SInTax: microsimulation of VAT and Excises in the context of EUROMOD*

André Decoster†, Kevin Spiritus

September 30, 2016

Abstract

This user manual describes the operation of the SInTax (SImulation of Indirect Taxes) microsimulation module, developed to run alongside EUROMOD. We explain how Engel curves are estimated based on a household budget survey, by first estimating total durable and total nondurable expenditures, and next estimating in a number of steps budget shares on a number of categories of nondurable expenditures. Next we describe how expenditures on different expenditures categories are imputed into a different dataset, e.g. into a dataset used by a direct tax-benefit simulator such as EUROMOD, enabling the combined simulation of direct and indirect taxes. We explain how aggregate implicit tax rates are calculated for categories of expenditures, enabling the simulation of household expenditures after an indirect tax reform. We show how welfare effects of such a reform are calculated. In the second part of this manual we explain how the model is used in practice, preparing all the estimations and configuration files necessary to run a simulation. The final section discusses the integration of the model into EUROMOD.

Keywords: Budget survey, expenditure estimation, Engel curves, EUROMOD.
JEL Classification: D12, D31, H24, H31

*This paper was written as part of the SBO-project "FLEMOSI: A tool for ex ante evaluation of socio-economic policies in Flanders", funded by IWT Flanders. See www.flemosi.be. We are indebted to the many people who have contributed to the development of EUROMOD and to the European Commission for providing financial support for it. This paper greatly benefited from valuable discussions with Peter Haan, and the audience of the EUROMODupdate2 project meeting 2012 in Bucharest. We are grateful to Vincent Mouton from Statistics Belgium for his valuable support with the Household Budget Survey. The results and their interpretation are the authors' responsibility.

†Both authors are at the Center for Economic Studies - Public Economics at KU Leuven. Contact: andre.decoster@kuleuven.be

Contents

1 Introduction

SInTax is a microsimulation module for indirect taxes, mainly based on Engel curves estimated on a Household Budget Survey. It can be used to predict expenditures and indirect tax liabilities for households with different incomes and demographic characteristics under a hypothetical commodity tax regime. The taxes covered by the model are VAT and specific and ad valorem excises. The model is designed to work adjoining to other microsimulation models, enabling the combined simulation of for example consumption and income taxes. Although the model can work independently and with a variety of other microsimulation models, SInTax was specifically designed to work together with the tax-benefit simulator EUROMOD[1]. As EUROMOD uses the EU-SILC dataset for its microsimulations, which contains no full information about household expenditures, the standard version of the model does not allow simulation of indirect taxes. This manual will explain how household expenditures can be imputed into the SILC data, and how EUROMOD can be combined with SInTax to enable the combined simulation of direct and indirect tax reforms.

SInTax is a partial equilibrium model, in that it assumes that producer prices remain fixed when a reform is simulated. The advantage of SInTax over a general equilibrium model is that it enables the user to assess the impact of a reform on the level of individual households, allowing distributional analysis.

This manual consists of two parts. The first part will give an overview of the model, discussing how it works and referring to the microeconomic theory behind it. The second part will explain how the model can be used.

Because the explanations in this manual are meant to be generally applicable to anyone using the model, it may come across as too abstract. In order to get a clearer view, the reader is advised to study the documentation of the Belgian[2] or the German[3] implementation. Do-files for the Belgian implementation are provided with the model as an illustration.

[1]EUROMOD is a tax-benefit microsimulation model for the European Union (EU) that enables researchers and policy analysts to calculate, in a comparable manner, the effects of taxes and benefits on household incomes and work incentives for the population of each country and for the EU as a whole. It is maintained, developed and managed by the Institute for Social and Economic Research (ISER) at the University of Essex in collaboration with national teams from the EU member states. References to EUROMOD in this document are based on version F6.36. See https://www.iser.essex.ac.uk/euromod for more information.

[2]See Decoster et al. [2014]

[3]See Decoster et al. [2013]

Part I
Engel curves and implicit indirect tax rates on aggregates

The SInTax model consists of two main components. In the first component a consumer demand model is estimated on a Household Budget Survey. The estimated parameters are then used in the second component to predict new consumption budget shares and indirect tax liabilities in a counterfactual situation. This counterfactual situation may involve any combined change of indirect tax parameters, demographic characteristics or disposable incomes – enabling the combined simulation of direct and indirect tax reforms.

A Household Budget Survey contains expenditure information for a large sample of households, typically on a monthly basis and for a large number of commodities. It also contains a number of demographic characteristics for each of the households. Section 2 will explain how this information is used to estimate the consumer demand system.

The typical Household Budget Survey covers a very large number of commodities. For many of those commodities there will be a large share of the population who do not consume positive amounts. This may be due to preferences, but also e.g. to seasonal effects[4]. In order to avoid the technical and conceptual problems that would arise if the demand system were estimated for each of the individual goods, expenditures are aggregated into a number of categories.

This aggregation in turn poses a new problem. It is possible that within the same category differentiated commodity taxes are levied on different goods, or one may wish to implement a commodity tax reform that only applies to a portion of the commodities in one group. The different tax rates on individual commodities will need to be aggregated into some concept of *implicit aggregate tax rates* on the categories. Tax reforms on individual goods will then impact the demand for the different categories through their influence on these implicit aggregate tax rates. This process of aggregation and the calculation of implicit tax rates will be the topic of section 3.

The user may wish to predict expenditures for a different dataset than the Household Budget Survey. If for example the SInTax model is used in combination with EUROMOD, household information will be based on

[4]A household filling out the survey in January may show a different consumption pattern on fruit and vegetables than if it fills out the survey in July.

the SILC dataset[5], which does not contain the necessary information on household expenditures. Before being able to predict expenditures shares in the counterfactual situation, one needs to impute baseline expenditures. This is possible if the input dataset contains the same covariates as were used for the estimations in the Household Budget Survey. This will be explained in section 4.

Section 5 will explain how information about the baseline is used to calculate implicit aggregate tax rates and consumer prices for the counterfactual situation. This allows prediction of the new budget shares, either in an arithmetical way – assuming constant budget shares or constant quantities, or taking into account behavioural responses using the estimated demand system parameters.

2 Estimating Engel curves

OLS regressions are used to estimate Engel curve parameters on the Household Budget Survey. The model assumes that there is one category aggregating all expenditures on durable goods – e.g. the purchase of transport vehicles – and a number of categories of non-durable expenditures – such as food, tobacco and clothing. The user could e.g. opt to use the COICOP specification[6].

The estimations occur in a number of steps. First both total durable and non-durable expenditures are estimated, with disposable income and a number of demographics as covariates. This is explained in subsection 2.1. Since the difference between disposable income and the sum of durable and total non durable expenditures is defined as savings, this first step also implies the estimation of a static savings function.

Expenditures on the different categories of non-durable expenditures are then estimated in two steps. For commodity groups for which many households have zero expenditures – such as tobacco and public transport – a two-step procedure is applied, as explained in subsection 2.2. The estimations for the remaining groups is explained subsection 2.3. Although the model leaves some liberty in choosing covariates used in the estimations, a natural choice would be to use the QUAIDS model. This is explained in

[5]European Union Statistics on Income and Living Conditions (EU-SILC). See http://epp.eurostat.ec.europa.eu/portal/page/portal/microdata/eu_silc for more information.

[6]Classification of Individual Consumption According to Purpose, a classification published by the United Nations Statistics Division. See http://unstats.un.org/unsd/cr/registry/regcst.asp?Cl=5 for more information.

subsection 2.4.

2.1 Estimations for total durable and non-durable expenditures

There will be many households in the data without durable expenditures in the surveyed period. For this reason a two-step procedure is applied in the estimation:

1. A probit estimation is used to estimate the probability that a household, given its characteristics, has non-zero durable expenditures (denote durable expenditures of the household h by e_D^h):

$$Pr\left(e_D^h > 0\right) = \beta_{PrD} X_{PrD}^{h\prime} + \varepsilon_{PrD}^h, \tag{1}$$

   where X_{PrD} are a number of covariates and β_{PrD} are the corresponding estimation parameters. The covariates typically include some function of disposable income and a number of demographic parameters.

2. If total durable expenditures are greater than zero, a function $f(.)$ of their amount is estimated on a number of covariates X_D, which are not necessarily the same as those used in the first step:

$$f\left(e_D^h\right) = \beta_D X_D^{h\prime} + \varepsilon_D^h. \tag{2}$$

For total non-durable expenditures the estimation only takes a single step, as no household will have zero expenditures in the surveyed period. Denoting by e_{ND}^h total non-durable expenditures the following model is estimated:

$$g\left(e_{ND}^h\right) = \beta_{ND} X_{ND}^{h\prime} + \varepsilon_{ND}^h. \tag{3}$$

The functions $f(.)$ and $g(.)$ will typically be logarithmic. The choice is left to the user though.

2.2 Estimations for groups with many zeroes

Among the categories of non-durable expenditures, there will be some for which many households are observed to have zero expenditures. Typical examples are rents, education, public transport and tobacco. For these categories, first probit estimations are used to determine the probability of positive expenditures:

$$Pr\left(e_G^h > 0\right) = \beta_{PrG} X_{PrG}^{h\prime} + \varepsilon_{PrG}^h, \tag{4}$$

where we denote the category under examination with a subscript G. The covariates X_{PrG} will typically include some function of total non-durable expenditures e^h_{ND} and a number of demographic characteristics.

For households with non-zero expenditures on a specific category G, the share $\omega^h_G = \frac{e^h_G}{e^h_{ND}}$ out of total non-durable expenditures is estimated:

$$\omega^h_G = \beta_G X^{h\prime}_G + \varepsilon^h_G. \tag{5}$$

One possible form for the covariates is the Engel curve part of a QUAIDS demand system (i.e. without relative price variation), as explained in more detail in subsection 2.4.

2.3 Estimations for the remaining groups

For the remaining categories a single-step estimation is performed. Let e^h_R denote the remaining amount of non-durable expenditures after expenditures on the groups treated in the previous subsection have been subtracted. The budget share out of these remaining non-durable expenditures, $\omega^h_G \equiv \frac{e^h_G}{e^h_R}$, is then estimated:

$$\omega^h_G = \beta_G X^{h\prime}_G + \varepsilon^h_G. \tag{6}$$

Here the covariates X_G typically include some function of remaining non-durable expenditures e^h_R and a number of demographic characteristics. Again a natural choice for the covariates is suggested by the QUAIDS model, explained in the next subsection.

2.4 Engel curves from the QUAIDS model

The Quadratic Almost Ideal Demand System (QUAIDS) model was introduced by Banks et al. [1997], as an extension to the Almost Ideal Demand System (AIDS) model introduced by Deaton and Muellbauer [1980]. The purpose of these models was to have restrictions to consumer demand from micro-economic theory readily built into a consumer demand model. The restrictions referred to are those of additivity, homogeneity and symmetry. An elaborate explanation is found in the book of Deaton and Muellbauer [1980]. The QUAIDS model was introduced in order to add more flexibility to the model, accommodating the non-linearity of Engel curves observed in consumer surveys, while retaining the advantages of the AIDS.

The SInTax model only uses the Engel curve part of the QUAIDS specification. It does not take into account relative price effects, but only real income effects. The Engel curve in QUAIDS is specified as:

$$\omega_G = \alpha_1 \log E + \alpha_2 (\log E)^2 + D'_G \delta_G + \varepsilon^h_G, \tag{7}$$

where E denotes the *real* value of the *total* expenditures used in the definition of ω_G, being the real values of e^h_{ND} in subsection 2.2 and e^h_R in subsection 2.3. The remaining covariates D_G will typically include a number of demographic characteristics and some interaction terms between total expenditures and these demographics. The fact that *real* expenditures are used has no importance for estimations in a cross-sectional budget survey. One should take into account changed price indices though when simulating an indirect tax reform.

3 Aggregation of disaggregated commodities into categories

In general, the commodity categories will contain goods which are taxed at different rates. Different foodstuff for example can be taxed as necessity or luxury items, or different excise rates may apply to different alcoholic products. It also should be possible to simulate reforms which apply differently to different products in the same category. Suppose for example that all food items are aggregated into one category, and the user wishes to introduce a tax on unhealthy foods. Different changes will then apply to different commodities in the category. In order to make this possible, it is necessary to introduce a concept of *implicit aggregate tax rates* on the categories.

We start this section by introducing in subsection 3.1 the concept of *implicit tax* rates on individual commodities. On each individual good, three types of indirect taxes apply: Value Added Taxes (VAT), ad valorem excises and specific excises. For simulations of consumer demand, only changes in the final consumer price are of importance. For this reason, the difference between the producer price and the consumer price of a good is summarized by the implicit tax rate on the good, summarizing the wedge introduced by the three indirect taxes. Subsection 3.2 will explain how these implicit rates can be aggregated to the level of the categories. Subsection 3.3 will show how this is applied in the SInTax model.

3.1 Implicit tax rates and tax liabilities on individual goods

For each household h the Household Budget Survey contains expenditures e^h_k on a number of commodities k. Three types of indirect taxes can be levied, which can in principle be different for each of the commodities: a Value Added Tax at rate t_k, levied on the producer price p_k augmented with the

any excises, an ad valorem excise at rate v_k, levied on the consumer price q_k, and a specific excise at rate a_k per unit. The consumer price q_k is related to the producer price p_k as follows:

$$q_k = (1 + t_k)(p_k + a_k + v_k.q_k). \tag{8}$$

Note how the consumer price appears on both sides of this equation. This is because the ad valorem excise is defined as a fraction of the consumer price.

The implicit tax rate τ_k on good k is defined by the formula:

$$q_k \equiv p_k(1 + \tau_k). \tag{9}$$

To be able to calculate the implicit tax rate, we need to express the specific tax rate a_k as a fraction α_k of the producer price:

$$\alpha_k \equiv \frac{a_k}{p_k} \tag{10}$$

We come back to the problem of the unobserved, and hence unknown producer price, in section 3.3. Inserting this into equation (8) and rearranging:

$$q_k = p_k \frac{(1 + \alpha_k)(1 + t_k)}{1 - (1 + t_k) v_k}, \tag{11}$$

such that, using definition (9):

$$\tau_k = \frac{(1 + \alpha_k)(1 + t_k)}{1 - (1 + t_k) v_k} - 1. \tag{12}$$

Knowledge of this implicit tax rate allows determining the household's tax liability T_k^h for the commodity. Denoting by x_k^h the quantity consumed by the household, and using $T_k^h \equiv \tau_k p_k x_k^h$, we find:

$$e_k^h = q_k x_k^h = (1 + \tau_k) p_k x_k^h \equiv p_k x_k^h + T_k^h. \tag{13}$$

Using the fact that $p_k x_k^h = \frac{e_k^h}{1+\tau^k}$ and rearranging:

$$T_k^h = \frac{\tau_k}{1 + \tau_k} e_k^h, \tag{14}$$

which shows how, once we have calculated the implicit tax rates τ_k, we can apply them to observed expenditures e_k to derive the tax liabilities.

One may wish to split the total indirect tax liability in (14) into different components. The VAT component $T_k^{h,t}$ is calculated as follows:

$$T_k^{h,t} = \frac{t_k}{1 + t_k} e_k^h = t_k (p_k + a_k + v_k.q_k) x_k^h. \tag{15}$$

The ad valorem excise liability $T_k^{h,v}$ is defined as follows:

$$T_k^{h,v} \equiv v_k e_k^h, \tag{16}$$

and the specific excise liability $T_k^{h,a} \equiv a_k x_k^h$ can be shown to be the remaining amount of the indirect tax liability:

$$\begin{aligned} T_k^h - T_k^{h,t} - T_k^{h,v} &= (q_k - p_k)\, x_k^h - t_k\, (p_k + a_k + v_k.q_k)\, x_k^h - v_k q_k x_k^h \\ &= (1 + t_k)\, (p_k + a_k + v_k.q_k)\, x_k^h - t_k\, (p_k + a_k + v_k.q_k) x_k^h \\ &\quad - p_k x_k^h - v_k q_k x_k^h \\ &= a_k x_k^h \\ &\equiv T_k^{h,a}. \end{aligned} \tag{17}$$

(18)

Note that the implicit tax rates are the same for all households in the population. Let dropping the index $.^h$ denote a sum over the entire population. Then equation (14) can be rewritten:

$$T_k \equiv \sum_h T_k^h = \frac{\tau_k}{1 + \tau_k} \sum_h e_k^h \equiv \frac{\tau_k}{1 + \tau_k} e_k. \tag{19}$$

The components T_k^t, T_k^v and T_k^a can be similarly defined.

3.2 Aggregating tax liabilities into categories

Suppose now that the commodities are completely partitioned into disjoint categories G. In what follows, let an index $._G$ denote a sum over all commodities k belonging to category G, for example $e_G \equiv \sum_{k \in G} e_k$ and $T_G \equiv \sum_{k \in G} T_k$.

Let the implicit aggregate tax rate τ_G on category G be defined by the identity:

$$T_G \equiv \frac{\tau_G}{1 + \tau_G} e_G. \tag{20}$$

The rate τ_G can then be calculated from previously derived quantities T_G and e_G, as follows:

$$\tau_G = \frac{T_G}{e_G - T_G}. \tag{21}$$

Similar quantities could be defined for the components VAT, ad valorem excise and specific excise.

3.3 Application in the SInTax model

When trying to apply the preceding results in the simulation model, we face two problems. The first arises from the fact that, in order to calculate the

specific excise rate α_k using equation (10), we need to know the producer price both in the baseline and in the counterfactual situation. The problem is that producer prices are not observed in the HBS. A second problem is that calculating the tax liabilities for individual goods using formula (19) assumes knowledge of the expenditures on the commodities. The problem though is that this is exactly what we are trying to predict.

To solve the first problem, that producer prices would be needed in order to determine α_k, a different procedure is used for the baseline and the counterfactual situation. Denote the baseline situation by an index $._{BAS}$ and use equation (9) to rewrite equation (10):

$$\alpha_{k,BAS} = a_{k,BAS}.\frac{1+\tau_{k,BAS}}{q_{k,BAS}}. \tag{22}$$

Substituting this into equation (12) and solving for the implicit tax rate:

$$\tau_{k,BAS} = \frac{1}{\frac{1}{1+t_{k,BAS}} - v_{k,BAS} - \frac{a_{k,BAS}}{q_{k,BAS}}} - 1, \tag{23}$$

which now expresses the implicit tax rates in terms of the three tax parameters (t_k,v_k and a_k) and the - at least in principle - more easily observable consumer price $q_{k,BAS}$.

In the counterfactual situation of changed indirect tax rates, however, the new consumer price $q_{k,SIM}$ is not yet known (where we denote the counterfactual situation by an index $._{SIM}$). To get around this, we assume that producer prices are unchanged, compared to the baseline. :

$$p_{k,SIM} \equiv p_{k,BAS} = \frac{q_{k,BAS}}{1+\tau_{k,BAS}}. \tag{24}$$

This allows to redefine the specific excise rate (10):

$$\alpha_{k,SIM} = \frac{a_{k,SIM}}{p_{k,SIM}} = a_{k,SIM}.\frac{1+\tau_{k,BAS}}{q_{k,BAS}}, \tag{25}$$

where both $\tau_{k,BAS}$, obtained using equation (23), and consumer price $q_{k,BAS}$ are available at the time of the simulation. We repeat that this implies that at the time of the simulation, the consumer prices $q_{k,BAS}$ need to be known for commodities where the specific excise a_k is not zero. With the known $\alpha_{k,SIM}$ we can now rely on (12) to derive the new implicit tax rate in the counterfactual situation:

$$\tau_{k,SIM} = \frac{(1+\alpha_{k,SIM})(1+t_{k,SIM})}{1-(1+t_{k,SIM})v_{k,SIM}} - 1. \tag{26}$$

In order to solve the second problem, that in order to calculate the implicit aggregate tax rates in the reform situation we need to know expenditures on the individual goods, we make an additional assumption. In the aggregation process we assume that the population average of the budget share spent on each good k within its category G, $\frac{e_k}{e_G}$, remains fixed with respect to the HBS situation. Thus:

$$e_{k,SIM} = \frac{e_{k,SIM}}{e_{G,SIM}}.e_{G,SIM} = e_{k,HBS}\frac{e_{G,SIM}}{e_{G,HBS}}, \tag{27}$$

where $e_{G,SIM}$ will be produced by the simulation of aggregate expenditures in the simulation part of the exercise.

Using equation (21) and the implicit tax rates on the individual commodities as calculated in (26), we can now calculate the aggregate implicit tax rate in the reform situation:

$$\begin{aligned}
\tau_{G,SIM} &= \frac{T_{G,SIM}}{e_{G,SIM} - T_{G,SIM}} \\
&= \frac{\sum_{k\in G} T_{k,SIM}}{\sum_{k\in G}\left(e_{k,SIM} - T_{k,SIM}\right)} \\
&= \frac{\sum_{k\in G} \frac{\tau_{k,SIM}}{1+\tau_{k,SIM}} e_{k,SIM}}{\sum_{k\in G} e_{k,SIM}\left(1 - \frac{\tau_{k,SIM}}{1+\tau_{k,SIM}}\right)} \\
&= \frac{\sum_{k\in G} \frac{\tau_{k,SIM}}{1+\tau_{k,SIM}} e_{k,HBS}}{\sum_{k\in G} e_{k,HBS}\left(1 - \frac{\tau_{k,SIM}}{1+\tau_{k,SIM}}\right)}.
\end{aligned} \tag{28}$$

Note how all the information we need with this assumption of fixed population budget shares, are the implicit rates on the individual goods and the population expenditures on the goods in the budget survey. Also note that this assumption is only made in determining the value of τ_G. Budget shares spent on the category are allowed to change during a reform, it is only the budget shares within the categories which are assumed not to change. The aggregate implicit tax rates found using equation (28) allow simulating tax liabilities and determining price indices for simulating expenditures based on real disposable incomes, as explained in section 5.

4 Imputation

Section 5 will show that, whichever method is chosen to simulate expenditures and tax liabilities, reference is made to the baseline expenditures of the households. If the user chooses not to use the household budget survey as

input file for his simulations, he will need to impute baseline expenditures. The same covariates should be available in the target database as those that were used during the estimation.

The imputation of baseline expenditures will follow the same steps as the estimation. First total durable and non-durable expenditures are imputed, then expenditures for all non-durable expenditure categories are imputed. The imputations are done using the parameters that were estimated in section 2. For durable expenditures and for non-durable categories with many zeroes, the two-step procedure is followed. The procedure to run an imputation is straightforward, and will be explained in Part II of this manual.

There are two complications though. A first point of attention is that in the estimation for total non-durable and durable expenditures, the dependent variable often is chosen to be logarithmic:

$$\log\left(e_D^h\right) = \beta_D X_D^{h\prime} + \varepsilon_D^h, \tag{29}$$

$$\log\left(e_{ND}^h\right) = \beta_{ND} X_{ND}^{h\prime} + \varepsilon_{ND}^h. \tag{30}$$

In order to obtain expected expenditure levels, one cannot simply take the exponent of the predicted values:

$$E\left[e^h\right] = E\left[e^{X^{h\prime}\beta+\varepsilon^h}\right] \neq e^{E\left[X^{h\prime}\beta+\varepsilon^h\right]} = e^{E\left[log\left(e^h\right)\right]}, \tag{31}$$

It can be shown[7] that the predicted values need to be adjusted by the HBS population averages of $\exp\left(\varepsilon_\cdot^h\right)$:

$$E\left[e^h\right] \approx e^{E\left[log\left(e^h\right)\right]}.E_{HBS}\left[\exp\left(\varepsilon^h\right)\right]. \tag{32}$$

In the second part of this manual it will be explained how these expected values are available at the imputation stage.

A second complication is that predictions of budget shares on non-durable commodity categories can be rather sensitive to errors made in preceding steps of the imputation. If some covariates are distributed differently in the target dataset than in the HBS, the distributions of the imputed total non-durable expenditures will differ from those in the HBS, which may lead to unexpected patterns in the distributions of imputed expenditures on the non-durable categories. In order to improve the quality of the imputations, some corrections may need to be performed.

One way to introduce such corrections is, when using disposable income as a covariate in the prediction of total non-durable and durable expenditures, to shift and rescale it such that its mean and variance correspond to those

[7] See Wooldridge [2003], pp. 207-210 and 276-280

of the corresponding variable in the HBS. Similarly, the distribution of total non-durable expenditures used in the imputations based on subsection 2.2 and the distribution of total remaining expenditures used in the imputations based on subsection 2.3 can be adjusted when predicting the budget shares spent on the non-durable expenditure categories. Again, total non-durable expenditures are left untouched, it is only when they are used as covariates to determine budget shares that a corrected variable is used.

5 Simulating expenditures and tax liabilities

At the time of a simulation in SInTax, the following information is available:

- Baseline expenditures $e^h_{G,BAS}$ for all commodity groups G, for all households h
- Implicit tax rates $\tau_{G,BAS}$ for groups G in the baseline and $\tau_{G,SIM}$ in the counterfactual situation, calculated internally as explained in section 3.

When expenditures are simulated for a counterfactual situation, we assume that quantities of durable goods remain fixed:

$$\begin{aligned} x^h_{D,SIM} &= x^h_{G,BAS} \\ \Leftrightarrow \frac{e^h_{D,SIM}}{(1+\tau_{D,SIM})\,.p_G} &= \frac{e^h_{D,BAS}}{(1+\tau_{D,BAS})\,.p_G} \\ \Leftrightarrow e^h_{D,SIM} &= e^h_{D,BAS}.\frac{1+\tau_{D,SIM}}{1+\tau_{D,BAS}}. \end{aligned}$$

There are then two possibilities:

1. We also assume constant quantities for all non-durable commodity groups G:
$$e^h_{G,SIM} = e^h_{G,BAS}.\frac{1+\tau_{G,SIM}}{1+\tau_{G,BAS}}. \tag{33}$$
2. We use the estimated Engel curves to derive new budget shares.

If we use estimated Engel curves, we assume constant savings. The reason is that, given the cross-sectional nature of a Household Budget Survey, it would be hard to make any other meaningful predictions about savings. An intertemporal model would be needed to say more. Since savings are the

leftover from disposable income after expenditures for durables and all non-durables have been subtracted, we get – denoting disposable income by y^h:

$$\begin{aligned} y^h_{SIM} - \left(e^h_{D,SIM} + e^h_{ND,SIM}\right) &= y^h_{BAS} - \left(e^h_{D,BAS} + e^h_{ND,BAS}\right) \\ \Leftrightarrow e^h_{ND,SIM} &= e^h_{ND,BAS} + e^h_{D,BAS}\left(1 - \frac{1+\tau_{D,SIM}}{1+\tau_{D,BAS}}\right) + y^h_{SIM} - y^h_{BAS}. \end{aligned}$$

When then total non-durable expenditures are known, there are two possibilities to determine the budget shares spent on the constituent categories:

1. The Engel curve parameters saved in the estimation phase are used. This comes down to the same procedure as the one used in the imputation phase, explained in section 4. There is one difficulty in this case, if the suggested correction is applied, which transforms the expenditures which are used as covariates such that their distributions conform to the respective distributions in the HBS. In the imputation of the baseline expenditures this procedure is correct, but in a simulation we want to allow these distributions to change. A correction would need to be implemented, in such a way that total non-durable expenditures and total remaining expenditures after subtraction of the categories with many zeroes are allowed to change, but that yields the unchanged baseline expenditures when baseline tax parameters are used.
   One way to accomplish this for total non-durable expenditures is to determine the differences between the counterfactual and the baseline non-durable expenditures, then to apply the usual corrections to the distribution of the baseline expenditures, and finally to add to this corrected variable the differences that ware found in the simulation. The resulting variable contains the total non-durable expenditures which can be used as a covariate for the prediction of the budget shares of the categories with many zeroes. These budget shares are then to be multiplied with the not-corrected total non-durable expenditures in order to determine the expenditure levels on these categories and the total remaining expenditures. An equivalent procedure can than be applied in order to correct the latter as covariates.

2. Constant budget shares are assumed:

$$\begin{aligned} \omega^h_{G,SIM} &= \omega^h_{G,BAS} \\ \Leftrightarrow e^h_{G,SIM} &= e^h_{G,BAS} \cdot \frac{e^h_{ND,SIM}}{e^h_{ND,BAS}}. \end{aligned} \tag{34}$$

Once the new expenditures are simulated, one can use equations (14)-(17) to determine the respective tax liabilities. To do so, we need to assume that the

budget shares within the categories are equal to those of the entire population in the HBS:

$$\frac{e^h_{k,SIM}}{e^h_{G,SIM}} \equiv \frac{e_{k,HBS}}{e_{G,HBS}} \equiv \omega_{k,HBS}. \tag{35}$$

This allows us to find the total tax liability, assuming constant producer prices p_k:

$$\begin{aligned} T^h_{G,SIM} &= \sum_{k\in G} T^h_{k,SIM} \\ &= \sum_{k\in G} \frac{\tau_{k,SIM}}{1+\tau_{k,SIM}} e^h_{k,SIM} \\ &= e^h_{G,SIM} \cdot \sum_{k\in G} \frac{\tau_{k,SIM}}{1+\tau_{k,SIM}} \omega_{k,HBS}. \end{aligned} \tag{36}$$

The liability for specific excises:

$$\begin{aligned} T^{h,a}_{G,SIM} &= \sum_{k\in G} T^{h,a}_{k,SIM} \\ &= \sum_{k\in G} a_{k,SIM} x^h_{k,SIM} \\ &= \sum_{k\in G} \frac{a_{k,SIM}}{(1+\tau_{k,SIM})\, p_{k,SIM}} \frac{e^h_{k,SIM}}{e^h_{G,SIM}} e^h_{G,SIM} \\ &= e^h_{G,SIM} \cdot \sum_{k\in G} \frac{1+\tau_{k,BAS}}{1+\tau_{k,SIM}} \frac{a_{k,SIM}}{q_{k,BAS}} \omega_{k,HBS}, \end{aligned} \tag{37}$$

The liability for ad valorem excises:

$$\begin{aligned} T^{h,v}_{G,SIM} &= \sum_{k\in G} T^{h,v}_{k,SIM} \\ &= \sum_{k\in G} v_{k,SIM} e^h_{k,SIM} \\ &= e^h_{G,SIM} \cdot \sum_{k\in G} v_{k,SIM} \omega_{k,HBS}. \end{aligned} \tag{38}$$

And the VAT liability:

$$\begin{aligned} T^{h,t}_{G,SIM} &= \sum_{k\in G} T^h_{k,SIM} \\ &= \sum_{k\in G} \frac{t_{k,SIM}}{1+t_{k,SIM}} e^h_{k,SIM} \\ &= e^h_{G,SIM} \cdot \sum_{k\in G} \frac{t_{k,SIM}}{1+t_{k,SIM}} \omega_{k,HBS}. \end{aligned} \tag{39}$$

6 Impact on welfare

Although policy makers are mostly interested in the financial impact of the reforms, researchers are more interested in the welfare effects implied by the changes in the quantities consumed. The SInTax model does not natively offer any welfare calculations. One possible measure for the welfare gain of a tax reform could be the following:

$$\begin{aligned} WG &= \sum_G q_{G,BAS} \cdot \left(x^h_{G,SIM} - x^h_{G,BAS}\right) \\ &= \sum_G \left(\frac{1+\tau_{G,BAS}}{1+\tau_{G,SIM}} e^h_{G,SIM} - e_{G,BAS}\right), \end{aligned}$$

where the assumption was used that producer prices remain constant.

Part II
Using the model

This part of the manual explains how the user can implement SInTax for a new country. It follows the order of the tasks to be performed. Readers who already have a working implementation of SInTax, and who just want to know how to run a simulation, are referred to subsection 8.2.

The SInTax program consists of two components, both written as Stata do-files[8]. The first component, an estimation program, takes as its input a household budget survey (HBS) dataset, and uses this to estimate Engel curves specified by the user. The second component, a prediction program, uses these estimates together with aggregate statistics from the HBS to impute expenditures into any other dataset, provided that the necessary covariates are available, and to use the imputed data to simulate expenditures and indirect tax liabilities in counterfactual situations.

The input files used by the model, as well as the output files generated by it, are text files with information on the level of individual households. This makes it possible to link to other microsimulation models, combining e.g. income tax and VAT reforms. The latest versions of the EUROMOD tax-benefit simulator have been adapted in order to facilitate the integration of the SInTax model.

First the user should choose a functional specification for the Engel curves and implement it in the estimation program. This step needs to be performed

[8]For more information about Stata, see http://www.stata.com/

only once for a given budget survey and specification of the Engel curves. It is the subject of section 7. Next the simulation program must be adapted such that it mirrors the operations in the estimation program. Once this is done, the simulation program can run independently from the estimation program. This is explained in section 8. The integration of SInTax into EUROMOD is explained in section 9.

7 Estimations of Engel curves

The estimation program estimates Engel curves using Household Budget Survey (HBS) data and writes the resulting parameters to Stata data-files (.dta). These parameters will later be read by the simulation program.

In section 2 we explained which steps are followed in the process of estimating Engel curves. These steps are performed internally by the SInTax estimation program using OLS regressions.

The user must adapt or create a number of do-files that perform the following tasks:

1. The expenditures in the HBS must be aggregated into a number of categories;
2. The covariates for the different regressions must be specified;
3. A file must be created containing statistics about the HBS;
4. The *userconfig.do* configuration file must be adapted to the new set-up;

Subsection 7.1 will further elaborate on each of these steps.

Once these steps have been completed, estimations can be performed using the do-file "*estimate.do*". More information about this step is given in subsection 7.2.

7.1 Preparing a new Household Budget Survey

The *_estimation* subdirectory of the SInTax folder contains a subdirectory *countries* which contains subdirectories for each country. The user needs to create a subdirectory for the country he is implementing. Within this country-specific directory the files are to be created that perform the steps listed above. These files are the topic of the current subsection. A more general overview of the directory structure of the model is given in appendix A.

Subsection 7.1.1 will explain how a do-file should be created that loads the HBS data, generates the necessary covariates for the estimations and aggregates expenditures on individual commodities into categories. Next, subsection 7.1.2 explains how a do-file is to be created that specifies which covariates are to be used in the estimations. Subsection 7.1.3 then explains how a commodity information file is to be created, containing the necessary information about the commodities in the HBS, such as population expenditures on individual commodities and the categories to which they belong. Finally everything must be put together in a user configuration file, explained in subsection 7.1.4.

7.1.1 Conversion of the HBS data

Two tasks are to be performed in this step: the variables available in the HBS should be converted to the format that is required by the choice of covariates, and the expenditures on individual commodities should be aggregated into a number of categories.

As the next subsection will explain in more detail, the user will be required to specify which covariates are to be used in the estimations. It is possible that these covariates are not available as such in the Household Budget Survey data. The user could opt for example to use the logarithm of disposable income as a covariate, but the Household Budget Survey only contains levels. Or the user could want to use category dummies, while the HBS encodes this information into single variables.

The second task to be performed in this step is aggregating individual commodity expenditures into a smaller number of categories. The user must define one category of durable expenditures, and a number of categories of non-durable expenditures.

A do-file *generate_data.do* performing these tasks must be created in the country-specific directory. It must create a new dataset based on the Household Budget Survey, containing one row for each household, containing all the necessary covariates, and containing aggregated expenditures per category. The do-file should save the resulting dataset using the following command:
save "$BUDGET_DATA_FILE", replace
where the global variable *$BUDGET_DATA_FILE* is defined internally in the estimation program.

7.1.2 Specification of the covariates used in the estimations

There are a number of different estimations that are done by the program:

1. OLS estimation of total non-durable expenditures;

2. Probit estimation of positivity of durable expenditures;

3. OLS estimation of durable expenditures if they are positive;

4. Probit estimation of positivity of expenditures on the non-durable categories with many zeroes;

5. OLS estimation of expenditure shares out of total non-durable expenditures on these categories with many zeroes, conditional on their positivity;

6. OLS estimation of expenditure shares out of remaining non-durable expenditures on the remaining categories.

The covariates for these regressions must be configured by the user in the file *define_covariates.do*, to be created in the country-specific directory for the estimations. This do-file must define the following global variables, referring to variables available in the data file generated in subsection 7.1.1:

- $INCOME: the variable containing disposable income
  E.g.: *global INCOME "disp_inc"*

- *$DUR_EXP*: the variable containing total durable expenditures.
  E.g.: *global DUR_EXP "agg_dur"*

- *$NONDUR_EXP*: the variable containing total non-durable expenditures.
  E.g.: *global NONDUR_EXP "totexpnondur"*

- *$DUR_DEP*: the dependent variable in the regression of total durable expenditures. This can be for example the level or the logarithm of total durable expenditures.
  E.g.: *global DUR_DEP "logtotexpdur"*

- *$NONDUR_*DEP: the dependent variable in the regression of total non-durable expenditures. This can be for example the level or the logarithm of total non-durable expenditures.
  E.g.: *global NONDUR_DEP "logtotexpnondur"*

- *$NONDUR_AGGS*: a list of the variables containing aggregated expenditures on the different categories of non-durable goods
  E.g.: *global NONDUR_AGGS "agg_1 agg_2 agg_3 ..."*

- *$ZEROES_AGGS*: the expenditure categories which are zero for many households, and thus require an extra probit estimation. Possible examples are rent, education, smoking and public transport.
  E.g.: *global ZEROES_AGGS "agg_3 agg_6 agg_10 agg_13"*

- $REMAINING_EXP: the variable containing remaining durable expenditures after subtraction of those with many zeroes
  E.g.: *global* REMAINING *"remaingexp"*

- *$COVLIST_DUR*: covariates for the estimations for total durable expenditures, as accepted by the Stata *reg* command

- *$COVLIST_NON_DUR*: covariates for the estimation of total non-durable expenditures, as accepted by the Stata *reg* command

- *$COVLIST_ZERO_GRP*: covariates for the estimations for the categories with many zeroes, as accepted by the Stata *reg* command

- *$COVLIST_REMAINING_GRP*: covariates for the estimations for the remaining categories, as accepted by the Stata *reg* command

7.1.3 Generating information about the HBS

A do-file *generate_commodity_info.do* should be created in the country-specific directory for the estimations. This do-file should create a dataset containing a row for each of the individual commodities in the Household Budget Survey (index k in Part I above), so not aggregated into categories. It should contain at least the following variables:

- *commodity_id*: a unique identifier for the commodity (index k above)

- *commodity_name*: an explanatory name for the commodity

- *category*: the aggregate category to which the commodity belongs (index G in Part I above)

- e: population-aggregated expenditures on commodity

The resulting dataset should be saved using the following command:
save "$HBS_COMMODITY_INFO_FILE", replace
where *$HBS_COMMODITY_INFO_FILE* is a global variable that has been created internally by the estimation program.

The variables *category* and e should be numeric. The numerical labels used for the *category* variable should uniquely identify the categories, but

commodity_id	commodity_name	e	category
111101	Rijst en rijstvlokken	4293082	Food, non-alcoholic beverages
111201	Tarwemeel, bloem, zelfrijzend meel	5228453	Food, non-alcoholic beverages
111202	Maïsmeel, rijstmeel en ander meel	823282.7	Food, non-alcoholic beverages
111203	Griesmeel	398118.4	Food, non-alcoholic beverages
111204	Havervlokken, havermout	294721.3	Food, non-alcoholic beverages
111302	Deegwaren verpakt met andere producten	548415.4	Food, non-alcoholic beverages
111303	Macaroni, spaghetti, vermicelli, enz.	1.11e+07	Food, non-alcoholic beverages
111304	Verse deegwaren	335640.5	Food, non-alcoholic beverages
111401		1.64e+07	Food, non-alcoholic beverages
111402		3301234	Food, non-alcoholic beverages
111403		2.31e+07	Food, non-alcoholic beverages
111404		4859312	Food, non-alcoholic beverages
111405		316032	Food, non-alcoholic beverages
111406		363411.9	Food, non-alcoholic beverages
111407		8726925	Food, non-alcoholic beverages
111408		2.38e+07	Food, non-alcoholic beverages
111411		689651.9	Food, non-alcoholic beverages
111412		2780200	Food, non-alcoholic beverages
111413		2324731	Food, non-alcoholic beverages
111414		6839388	Food, non-alcoholic beverages
111415		1828291	Food, non-alcoholic beverages
111426	Rozijnenbrood	2743986	Food, non-alcoholic beverages

Figure 1: An example of a file containing commodity information.

have no further significance. The user is free to create labels for them, which can help keeping the resulting commodity information file readable:
label define categoryLabel 'categoryNr' "'categoryName'", add
...
label values category categoryLabel

Figure 1 shows an example of what this file could look like.

7.1.4 User configuration

The *_estimation* directory should contain a configuration file *userconfig.do*. This file should contain the following global macro variables:

- *$COUNTRY_FULL*: Full name of the country under investigation. This string should be equal to the name of the country-specific directory. E.g.: *global COUNTRY_FULL "belgium"*
- *$BUDGET_YEAR:* The expenditure-reference year from the HBS data. E.g.: *global BUDGET_YEAR = 2005*

All these global variables are available throughout the do-files to be created by the user.

7.2 Running an estimation

Now that everything is set up, running an estimation is straightforward. Use the Stata command *cd* to change the working directory to the one containing the file "*estimate.do*", which should be in the */_estimation* subdirectory of the SInTax package, and run the estimation:

run estimate.do

The estimations are performed and auxiliary files are created, all of which will be placed in the "*/_prediction/*countries*/$COUNTRY_FULL*" directory.

The following files are created:

- *hbs_category_info_bud$BUDGET_YEAR.dta*:
  population expenditures and implicit tax rates per category; file created by the model
- *hbs_commodity_info_bud$BUDGET_YEAR.dta*:
  the commodity information file discussed in subsection, with implicit tax rates added to it by the model
- *hbs_income_distribution_bud$BUDGET_YEAR.dta*:
  mean value and variance for disposable income in the HBS, can be used for correcting the distribution as described in sections 4 and 5.
- *hbs_remaining_expenditures_distribution_bud$BUDGET_YEAR.dta*:
  mean value and variance for remaining non-durable expenditures after subtraction of the categories with many zeroes in the HBS, can be used for correcting the distribution as described in sections 4 and 5.

and in the *estimation_parameters* subdirectory all the estimation parameters are saved, with an indication of *$BUDGET_YEAR* in the file name.

8 Simulating consumer demand

The information in the files saved by the estimation program is used to predict expenditures for any household in a target dataset (such as the SILC-dataset in EUROMOD), as long as the necessary covariates are present. This section first explains how a new country can be implemented for this purpose. Note that before the user can run a simulation, the baseline expenditures need to be imputed into the target dataset. This must be done, whichever of the methods of simulation described in section 5 is chosen. Subsection 8.1 will explain how do-files should be created to prepare the target dataset and to

use the estimated Engel curves for the imputation of baseline expenditures and the simulation of counterfactual situations. Subsection 8.2 will then explain how, once the system is set up, the user can impute the baseline expenditures and run a simulation.

8.1 Preparing the simulation program for a new country

Before being able to run a simulation, a number of preparations need to be done. The target household data will need to be manipulated so that it contains the same covariates as the ones used in the estimations. If the data are from a different year than the HBS, amounts need to be adjusted for inflation where applicable. This preparation of the target household data will be explained in subsection 8.1.1. Next the user needs to specify how expenditures are to be predicted, both for imputations and simulations, using the Engel curves parameters saved by the estimation program. This will be the topic of subsection 8.1.3. Subsection 8.1.4 discusses how the resulting expenditures should be uprated again to the income year of the target household data. And finally, just like with the estimation program, a user configuration file *userconfig.do* needs to be created to put everything together. This is explained in subsection 8.1.5.

8.1.1 Preparation of the target household data

The target household data can come from any source, as long as the necessary variables for the predictions, i.e. the explanatory variables from the estimations, are available. One specific example is the output of a tax-benefit microsimulation model such as EUROMOD. These target data need to be converted to a format that can be used by the prediction program. An advantage of this flexibility is that the sequential usage of different microsimulation models becomes possible, enabling e.g. the combined simulation of direct tax-benefit reforms and indirect tax reforms.

This flexibility though makes that the conversion of the target dataset must happen again, each time a simulation is run in SInTax. After all, there is no reason why the output from preceding models should always be the same.

To do so, the simulation program calls a do-file, configured in the global variable
$HH_CONVERT_FILE. This do-file is to be created by the user in the country-specific directory */_prediction/*countries*/YOUR_COUNTRY*. The global variable referred to, as well as the ones referred to below, are given by

the user as command line parameters, or are configured in the configuration file *userconfig.do, as* explained in subsection 8.1.5.

A global variable *$HH_IN_FILE* is configured in the same way, containing the path to the household data file to be converted. The do-file $HH_CONVERT_FILE should open this file, and perform the following steps:

1. Create the covariates that are necessary for the predictions. These should mirror the covariates that were used in the estimations. The variables should have the same names.
   It is possible that at this point some covariates cannot yet be created. For example, to be able to create the variable containing the logarithm of total non-durable expenditures, needed if e.g. the Engel curves from the QUAIDS system are used, first total non-durable expenditures need to be predicted. These variables will later be created in the prediction phase, when the necessary variables are available, as explained in subsection 8.1.3.

2. Create scalars *hhPriceIndex* and *budgetPriceIndex*, containing the price levels for the year of the target household dataset and of the HBS data that were used for the estimations. The former year is available to the user as a global variable *$HH_IN_YEAR*, the latter as *$BUDGET_YEAR*, both configured as explained above. Suggestions for price indices are consumer price indices or GDP deflators, as available from statistical offices of the country in question. The scalars created will be available to the other do-files throughout the simulation.

3. Uprate amounts where necessary, using the scalars created in the previous point. The QUAIDS system Engel curves for example have the logarithm and the logarithm squared of real disposable income as covariates. In order to correct for the difference in price levels in the budget survey year and the year of the of the counterfactual input data, divide amounts in the input data by *hhPriceIndex* and multiply by *budgetPriceIndex.*[9].

Lastly, the resulting dataset should be saved as follows:
save "$HH_CONVERTED", replace
where the global variable *$HH_CONVERTED,* referring to a temporary file for internal use, is defined internally by the prediction program.

[9]Note that this is different from correcting for the change in prices caused by a tax reform. This is done in the do-files that handle the simulation, as explained in subsection 8.1.3.

8.1.2 Creating the baseline tax code file

To be able to simulate the impact of a tax reform on consumer demand, first it needs to be known which tax parameters were in place in the baseline.

For each of the individual commodities, the following variables should be provided:

- *vat*: the VAT rate raised on this commodity in the baseline scenario
- *excise_specific*: the specific (per unit) excise rate raised on this commodity. Set this to zero if you do not plan to simulate excises.
- *unit*: the unit for which the specific excise is defined (e.g. a litre of beer or a single cigarette). This is a merely informative text variable.
- *excise_ad_valorem*: the ad valorem excise levied on the commodity. Set this to zero if you don't plan to simulate excises.
- q: the consumer price of this commodity, in the units indicated in the *unit* column and referred to in the *excise_specific* column. This can be set to 1 for goods on which no specific excise is levied, as this variable is only used to calculate the implicit tax rates. For more information please refer to part I of this manual.

Except for the variable q, all of these variables should be formatted as text. These can contain numbers, but they can also contain place-holders, which are to be replaced when the user runs a simulation. For example, VAT rates of "0.21" could be replaced by a place-holder "$VAT3", denoting the third and regular VAT rate in Belgium.

A do-file *create_baseline_commodity_info.do* should be created in the country-specific directory which opens the HBS commodity file:
use "$HBS_COMMODITY_INFO_FILE", clear
and adds the tax information variables listed above. The resulting dataset should be saved to the baseline commodity information file:
save "$BAS_COMMODITY_INFO_FILE", replace

When running a simulation, the user is required to indicate the values to be used for the place-holders, for example *$VAT3=0.21*. More information about this will be given in subsection 8.2.

8.1.3 Predictions

There are several possible methods to predict household expenditures into the target household data, using either constant quantities or using Engel

curves. These methods were explained in section 5. Whichever method is chosen, the expenditures in the baseline tax system must first be imputed for each household. This is done by using the Engel curves parameters produced in the estimation stage, and described in section 7, to predict expenditures in the baseline indirect tax situation.

If the target dataset into which the baseline expenditures are to be imputed is the output of another model, the user should make sure to use the output of a baseline run of that model.

The user should create two do-files in the country-specific directory. The first, *impute.do*, does the imputations into the target dataset, following the procedure explained in section 4, and using the Engel curves parameters from the estimation phase. The second do-file, *simulate.do,* is similar to the first, now following the procedure explained in section 5, again using the Engel curves parameters. The resulting expenditure variables for each household should in each case be written to a Stata datafile defined in global variable *$EXPENDITURE_OUT_FILE*, to be specified as a command line argument or in the user configuration file.

In both do-files, the user has at her disposal a number of Stata functions, defined internally by the simulation program. These functions will automatically retrieve the estimation parameter files written by the estimation program, so the user does not need to worry about this. For the predictions to work, the same covariates as those used in the estimations must be available, with the same names, as should have been prepared following the instructions in subsection 8.1.1.

- **predictNondur** *varname*: predict non-durable expenditures
  Using the parameters from the Engel curves estimation and the covariates available, the dependent variable for non-durable expenditures is predicted. The resulting variable is saved as *varname*. This variable does not need to be in levels, it can also be e.g. in logarithms, depending on what was estimated in section 7. This function should only be used in the imputation stage, as in the simulation stage total non-durable expenditures are determined as a residual after the assumption of constant savings and constant quantities of the durable expenditures.
  Besides the predicted non-durable expenditures, the following return values will be available:

  - *r(meanExpResid)*: the mean of the exponents of the residuals in the estimation. Refer to section 4 to see why this is useful for correcting predicted expenditure levels.
  - *r(varResid)*: the variance of the residuals in the estimation. This

may be useful if other correction methods are used than those explained in section 4.

- **getHbsIncomeDistribution**: retrieve information about the HBS distribution of disposable income
  This can be used in order to correct the distribution of the covariates in the imputations and the predictions, as explained in section 4. The properties of the distribution are available as the following return values:
  - r(income_mean): mean
  - r(income_sd): standard deviation

- **getTotExpNonDurDistribution**: retrieve information about the HBS distribution of total non-durable expenditures
  This can be used in order to correct the distribution of the covariates in the imputations and the predictions, as explained in section 4. The properties of the distribution are available as the following returns values:
  - r(nondur_mean): mean
  - r(nondur_sd): standard deviation
  - r(nondur_max): maximum value

- **predictDurPositive** *varname*: predict the probability of positive durable expenditures
  Using the parameters from the estimated Engel curves and the covariates available, it is predicted for each household whether they have strictly positive durable expenditures. The resulting variable is saved as *varname*, a value of 1 indicating positive expenditures, a value of 0 indicating zero expenditures.
  This function should only be used in the imputation stage, as in the simulation stage constant quantities as assumed for durable consumption.

- **predictDur** *varname positive_dur*: predict durable expenditures
  Using the parameters from the estimated Engel curves and the covariates available, the dependent variable for durable expenditures is predicted, conditional on the variable *positive_dur* written by the function *predictDurPositive*. The resulting variable is saved as *varname*. This variable does not have to be in levels, it can also be e.g. in logarithms,

depending on what was estimated in section 7. This function should only be used in the imputation stage, as in the simulation stage constant quantities as assumed for durable consumption.
Besides the predicted non-durable expenditures, the following return values will be available:

  - *r(meanExpResid)*: the mean of the exponents of the residuals in the estimation. Refer to section 4 to see why this is useful for correcting predicted expenditure levels.
  - *r(varResid)*: the variance of the residuals in the estimation. This may be useful if other correction methods are used than those explained in section 4.

- **getTotExpDurDistribution**: retrieve information about the HBS distribution of total durable expenditures
  This could be used in order to correct the distribution of the covariates in the imputations and the predictions, as explained in section 4. The properties of the distribution are available as the following return values:

  - r(dur_mean): mean
  - r(dur_sd): standard deviation
  - r(dur_max): maximum value

- **getZeroesAggs**: get list of categories of non-durable expenditures with many zeroes
  Returns as *r(zeroesAggs)* a list of expenditure categories with many zeroes. For these categories, a separate probit estimation was performed by the estimation program. If there were no such groups selected by the user, the string value "*NONE*" is returned.
  The user should loop through these categories, and use the following functions to predict expenditure shares on them.

- **predictZeroExpPositive** *varname category*: predict the probability of positive expenditures for a category with many zeroes
  Using the parameters from the estimated Engel curves and the covariates available, predict whether expenditures on *category* are zero, and save the result in *varname*. A value of 1 indicates positive expenditures, a value of 0 indicates zero expenditures.

- **predictZeroExpShare** *varname positive_exp category*: predict expenditure share for a category with many zeroes
  Using the parameters from the estimated Engel curves and the covariates, predict the share of non-durable expenditures on *category* for households where *positive_exp* is equal to one. *positive_exp* should be the variable generated by *predictZeroExpPositive*. The resulting expenditure shares are saved in *varname*. Be careful, predicted shares can be negative, and you may want to set them to zero.

- **getRemainingExpDistribution**: retrieve information about the HBS distribution of remaining non-durable expenditures after subtraction of those for which a probit estimation was used
  Using information from the functions above, namely total non-durable expenditures and budget shares on categories for which probit estimations were used, it is possible to determine the levels of the total remaining expenditures. When these are to be used as a covariate in a prediction, the information from the following return values can be used in order to correct the distribution, as explained in section 4:

  - r(remainingExp_mean): mean
  - r(remainingExp_sd): standard deviation

- **getNonDurVars**: get categories of non-durable expenditures
  Returns as *r(nonDurVars)* a list of all non-durable expenditure categories, including those with many zeroes.
  The user should loop through the categories for which no expenditure shares have been predicted yet, and do so.

- **predictExpShare** *varname nonDurVar*: predict expenditure share for the given category
  Using the estimation parameters and the covariates, predict the expenditure share on *nonDurVar*, and store the results in *varname*. Be careful, predicted shares can be negative, and you may want to set them to zero and renormalise the new shares such that they sum to one.

Make sure that before calling any of these functions, the necessary covariates are available. E.g. before predicting the expenditure shares using *predictExpShare*, make sure that the remaining expenditure variable and any of its required transformations and interaction terms have been constructed.

When you are done with all this, calculated total durable and non-durable expenditures and non-durable expenditure shares to calculate the expenditures on all of the categories, saving them as variables *x'suffix'_s*, where *suffix* is the same indicator for the category as was used in the commodity information files, as discussed in subsection 7.1.3. In this process, consider applying the corrections suggested in sections 4 and 5, as demonstrated in the Belgian code examples.

The indirect tax reform itself may change the overall price level. If for example all VAT rates increase, while disposable income remains the same, the price level increases and real incomes decrease. Since the estimations may depend on real values of the covariates – as is the case e.g. with QUAIDS – this needs to be corrected for. The user is advised to calculate a deflator in his file *simulate.do* to handle this. Such deflator could for example take the following, household-specific value:

$$\prod_{G} \left(\frac{1 + \tau_{G,SIM}}{1 + \tau_{G,BAS}} \right)^{\omega^{h}_{G,BAS}} . \tag{40}$$

The reader is referred to part I for information about the notation.

The resulting expenditures should be written to the file specified in the global variable *$EXPENDITURE_OUT_FILE*. Only the the variables *idhh*, containing a household identifier, and *x**, containing the predicted expenditures, should be kept.

8.1.4 Uprating

The predicted expenditures are now in the price level of the HBS reference year. The user must create create a do-file *uprate.do* in the country-specific directory which uses the scalars *hhPriceIndex* and *budgetPriceIndex* that were discussed in subsection 8.1.1 to bring the incomes and expenditures to the price level of the target household file. Tax liabilities will be calculated internally based on the uprated expenditures, following the procedures in section 5.

8.1.5 User configuration

As will become clear in subsection 8.2, a number of configuration parameters are set from the command line:

- the country for which to run a simulation
- the income year for the household file

- the year of the household budget survey from which estimations should be used
- the year of the indirect tax code
- the method of simulation

These arguments must be passed from the command line, because they will generally be different for each simulation. Some parameters though are less likely to change. Standard values for these should be configured as global macro variables in a configuration file *userconfig.do*, in the directory */_simulate*. The user can still change their values through the command line. These global variables will be available throughout the do-files in the prediction program, and some have been referred to in preceding sections:

- *$OUT_PATH*: directory where the output file is to be written to
  E.g.: *global* OUT_PATH = "output"
  Both log files and resulting expenditure files will be written in this directory. Either a full path, or a directory relative to *c:/sintax/*.
- *$EXPENDITURE_OUT_FILE:* file to save results of the simulation.
  E.g.: *global EXPENDITURE_OUT_FILE "be_${HH_IN_YEAR}_tco.txt"*
  Here *$OUT_PATH* will be added internally to form a full path.
- *$IN_PATH*: path containing the target household dataset
  E.g.: *global IN_PATH = "'emOutDir'" // directory in which target household dataset is to be found*
- *$HH_IN_FILE*: file containing household information, referred to as "the target dataset", e.g. output from EUROMOD
  E.g.: *global HH_IN_FILE "tmp_${COUNTRY_FULL}_${HH_IN_YEAR}_TCO.txt"*
  The global variables *$COUNTRY_FULL* and *$HH_IN_YEAR* are defined in subsection 8.2. Using these global variables in the path allows for some flexibility in using different input files for different countries and income years.
- *$TAX_IN_FILE*: file containing tax parameters
  E.g.: *global TAX_IN_FILE ...*
  "c:/em/output/${COUNTRY_FULL}_${HH_IN_YEAR}_TAX.txt"
  This file should be in the format explained in subsection 8.2.1.
- *$BAS_EXPENDITURE_FILE*:
  E.g.: *global BAS_EXPENDITURE_FILE "'baselinePath'/'baselineFile'"*

File containing baseline expenditures; necessary for simulations of counterfactual situations. This is a file that was defined as *$EXPENDITURE_OUT_FILE* in a run with "impute" as method (see section 8.2).

- *$HH_CONVERT_FILE*: Do-file for converting target household dataset to format required by the prediction program
  E.g.: *global HH_CONVERT_FILE "convertHH_EUROMOD.do"*

8.2 Running a prediction

Once the preparatory steps set forth in subsection 8.1 are done, a simulation can be performed using the do-file *predict.do* found in the directory */SInTax/_prediction*.

This do-file requires a number of command line arguments, as explained in subsection 8.2.2. Before this do-file can run, a tax parameter file should be in place. We will discuss this first, in subsection 8.2.1.

8.2.1 The tax parameter file

This text file contains the tax parameters that are to be used in the simulation. It contains two rows: the first row contains text strings, the second row contains the values corresponding to those text strings. These are instructions to the model for how to replace the place-holders that were created (see subsection 8.1.2). An example of what the file could look like:

$vat0	$vat1	$vat2	$vat3
0	6	12	21

This file could be written within and outputted from EUROMOD, so that the tax parameters can be specified in the EUROMOD user interface. How to do this is explained in section 9.

When the user is running an imputation, the baseline tax rates should be used.

8.2.2 Calling predict.do

Before executing *predict.do*, the working directory must be changed to the directory containing the do-file (typically */_prediction*). This do-file then is to be called with arguments:

- ***country***: the full name of the country for which an estimation is performed. This tells the simulation program in which country directory to look for the parameters that were saved by the estimation program.

The value of this argument is available throughout the simulation as the global variable *$COUNTRY_FULL*.

- ***income***: the income year of the household data on which the simulation will be performed. This may be the policy year from EUROMOD, if SInTax is coupled to EUROMOD. The value of this argument is available throughout the simulation as the global variable *$HH_IN_YEAR*.

- ***budget***: the expenditure reference year from the HBS data from which the estimation parameters should be used in the simulation. This helps the simulation program to select the correct estimation parameters. This parameter is also used for uprating. The value of this argument is available throughout the simulation as the global variable *$BUDGET_YEAR*.

- ***ityear***: the year of the indirect tax code to use. This allows the simulation program to find the commodity information files that were created as explained in subsection 7.1.3, with tax parameters and placeholders added as explained in subsection 8.1.2. The value of this argument is available throughout the simulation as the global variable *$IT_TAX_CODE_YEAR*.

- ***method***: an indication of which algorithm to use in order to predict expenditures. Current possibilities are:

  - *impute*: predict baseline expenditures into the target household dataset
  - *const_quant*: simulate counterfactual situation, assuming quantities are the same as in the baseline
  - *const_shares*: simulate counterfactual situation, assuming budget shares are the same as in the baseline
  - *simulate*: simulate counterfactual situation, using the parameters from the estimated Engel curves

Examples:

- Run a normal simulation on the latest EUROMOD output (assuming that all paths are correctly specified in *userconfig.do*):
  run "predict.do" country=Belgium income=2012 budget=2009 ityear=2011 method=SIMULATE

- Impute baseline expenditures into baseline output from EUROMOD: *run "predict.do" country=Belgium income=2012 budget=2009 ityear=2011 method=IMPUTE*

When running a simulation, the following steps are performed by the program:

1. The *userconfig.do* file is loaded (see subsection 8.1.5)

2. From this, it knows where to find the file containing the tax rates for the simulation. This, together with the files created in subsections section 7.1.3 and 8.1.2, allows the program to replace the place-holders in the commodity information file, and to calculate the new aggregate implicit tax rates. These baseline and simulation aggregate implicit tax rates are available throughout the simulation program in scalars *basImplTax_* suffix and *simImplTax_* suffix, where *suffix* refers to the category number defined in the commodity file.

3. New expenditures are predicted using one of the files created in subsection 8.1.3, depending on the method specified.

4. Indirect tax liabilities are inferred and stored in the variables *tva'suffix'_s*, *texav'suffix'_s* and *texsp'suffix'_s*, conforming to EUROMOD standards.

5. The household identifier variable *idhh* is renamed to *idhh_head*. The results are saved to the file *$EXPENDITURE_OUT_FILE* specified by the user.

A log of the simulation is written to */_prediction/output/simulate.log*

9 Integrating SInTax into EUROMOD

A major advantage of the SInTax simulation program is that it can be called from within EUROMOD. In order to do so, a new policy sheet should be created in the EUROMOD parameter file, at a point in the spine where disposable incomes are available. In this policy sheet, the following EUROMOD functions should be present[10]:

[10]Not all of the EUROMOD functions used are available in earlier versions of EUROMOD. The method described below should work correctly for those working in the new, non-Excel, user interface.

Figure 2: How to write the tax code file

	Policy	Gr...	BE_2013	Comment
38.1	DefConst		on	VAT rates
38.1.1	$VAT0		0	
38.1.2	$VAT1		6	
38.1.3	$VAT2		12	
38.1.4	$VAT3		21	
38.2	Totals		on	
38.2.1	TAX_UNIT		tu_household_be	
38.2.2	Varname_Min		min	
38.2.3	Agg_Var		idhh	
38.3	Elig		on	
38.3.1	Elig_Cond		{idhh=min_idhh}	
38.3.2	TAX_UNIT		tu_household_be	
38.4	DefOutput		on	
38.4.1	Who_Must_Be_Elig		one	
38.4.2	File		tmp_be_2013_TAX	
38.4.3	VarGroup		$VAT*	
38.4.4	TAX_UNIT		tu_household_be	

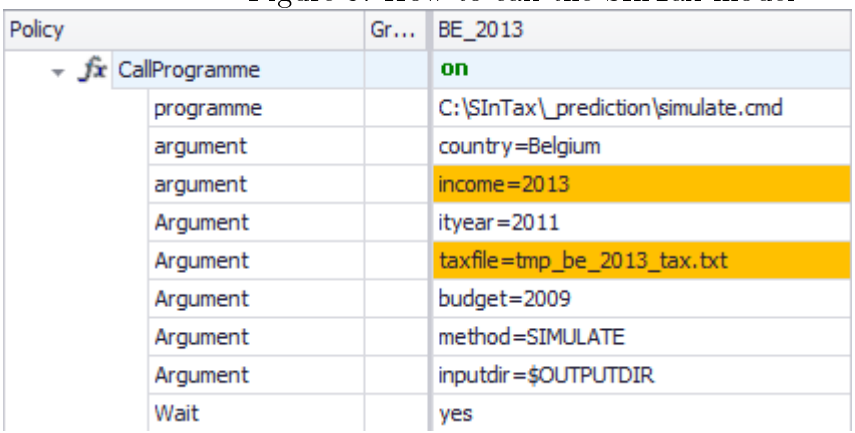
- Use function *DefOutput* to write all variables necessary for the simulation to a file. This will typically include the household id, disposable income and demographic parameters. This is the file to be configured as *$HH_IN_FILE* in *userconfig.do* or in the input parameters of the prediction program, as explained in subsection 8.2.2.

- Use function *DefOutput* in order to write the tax code file to the path configured as *$TAX_IN_FILE* in *userconfig.do*. One way to do this is demonstrated in figure 2.

- Use the *CallProgramme* function to call the windows batch file */_prediction/simulate.cmd*, with as arguments the country, the year of the budget survey data and the method to use for the simulation. The command line arguments for this function are the same as for *predict.do*. See figure 3 for an example.

The batch file *simulate.cmd* refers to the System Environment Variable "STATA_EXE". This variable should refer to the path to your Stata executable (e.g. "C:/Stata12/statase.exe", including the quotation marks). Refer to your IT department in order to set

Figure 3: How to call the SInTax model



this variable on your PC.

This batch file also assumes that your *SInTax* directory is installed in the root of your *c:*-drive. If this is not the case, use the following command to link the *c:/SInTax* path to the actual path to your SInTax directory[11]:
mklink /j c:/SInTax x:/yourpath/SInTax
Alternatively, change the references to it in the batch file and the EUROMOD policy sheet.

If the user wishes to do so, he can merge the results of the SInTax simulation back into EUROMOD, using the *DefInput* function.

References

James Banks, Richard Blundell, and Arthur Lewbel. Quadratic engel curves and consumer demand. *The Review of Economics and Statistics*, 79(4): 527–539, November 1997. ISSN 00346535. URL `http://www.jstor.org/stable/2951405`.

Angus Deaton and John Muellbauer. *Economics and Consumer Behavior*, volume 9780521296762 of *Cambridge Books*. Cambridge University Press, 1980.

André Decoster, Richard Ochmann, and Kevin Spiritus. Integrating indirect taxation into euromod. documentation and results for germany. Technical report, EUROMOD at the Institute for Social and Economic Research, 2013.

[11] This command may not work in versions of Windows before Windows Vista. Please refer to your system administrator for how to create such a junction on your system.

André Decoster, Richard Ochmann, and Kevin Spiritus. Integrating indirect taxation into euromod. documentation and results for belgium. Flemosi Discussion paper 32, KU Leuven, 2014.

J. M. Wooldridge. *Introductory Econometrics - A Modern Approach*. Thomson, 2003.

A Directory structure

This is an overview of the directories in the SInTax folder, and the files that are to be created or adapted by the user.

- [_estimation]
  - [_internals]
  - [countries]
    - [EachCountry]
      - define_covariates.do (section 7.1.2)
      - generate_data.do (section 7.1.1)
      - generate_commodity_info.do (section 7.1.3)
  - estimate.do (section 7.2)
  - userconfig.do (section 7.1.4)
- [_prediction]
  - [_internals]
  - [countries]
    - [EachCountry]
      - [_estimation]
      - create_baseline_commodity_info.do (section 8.1.2)
      - convertHH.do (section 8.1.1)
      - impute.do (section 8.1.3)
      - simulate.do (section 8.1.3)
      - uprate.do (section 8.1.4)
  - [output]
  - simulate.cmd (section 9)
  - predict.do (section 8.2)
  - userconfig.do (section 8.1.5)