system, is disposable income divided by a Stone price index. This price index is simply a weighted geometric mean of consumer prices of each commodity. Weights are income shares of expenditures of these commodities. Saving is included in this price index as well and its price is set to one. The Stone Price index reads as:

$$P_w(\mathbf{q}) = \prod_k (q_k)^{w_k}. \tag{95}$$

Notice that the price indices used here are household specific as income shares of expenditures are evidently not the same for all households. An expenditure function associated constant income share demand system consists of deflating household disposable income with the price index $P_w(\mathbf{q})$[35]:

$$Q_y = \frac{y}{P_w(\mathbf{q})}. \tag{96}$$

Actually, this measure is a money metric utility (MMU) index associated with the constant income share demand system. Alternatively, one can call this welfare index 'real disposable income'.

Afterwards, we calculate the change in this MMU index following the income shock and accompanying VAT reduction. The relative size of this change will be used to define losers and winners of these shifts in the environment. We group the people in five categories, depending on the size of the percentage change in their MMU index. The first two groups are the losers of these shifts. The people who are negatively affected the most are those who lost more than 5% of their real disposable income and the second group of losers are the individuals who lost 1 up to 5% of their real income. The third group constitutes of individuals who are not really effected by these changes in the environment. Their losses (or gains) are less than 1%. The fourth and fifth categories are the winners. The fourth group experiences a 1 up to 5% increase in real income. The people in the fifth group are the biggest winners. Their real income increases with more than 5%.

Table 20 presents how certain sub–populations are affected by these changes in real disposable income. Our welfare analysis specifically focuses on age, gender, education, and economic activity of the individual. The first column of Table 20 shows the total number of people with the given characteristic. The second and third columns are the relative sizes of losers with the given characteristic. The fourth column represents the people for whom the welfare changes are negligible. Finally, fifth and sixth columns represent the relative size of winners with the given characteristic.

Our welfare analysis covers 8.5 million people who are 18 year or older. The people who lost more than 5% of real income constitute 10.6% of this population. People whose loss is 5% to 1% are quite similar in size (9.7%). The third group, where real income loses are negligible, are the second biggest group with 30.2%. Almost half of the population (48.2%) experiences a 1 to 5% increase in real disposable income, while for only 1.4% of the population this increase is above 5%.

[35] Developing a quantity index based on expenditures instead of income is rather straightforward. This measure deflates total expenditures, E, using expenditure shares as weights in the price index: $Q_E = \frac{E}{P_\omega(\mathbf{q})} = \frac{E}{\prod_k (q_k)^{\omega_k}}$.

Recall that total expenditures do not include saving.

Overall, there is a clear tendency with respect to the size of winners and individual characteristics. The share of winners are higher in population subgroups which are more prone to poverty such as elderly, women and people with lower educational attainment. Except for the lowest two age categories, the size of the groups of winners is increasing gradually with age. For the two oldest groups, people who are aged between 55 and 64 years and people who are 65 years or older, the size of winners are 55% and 69.1% respectively, which are quite high compared to the population average of 49.6%. The gender variation is less striking but we still observe a 3.8 percentage point difference between the share of winners among men (47.6%) and women (51.3%). People who have less than upper secondary education are more likely to experience welfare gains. The share of winners is highest among people with primary education (64.4%). Those who did not complete primary education and who only hold a lower secondary degree have the second and third largest share of winners, 58.3% and 52.4% respectively. We observe similar patterns with respect to economic activity. Only 38.5% of the people who are in the workforce experience welfare gains, while the same number is 53% for those who are unemployed and 67.5% for those retired. This stands to reason, as persons not working are never hit by the negative income shock.

Table 20: Changes in real disposable income by individual characteristics

	Total number of people	**Losers**		**Neutral**	**Winners**	
		$(-\infty, -0.05)$	$[-0.05, -0.01]$	$(-0.01, 0.01)$	$[0.01, 0.05]$	$(0.05, \infty)$
Age group						
18-25	916 279	14.2%	15.5%	29.7%	39.3%	1.3%
25-34	1 385 033	15.5%	14.2%	31.7%	37.7%	0.9%
35-44	1 534 534	14.8%	11.2%	31.8%	41.6%	0.7%
45-54	1 572 136	12.8%	11.2%	30.7%	44.4%	1.0%
55-64	1 304 321	7.0%	8.0%	29.9%	52.7%	2.3%
65+	1 731 591	2.0%	1.4%	27.5%	67.1%	2.0%
Gender						
Male	4 108 726	11.5%	9.8%	31.1%	46.3%	1.3%
Female	4 335 168	9.8%	9.5%	29.3%	49.9%	1.4%
Education						
Not completed Primary	460 018	9.0%	6.4%	26.2%	56.4%	1.9%
Primary	942 803	3.3%	6.2%	26.1%	62.3%	2.1%
Lower Secondary	1 613 939	9.0%	10.9%	27.8%	51.1%	1.3%
Upper Secondary	2 662 822	12.9%	11.4%	27.4%	46.9%	1.4%
Post Secondary	195 725	16.2%	13.9%	34.2%	34.7%	1.0%
Economic Activity						
Workforce	4 588 984	15.7%	14.2%	31.9%	37.3%	0.8%
Unemployed	548 376	9.8%	11.4%	25.8%	52.5%	0.5%
Retired	1 980 505	2.3%	2.7%	27.5%	65.4%	2.1%
Total	8 443 894	10.6%	9.7%	30.2%	48.2%	1.4%

9 Conclusion

This report presented the results of the project JRC/SVQ/2018/B.2/0021/OC. In the course of this project we developed a new Indirect Tax Tool for EUROMOD (ITTv3). ITTv3 modifies the existing indirect tax tool (ITTv2) in the following respects:

- the number of countries for which indirect tax simulation can be performed is increased to 18;
- the imputation of expenditure variables from the Household Budget Surveys (HBS) to the European Union Statistics on Income and Living Conditions (EU–SILC) is performed at the most detailed level of aggregation available (roughly 200 good categories). This enables the simulation of tax rate changes on narrowly defined good or service categories;
- the ITT which previously was an add–on, is now fully integrated into EUROMOD.

The following tasks have been executed throughout the project:

- the latest releases of the Eurostat versions of the national HBS's micro–data (2010) and the EU–SILC micro–data for the corresponding (or closest available) year were gathered for almost all member countries of the EU. A selection of 18 countries was made on the basis of relevance of the country and a first quality check of the available data. The datasets of these countries are prepared for imputation by means of standard data cleaning and harmonisation procedures;
- an imputation method has been developed in order to meet the challenges of imputing expenditure variables at highly disaggregated levels. A tool kit for evaluating the imputation results of this method has been developed and was applied to the imputation results for the 18 selected countries;
- ITTv3 has been integrated into the EUROMOD microsimulation model. Three alternative behavioural assumptions for simulating expenditure reactions to price and income fluctuations and associated changes in individual indirect tax burdens are available: constant income shares, constant quantities, and constant expenditure shares. The EUROMOD Statistics Presenter has been expanded to allow for analyses of output produced by ITTv3.

We summarise hereafter the main findings and results of our project.

- During the detailed data preparation work undertaken for this project, we identified a number of inconsistencies in the EUROSTAT versions of the 2010 Household Budget Survey micro–data on expenditures. At numerous occasions, the sum of expenditures at a more detailed level of aggregation is not equal to the expenditures reported at a higher level of aggregation. We therefore recommend to use these data cautiously. We also suggest to inform EUROSTAT about the problem and ask them to check their data manipulation and validation codes which create the current versions of the HBS rendered available for research purposes. The inconsistencies spotted in HBS had to be remedied by making certain assumptions in order to proceed with

imputations. We are unable to say at this stage how seriously this might affect the results of simulations with the new indirect tax tool of EUROMOD.

Another reason for being cautious while using HBS micro–data concerns the compilation guidelines of the samples composing these data. EUROSTAT renders available household budget surveys every five years, starting from the year 2005. Not all countries do, however, organise a budget survey in exactly those years. Therefore, EUROSTAT allows the transfer of data from surveys up to two years earlier than the reference year of the HBS. They also allow merging several datasets that belong to different years if no important methodological changes in data collection have taken place during that period. We know, for instance, that this is what Belgium has done for the 2010 EUROSTAT version. We believe that this practice should not pose a big problem as long as there are no major indirect tax reforms or producer price changes over the years the dataset spans. But a better description of how the final samples for each country are composed would be welcomed.

– The imputation method we developed is able to meet the objective to impute expenditures information into the SILC–datasets underlying EUROMOD at the most detailed level of aggregation. At such a detailed level of aggregation, expenditure data, as recorded in Household Budget Surveys, tend to be highly volatile, exhibiting many zero observations, and some positive outliers. This makes it practically impossible to estimate expenditures at very disaggregated levels. Hence, our previous approach in ITTv2, where we imputed expenditures with predicted values of a regression model, is not well suited for the imputation of expenditures at very high levels of disaggregation.

  We therefore returned to matching based imputation methods. Basically,for each household in the recipient dataset such methods determine how to find the closest resembling household in the source dataset, and then impute the missing variables in the recipient dataset from the observations for those closest households in the source data, on those variables. Such methods certainly enable it to impute expenditure values of an EU–SILC household with the values of the matched HBS household in full detail. However, many household characteristics are candidate to determine how closely resembling one household is to another, and it is unclear which weight one should give to each of these characteristics. The method we developed makes in a first step a regression analysis which allows to determine in how far such household characteristics are able to explain expenditures on broadly defined categories of goods, which are much less volatile. We use then fitted values of these regression models to determine the pairs of closest resembling households in source and recipient dataset. Once matched pairs were determined, imputation of missing values at the most detailed level of good aggregation is possible.

– The application of this new method to the 18 selected countries gave satisfying results. The detailed evaluation and macro–validation we perform in Section 6 revealed that the present imputation performs as well as the approach followed in ITTv2 (based on Engel curve

estimation).

- The EUROMOD implementation of a user friendly tool that allows running simulations, possibly varying over several hundreds of indirect tax instruments was not straightforward, of which the detailed explanation in Section 7 testifies. An illustrative simulation exercise in Section 8 shows therefore how the tool practically works.

  The integration of an indirect tax module into an income tax and benefit calculator has promising features, both from a theoretical and policy analysis point of view. The interaction of different tax instruments is often too complex to be fully characterised by theoretical models alone. Simulating joint reforms in the direct and indirect system – popular in current policy debates on shifting the tax burden from labour to other tax bases – may therefore reveal insights which could not be assessed purely analytically.

- We investigated the implications of relaxing the fixed producer prices assumption and therefore, the possibility to build in a pass–through parameter into the model. We found that the assumption of fixed quantities does not allow for flexible pass–through. Furthermore, for the two versions of constant shares, we showed that a flexible pass–through parameter had no implications for indirect tax revenues from goods on which no specific excises are levied. Unfortunately, for goods with specific excises then analysis reveal that iterative methods are required which the current EUROMOD architecture does not foresee. More importantly, a flexible pass–through may have important welfare implications, a point to which we come back below.

A number of issues are certainly worth to be further investigated.

- The option to impute expenditures at the most detailed level of aggregation does not resolve the issue of aligning statutory rates perfectly with the more detailed good categories. Aggregation of different tax rates applicable to one good category remains necessary and the implications of the choices which had to made need certainly to be investigated further.

  Moreover, such an imputation at the most detailed level of aggregation raises a number of questions. At such a detailed level of aggregation, the design of Household Budget Surveys where people record detailed expenditures during a rather limited period of the year, makes it difficult to discriminate between zero expenditure observations which are due to infrequent purchasing (*e.g.* due to bulk purchasing or love for variety) from 'true' zeros (being goods not consumed by that household). However, the detailed imputation approach considers all zero expenditures observed at certain moment in time as 'true' zeros. Depending on the purpose of the simulation, this might have serious implications. If, for example, the implications of tax reforms at this detail are to be investigated, one can obtain seriously biased results if zero purchases are confounded with zero consumption.

  The comparison of using several levels of aggregation in the imputation is therefore certainly worthwhile. The advantage of our imputation method is that it is flexible in this respect. Even

more so, simulating with higher levels of aggregation with the current imputation is in principle possible.

– We opted to impute income shares of expenditures as observed in the HBS into the EU–SILC of the corresponding year of observation (or the most closely available year). When one wants to perform the analyses for other policy years, the uprating system built into EUROMOD (see EUROMOD Modelling Conventions, 2015) can be used.

  This is however not the only option. One can use the present imputation method, and even the present imputation, to impute values into SILCs of another year than that of the budget survey from which the data stem. It is important to note that we opted to impute income shares of expenditures. If one assumes that the observed distribution of such income shares is relatively stable over time, such an approach is certainly warranted. A comparison of this approach with what comes out of the uprating praxis is certainly worth to be examined.

– The ultimate aim of microsimulation models is to do welfare analysis. We develop currently two consumption welfare measures to do so. They will be implemented when the ITTv3 goes public. But, as was pointed out earlier, the fruitfulness of an indirect tax model with a direct tax model relies on the ability to analyse the effects of the interaction of both instruments. The welfare implications of such interactions are not limited to how it affects persons' consumption opportunities. Its impact on the labour/leisure time trade–off needs certainly to be integrated into a welfare measure.

.

References

[1] Akoğuz Cansu E., B. Capéau, A. Decoster, L. De Sadeleer, & Toon Vanheukelom (2019). "A new indirect tax tool for EUROMOD JRC Project no. JRC/SVQ/2018/B.2/0021/OC Deliverable 1," document number Ares(2019) 6901733.

[2] De Agostini P., B. Capéau, A. Decoster, F. Figari, J. Kneeshaw, C. Leventi, K. Manios, A. Paulus, H. Sutherland, and T. Vanheukelom (2017). "EUROMOD extension to indirect taxation: final report," *EUROMOD technical note series*, EMTN/3.0.

[3] Decoster, A., D. Vandelanoote, T. Vanheukelom, and G. Verbist (2014), "Gross incomes in the Belgian SILC–dataset: An analysis by means of EUROMOD," Flemosi Discussion Paper DP33.

[4] Decoster A., J. Vanderkelen, W. Van Lancker, and T. Vanheukelom (2020), "Sociaaleconomische kenmerken van werknemers en zelfstandigen in sectoren getroffen door de lockdown," *Leuvense Economische Standpunten, 177*, Leuven.

[5] D'Orazio M., M. Di Zio, and M. Scanu (2006). *Statistical matching: Theory and practice*, John Wiley & Sons Ltd.: Chichester.

[6] EUROMOD Modelling Conventions (2015). download on 23/08/2020 from `https://www.euromod.ac.uk/sites/default/files/EUROMOD%20Modelling%20Conventions%20_26102015.pdf`

[7] Eurostat (2012). "Description of the data transmission for HBS (Reference Year) 2010," Luxembourg.

[8] Gorman W.M. (1981). "Some Engel curves," in: A. Deaton (ed.) *Essays in the theory and measurement of consumer behaviour in honour of Richard Stone*, Cambridge University Press: Cambridge, 7–29.

[9] International Monetary Fund (2020). "Policy Responses to Covid–19," Online.

[10] Lamarche P. (2017). "Measuring income, consumption and wealth jointly at the micro–level," Eurostat, *mimeo*.

[11] Leulescu A. and M. Agafitei (2013). "Statistical matching: a model based approach for data integration," *Eurostat – Methodologies and Working papers*, Publications Office of the European Union: Luxembourg.

[12] Raghunathan T.E., J.M. Lepkowski, J. Van Hoewyk, and P. Solenberger (2001). "A multivariate technique for multiply imputing missing values using a sequence of regression models," *Survey Methodology*, 27(1) (Statistics Canada Catlogue no. 12–001), 85–95.

[13] Serafino P. and R. Tonkin (2017). "Statistical matching of European Union statistics on income and living conditions (EU–SILC) and the household budget survey," *Eurostat – Statistical Working Papers*, Publications Office of the European Union: Luxembourg.

[14] van Buuren S. and K. Groothuis–Oudshoorn (2011). "MICE: Multivariate Imputation by Chained Equations in R," *Journal of Statistical Software*, 45(3), 1–67.

Appendices

Appendix I Country names and codes

Table 21: List of country abbreviations

Abb.	Country
BE	Belgium
CY	Cyprus
CZ	Czech Republic
DE	Germany
DK	Denmark
EL	Greece
ES	Spain
FI	Finland
FR	France
HU	Hungary
IE	Ireland
IT	Italy
LT	Lithuania
PL	Poland
PT	Portugal
RO	Romania
SI	Slovenia
SK	Slovakia

Appendix II Summary files of the imputations

For each country we provide an Excel summary file `summary XX.xlsx` (where `XX` stands for the country code defined in Appendix I) with information on the imputation results. The content of this file is structured as follows.

- Sheet 1: `descriptive statistics XX`

  The sheet contains sample statistics for the HBS and SILC variables used in the imputation.

  For HBS, we give the following statistics:

  - HBS – number of households with non-positive incomes: cells `2R:4X`;
  - HBS – number of households with negative expenditures on a good or good aggregate: cells `3X:4Y`;
  - HBS – is the reference person a farmer? : cells `29X:32X`;
  - HBS – region : cells `2AA:13AB`;
  - HBS – numerical covariates : cells `14R:26AF`;
  - HBS – expenditure levels: cells `34R:46AN`;
  - HBS – income shares of expenditures : cells `48R:60AN`.

  For SILC, we give the following statistics:

  - SILC – number of households with non-positive incomes : cells `2A:4G`;
  - SILC – is the reference person a farmer? : cells `29G:32G`;
  - SILC – region : cells `2J:13K`;
  - SILC – numerical covariates : cells `14A:26O`.

- Sheet 2: `regression results XX`
  - Pseudo–R^2 values of regressions per broad category: cells `1A:21B`.
  - Covariates used in the regression: cells `1D:19D`.
  - Covariates not used in the regression: cells `1F:19F`.
  - Detailed regression results for 20 broad categories: probits start at cells `25A:25H`; linear regressions start at cells `25J:25T`.

  Categories retained as input for the distance function (pseudo–R^2>.1) are highlighted in green.

– Sheet 3: `ventile tables and graphs XX`

  Contains summary statistics of income shares of expenditures on broad categories including:

  – weighted mean, minimum and maximum values of income shares of expenditures, overall and per income ventile,

  – overall and per income ventile 5th, 25th, 50th, 75th and 90th population percentiles of income shares,

  – weighted mean household disposable income overall and per income ventile, and

  – an alternative mean income share, calculated as population total expenditure (overall or per ventile) divided by total income (overall or per ventile), denoted by `mean2` in the sheet.

Table 22: Position of the statistics for SILC (imputed expenditure values)

Food and non–alcoholic beverages:	cells 2A: 27K,
Housing (rental):	cells 30A: 55K,
Housing (goods and services):	cells 58A: 83K,
Utilities:	cells 86A:111K,
Communications:	cells 114A:139K,
Culture and recreation:	cells 142A:167K,
Personal care:	cells 170A:195K,
Insurance:	cells 198A:223K,
Alcoholic beverages:	cells 226A:251K,
Tobacco:	cells 254A:279K,
Private transportation:	cells 282A:307K,
Public transportation:	cells 310A:335K,
Travelling and holiday:	cells 338A:363K,
Education:	cells 366A:391K,
Vehicles:	cells 394A:419K,
Housing (durables):	cells 422A:447K,
Clothing and personal items:	cells 450A:475K,
Health and care:	cells 478A:503K,
Restaurants:	cells 506A:531K,
Other:	cells 534A:559K,
Total expenditure:	cells 562A:587K,
Saving:	cells 590A:615K.

Table 23: Position of the statistics for HBS (observed expenditure values)

Food and non–alcoholic beverages:	cells 2M: 27W,
Housing (rental):	cells 30M: 55W,
Housing (goods and services):	cells 58M: 83W,
Utilities:	cells 86M:111W,
Communications:	cells 114M:139W,
Culture and recreation:	cells 142M:167W,
Personal care:	cells 170M:195W,
Insurance:	cells 198M:223W,
Alcoholic beverages:	cells 226M:251W,
Tobacco:	cells 254M:279W,
Private transportation:	cells 282M:307W,
Public transportation:	cells 310M:335W,
Travelling and holiday:	cells 338M:363W,
Education:	cells 366M:391W,
Vehicles:	cells 394M:419W,
Housing (durables):	cells 422M:447W,
Clothing and personal items:	cells 450M:475W,
Health and care:	cells 478M:503W,
Restaurants:	cells 506M:531W,
Other:	cells 534M:559W,
Total expenditure:	cells 562M:587W,
Saving:	cells 590M:615W.

Plots of income shares of expenditures against disposable income are included to the right of the table for HBS (columns Y:AM).

- Sheet 4: `correlation differences XX`
  - Mean difference in weighted correlation in HBS and SILC:

within covariates:	cell `2B`,
between covariates and expenditure categories:	cell `3B`,
within expenditure categories:	cell `4B`.

  - Difference in weighted correlation matrix of HBS and SILC.

    The start position of the matrix is cell `8B`, and the matrix is structured as follows:

difference in correlation	
within covariates	between covariates and expenditure categories
between expenditure categories and covariates	within expenditure categories

    For the covariates part we excluded the square and cube of log income and the categorical variables.

    For the expenditures part, we included entries for total expenditures and saving, but these are not retained in the calculations of the means.

    Cells of the matrix are shaded in darker red the larger the value of the cell (absolute value of difference in estimated correlation for that cell between SILC and HBS).

Appendix III Definition of the 20 broad expenditure categories

Table 24 and Table 25 display the composition of the 20 broad aggregates on which we performed the regressions for each country. We indicate the composition by means of the HBS–nomenclature for the commodities and aggregates, but notice that we sometimes overwrote the variables as recorded in the HBS according to the rules described in Section 4.4.1.4.

A number of descriptive statistics on the levels and income shares of the expenditures on those broad categories is provided in the sheet `descriptive statistics XX` of the summary file of the imputation for each country (`summary XX`), where `XX` stands for the country code (see Appendix I). *Total expenditure* is defined as the sum of expenditures on these broad categories, and *saving* denotes the difference between household disposable income and total expenditure.

Table 24: Composition of consumption categories

Category	Level of aggregation							
	first		second		third		fourth
1. Food and non–alcoh. beverages	EUR_HE01						
2. Housing: actual rentals		+	EUR_HE041				
3. Utilities		+	EUR_HE045 (Electricity, gas, and other fuels)	+	EUR_HE0441 (Water supply)		
				+	EUR_HE0442 (Refuse collection)		
				+	EUR_HE0443 (Sewerage collection)		
4. Communication	EUR_HE08						
5. Personal care		+	EUR_HE121				
6. Insurance		+	EUR_HE125				
7. Alcoh. beverages		+	EUR_HE021				
8. Tobacco		+	EUR_HE022				
9. Private transport		+	EUR_HE072 (Operation of personal transport equipment)				
10. Education	EUR_HE10						
11. Clothing and personal items	EUR_HE03 (Clothing and footwear)	+	EUR_HE123 (Personal items)				
12. Health & care	EUR_HE06 (Health products and services)	+	EUR_HE124 (Social protection services)				
13. Restaurants		+	EUR_HE111 (Catering services)				

Table 25: Composition of consumption categories (continued)

Category	Level of aggregation			
	first	second	third	fourth
14. Housing: goods and services	EUR_HE05 (Furnishings, household equipment, and routine maintenance of the house)	+ EUR_HE043 (Maintenance and repair of the dwelling)	+ EUR_HE0444 (Other services relating to the dwelling) – EUR_HE0531 (Large household appliances) – EUR_HE0532 (Small electrical household appliances) – EUR_HE0551 (Big tools for garden)	– EUR_HE05111 (Furniture and furnishings) – EUR_HE05112 (Carpets and other floor coverings)
15. Housing: durables			+ EUR_HE0531 (Large household appliances) + EUR_HE0532 (Small electrical household appliances) + EUR_HE0551 (Big tools for garden)	+ EUR_HE05111 (Furniture and furnishings) + EUR_HE05112 (Carpets and other floor coverings)
16. Culture and leisure	EUR_HE09 (Recreation and culture)	– EUR_HE096 (Package holidays)		
17. Public transport		+ EUR_HE073 (Transport services)		
18. Vehicles		+ EUR_HE071 (Purchase of vehicles)		
19. Travelling and holiday		+ EUR_HE096 (Package holidays) + EUR_HE112 (Accommodation services)		
20. Other		+ EUR_HE127 (Other services) + EUR_HE126 (Financial services)		

Appendix IV Covariates used in different countries

As explained in the main text, we use a number of socio–demographic variables as covariates in the regressions. Recall that the covariates for all regressions (20 linear OLS regressions for positive expenditures and 20 probits for estimating the probability of positive expenditures) are the same per country. Table 26 contains a full list of the variables that were considered to be included as explanatory variable. A list of descriptive sample statistics of these variables for both HBS and SILC can be found in the sheet `XX descriptive statistics` of the summary file for each country (`summary XX`). For reasons explained in the main text, age dummies do not occur in that list.

A variable is excluded from the regressions if the information on the variable is absent or not trustworthy in either the SILC or the HBS dataset. In the next table we indicate by a zero which variables are not included in the regressions for each of the countries.

Table 26: Full list of explanatory variables by country

Variable	BE	CY	CZ	DE	DK	EL	ES	FI	FR	HU	IE	IT	LT	PL	PT	RO	SI	SK
HH disposable income (3^{rd} degree polynomial)																		
n adult male HH members																		
n HH members age <=14																		
n HH members 15 – 29																		
n HH members 30 – 44																		
n HH members 45 – 59																		
n HH members age >= 60																		
n employed HH members																		
n unemployed HH members																		
n pensioned HH members																		
n disabled HH members				0			0	0							0			
n student HH members age >14																		
n with higher education					0			0									0	
n non–EU citizens						0		0					0			0	0	
reference person farmer	0	0	0	0	0	0	0	0	0	0	0	0			0	0	0	0
region dummies		0	0	0	0	0		0	0		0	0	0		0		0	0