INTERNATIONAL JOURNAL OF MICROSIMULATION (2015) 8(3) 137-147

INTERNATIONAL MICROSIMULATION ASSOCIATION

Online appendix to:

Accounting for Tax Evasion Profiles and Tax Expenditures in Microsimulation Modelling. The BETAMOD Model for Personal Income Taxes in Italy

Andrea Albarea

Department of Economics, Ca' Foscari University
San Giobbe, Cannaregio 873, 30121 Venice, Italy
e-mail: andrea.albarea@unive.it

Michele Bernasconi

Department of Economics, Ca' Foscari University
San Giobbe, Cannaregio 873, 30121 Venice, Italy
e-mail: bernasconi@unive.it

Cinzia Di Novi

Department of Economics, Ca' Foscari University
San Giobbe, Cannaregio 873, 30121 Venice, Italy
e-mail: cinzia.dinovi@unive.it

Anna Marenzi

Department of Economics, Ca' Foscari University
San Giobbe, Cannaregio 873, 30121 Venice, Italy
e-mail: anna.marenzi@unive.it

Dino Rizzi

Department of Economics, Ca' Foscari University
San Giobbe, Cannaregio 873, 30121 Venice, Italy
e-mail: rizzid@unive.it

Francesca Zantomio

Department of Economics, Ca' Foscari University
San Giobbe, Cannaregio 873, 30121 Venice, Italy
e-mail: francesca.zantomio@unive.it

A. STATISTICAL MATCHING BETWEEN THE IT-SILC AND SHIW DATASETS

We describe here how the statistical matching between the IT-SILC dataset with the Bank of Italy's Survey on Households Income and Wealth (SHIW) at the household level was performed. First, two constraints need be satisfied to make matching feasible: (i) the two surveys must be random samples from the same population; (ii) there must be a common set of conditioning variables. In our case, the first condition is met by design, since both the IT-SILC 2011 and the SHIW 2012 data are representative of the Italian population. As far as the second constraint is concerned, the variables (X) common to each dataset and chosen for the process of imputation of self-reported asset value of the main residence, insurance premiums and house refurbishments expenditures are: equivalent household income, the percentage of household members with more than upper secondary educational qualification, a set of household composition dummies, and the main earner's employment status. The final sample is made up of 7.951 households from the SHIW survey and 19.399 households from the IT-SILC Survey.

The dataset, integrated by IT-SILC -Bank of Italy was created using the Mahalanobis Distance Matching Method (MDMM), a statistical method which allows individuals with similar characteristics but from different datasets to be paired (Rosenbaum and Rubin, 1983). In order to obtain a more precise matching, the sample was stratified in cells according to the main residence homeownership, other properties homeownership and geographical area so that exact matching on these variables is ensured; then, within each stratum, the donor household has been selected based on the Mahalanobis distance metric, measured on the other X variables. The Mahalanobis metric is a measure of dissimilarity between observation which measures the distance between units i from the recipient dataset IT-SILC and j from the donor dataset SHIW weighting each coordinate of X in inverse proportion to the variance of that coordinate:

$$d(i,j) = (x_i - x_j)\Sigma^{-1}(x_i - x_j)' \qquad [A.1]$$

where Σ is the variance–covariance matrix of X.

Matching has been performed at the household level and with replacement, that is allowing the same SHIW household to act as donor for multiple IT-SILC households, if deemed as the most adequate, rather than being discarded after having served once as donor. Once the matching procedure was complete, we check the quality of the matching. The quality of matching was evaluated in terms of maintaining the asset value of the main residence, insurance premiums and

house refurbishments expenditures distributions, both in terms of preserving the pre-existing variables distribution as well as in terms of pre-existing relations between variables of interest.

The next step was i) the comparison between the asset value of the main residence, insurance premiums and house refurbishments expenditures distributions in the integrated dataset and the pre-existing SHIW one, ii) the calculation of the correlation between asset value of the main residence, insurance premiums and house refurbishments expenditures distributions and the X vector to verify the maintenance of the sign recorded in the "donor set". The differences between the common-fusion correlations in the SHIW data set versus the fused IT-SILC data set were well preserved for most variables. For the sake of brevity, tables showing distributions and correlations are not included but they are available on request.

Finally, the quality of the matching has been evaluated in terms of "balancing test": we compared the mean covariate values in the recipients and matched donors i.e. each of the observable covariates within the recipients has the same average value within the matched donors. Before matching we expect differences, after matching the variables should be balanced in both groups and significant differences should not persist. The covariate balancing test, included in Table A1, shows that the matching is effective in removing differences in observable characteristics between the recipients and matched donors. In particular, the median absolute bias is reduced by approximately 82%-98%. The Pseudo R-squared after matching is always close to zero, correctly suggesting that the covariates have no explanatory power in the matched samples. The chi-square test conducted before and after matching, proves that the propensity score removed bias due to differences in covariates between the recipients and matched donors.

Table A.1 Balancing test

Property 1st	Property 2nd	Region N-W	Region N-E	Region C	Region S	Sample	Pseudo R2	LR test	p-values	Median bias	% reduction in median bias
✓		✓	✓			Before matching	0.013	99.72	0.000	6.4	0.969
						After matching	0.000	0.38	0.996	0.2	
✓				✓		Before matching	0.012	42.77	0.000	5.0	0.980
						After matching	0.000	1.23	0.942	0.1	
✓					✓	Before matching	0.019	102.27	0.000	5.8	0.931
						After matching	0.000	1.30	0.935	0.4	
✓	✓	✓				Before matching	0.007	11.84	0.037	6.7	0.955
						After matching	0.000	0.07	1.000	0.3	
✓	✓		✓			Before matching	0.001	2.50	0.776	4.2	0.929
						After matching	0.001	2.27	0.810	0.3	
✓	✓			✓		Before matching	0.007	14.36	0.013	6.0	0.933
						After matching	0.000	0.20	0.999	0.4	
✓	✓				✓	Before matching	0.004	9.31	0.097	7.0	0.971
						After matching	0.000	0.63	0.987	0.2	
	✓	✓	✓	✓	✓	Before matching	0.004	4.14	0.529	4.6	0.824
						After matching	0.001	1.46	0.917	0.8	
		✓	✓			Before matching	0.042	136.58	0.000	6.1	0.984
						After matching	0.000	1.95	0.857	0.1	
				✓	✓	Before matching	0.029	122.41	0.000	10.0	0.980
						After matching	0.000	0.51	0.992	0.2	

B. ESTIMATION OF TAX EVASION RATE

B.1. Tax evasion rates by income source type and geographical area

From official tax returns data (MEF) we know the total amount of reported income and the number of taxpayers by four *main* income source type (*EMP*=employment income, *PEN* =pensions, *IMM*=rental income from immovable property, *SELF*=self-employment income) and, separately, by four geographical area. From the BETAMOD simulated true gross incomes we compute the total amount of reported income and the number of taxpayers for the same characteristics. Then, it is possible to compute two sets of average tax evasion rates by *main* source of income *i:*

$$\bar{e}_i^S = \frac{\bar{y}_i^S - \bar{y}_i^{S\,MEF}}{\bar{y}_i^S} \qquad [B.1]$$

and by geographical area *j*:

$$\bar{e}_j^A = \frac{\bar{y}_j^A - \bar{y}_j^{A\,MEF}}{\bar{y}_j^A} \qquad [B.2]$$

To convert the tax evasion rates by main income source type of taxpayers ($\bar{e}_i^S$) into rates referred to types of income received ($\bar{e}_j^R$), we use the BETAMOD estimated true gross incomes to build a 4×4 matrix *B*, in which each element B_{ji} is total amount of true gross income of type *j* ($j = 1,...,4$) received by taxpayers with main source of income type *i* ($i = 1,...,4$). The total amount of unreported income by main source of income type *i* is computed as:

$$U_i^S = \left(\bar{y}_i^S - \bar{y}_i^{S\,MEF}\right) N_i^S \qquad [B.3]$$

where N_i^S is the number of taxpayers with main source of income type *i*.

With this information it is possible to compute the tax evasion rates by income source by solving the linear system:

$$\begin{bmatrix} B_{EMP,EMP} & B_{PENS,EMP} & B_{IMM,EMP} & B_{SELF,EMP} \\ B_{EMP,PENS} & B_{PENS,PENS} & B_{IMM,PENS} & B_{SELF,PENS} \\ B_{EMP,IMM} & B_{PENS,IMM} & B_{IMM,IMM} & B_{SELF,IMM} \\ B_{EMP,SELF} & B_{PENS,SELF} & B_{IMM,SELF} & B_{SELF,SELF} \end{bmatrix} \begin{bmatrix} \bar{e}_{EMP}^R \\ \bar{e}_{PENS}^R \\ \bar{e}_{IMM}^R \\ \bar{e}_{SELF}^R \end{bmatrix} = \begin{bmatrix} U_{EMP}^R \\ U_{PENS}^R \\ U_{IMM}^R \\ U_{SELF}^R \end{bmatrix} \qquad [B.4]$$

so $\bar{e}^R = \left(\bar{e}_{EMP}^R, \bar{e}_{PENS}^R, \bar{e}_{IMM}^R, \bar{e}_{SELF}^R\right) = B^{-1}U$.

The amount of unreported income by source type is:

$$U_i^R = \bar{e}_i^R Y_i^R \qquad [B.5]$$

and Y_i^R is the total amount of type *i*'s received income. The amount of unreported income by geographical area is instead:

$$U_j^A = \bar{e}_j^A Y_j^A \qquad [B.6]$$

and Y_j^A is the total amount of area *j*'s received income.

From the BETAMOD simulated true gross incomes we compute the total amount of individual incomes by main income source type ($i = 1,...,4$) and by geographical area, ($j = 1,...,4$), obtaining the 4×4 matrix $Y = \{y_{ij}\}$.

By using matrix *Y* and the marginal distribution of unreported income by source type, U^R, and by geographical area, U^A, with the use of the RAS technique we first obtain the joint distributions of total unreported income by income source and by area, $U = \{u_{ij}\}$, and, secondly, the 4×4 matrix of average tax evasion rates, $\bar{e} = \{\bar{e}_{ij}\}$, by source type of received income and by geographical area:

$$\bar{e}_{ij} = \frac{u_{ij}}{y_{ij}} \qquad [B.7]$$

The matrix $\bar{e}$ is shown in Table 12.

B.2. Tax evasion profiles by classes of true gross income

Each average tax evasion rate $\bar{e}_{ij}$ is then modulated in order to obtain a profile of tax evasion by classes of true gross income, i.e. a vector of tax evasion rates associated with 13 income classes (see Table 9). Define $\bar{e}_{ijk}$ as the average tax evasion rate for income class *k*, source type *i* and area *j* with the following function:

$$\bar{e}_{ijk} = \frac{k_i^e \bar{e}_{ij}}{1 + (k_i^e - 1)\left(\frac{y_{ijk}}{k_i^y \bar{y}_{ij}}\right)^{z_i}} \qquad [B.8]$$

where:

y_{ijk} = mean gross true of class *k*, source type *i* and area *j*;

$\bar{y}_{ij}$ = mean gross true of source type *i* and area *j*;

$\bar{e}_{ij}$ = average tax evasion rate for source type *i* and area *j*;

and the parameters to estimate are:

k_i^e determines the ordinate intercept;

k_i^y determines the level of income for which $\bar{e}_{ijk} = \bar{e}_{ij}$;

z_i determines the curvature of the function.

With this formulation we need to estimate 12 parameters: k_i^e, k_i^y, z_i with $i = 1,...,4$. As we assume that pensions cannot be concealed, the number of parameters reduces to 9. The method used by BETAMOD is a procedure of numeric optimization that assigns randomly the value of the 9 parameters and choose the best combination that minimize the distance function

$$D = \sum_k \left| Y_k^{MEF} - Y_k^{BETAMOD} \right| \quad [B.9]$$

where Y_k^{MEF} and $Y_k^{BETAMOD}$ are respectively the official returns and the BETAMOD total amount of reported income by classes. The profiles obtained are shown in Figure 8.

B.3. Assignment of individual tax evasion rates

The average tax evasion rate $\bar{e}_{ijk}$ (for the *i*-th income source type, the *j*-th geographical area and the *k*-th class of true gross income) is defined as the ratio between the unreported income U_{ijk} and the true income Y_{ijk} of the cell:

$$\bar{e}_{ijk} = \frac{U_{ijk}}{Y_{ijk}} \quad [B.10]$$

And can be seen as the product:

$$\bar{e}_{ijk} = \frac{U_{ijk}}{Y_{ijk}} = \frac{U_{ijk}}{Y_{Eijk}} \frac{Y_{Eijk}}{Y_{ijk}} = \bar{e}_{Eijk} H_{Yijk} \quad [B.11]$$

where:

Y_{Eijk} is the total amount of tax evaders' income in cell *i,j,k*

$\bar{e}_{Eijk} = \frac{U_{ijk}}{Y_{Eijk}}$ is the average tax evasion of tax evaders in cell *i,j,k*

$H_{Yijk} = \frac{Y_{Eijk}}{Y_{ijk}}$ is the share of tax evaders income in cell *i,j,k*

The values of $\overline{e}_{E\,ijk}$ and $H_{Y\,ijk}$ are unknown, but we know their product $\overline{e}_{ijk}$ and their maximum value (i.e. 100%). In the absence of further information, we assume that the two values are equal, so:

$$H_{Y\,ijk} = \overline{e}_{E\,ijk} = \sqrt{\overline{e}_{E\,ijk}} \qquad [B.12]$$

For instance, if the average tax evasion rate in a cell is 0.25, then we assume that $H_{Y\,ijk} = \overline{e}_{E\,ijk} = \sqrt{0.25} = 50\%$, i.e. tax evaders own the 50% of the true gross income in the cell and that their tax evasion rate is 50%.

To assign individual tax evasion in BETAMOD we proceed in the following way:

1. for each value $\overline{e}_{ijk}$ we compute the two values $\overline{e}_{E\,ijk}$ and $H_{Y\,ijk}$;
2. we randomly assign a probability to be a tax evader to each taxpayer in the sample (by means of a uniform distribution) and we order taxpayers in decreasing order of probability;
3. starting with the taxpayer with the highest probability, we assign a random tax evasion rate drawn from a *beta* distribution with mean $\overline{e}_{E\,ijk}$ and a standard error varying with the mean;
4. we proceed to assign tax evasion rates to taxpayer with lesser probability until we reach the total amount of unreported income of the cell U_{ijk}.

The *beta* distribution used to assign a tax evasion rate e_{tijk} to the income of source type *i* of the taxpayer *t* with characteristics *j,k* is then $e_{tijk} \sim beta\left(\frac{\theta \overline{e}_{E\,ijk}}{1-\overline{e}_{E\,ijk}}, \theta\right)$. This *beta* distribution has expected value equal to $E(e_{tijk}) = \overline{e}_{E\,ijk}$ and standard deviation equal to $sd(e_{tijk}) = \sqrt{\frac{\overline{e}_{E\,ijk}(1-\overline{e}_{E\,ijk})^2}{k+1-\overline{e}_{E\,ijk}}}$.

The standard deviation is close to zero when there is no tax evasion ($\overline{e}_{E\,ijk} = 0$) or when all income is concealed ($\overline{e}_{E\,ijk} = 1$), and is negatively correlated with the parameter θ. We assigned to θ a value of 5 in order to obtain a maximum value of the standard deviation approximatively equal to 1/6 when the average tax evasion rate is about 1/3.

C. INCIDENCE OF TAX RELIEFS ON REPORTED INCOME

Figure C.1 Total tax deductions as a proportion of reported income (%)

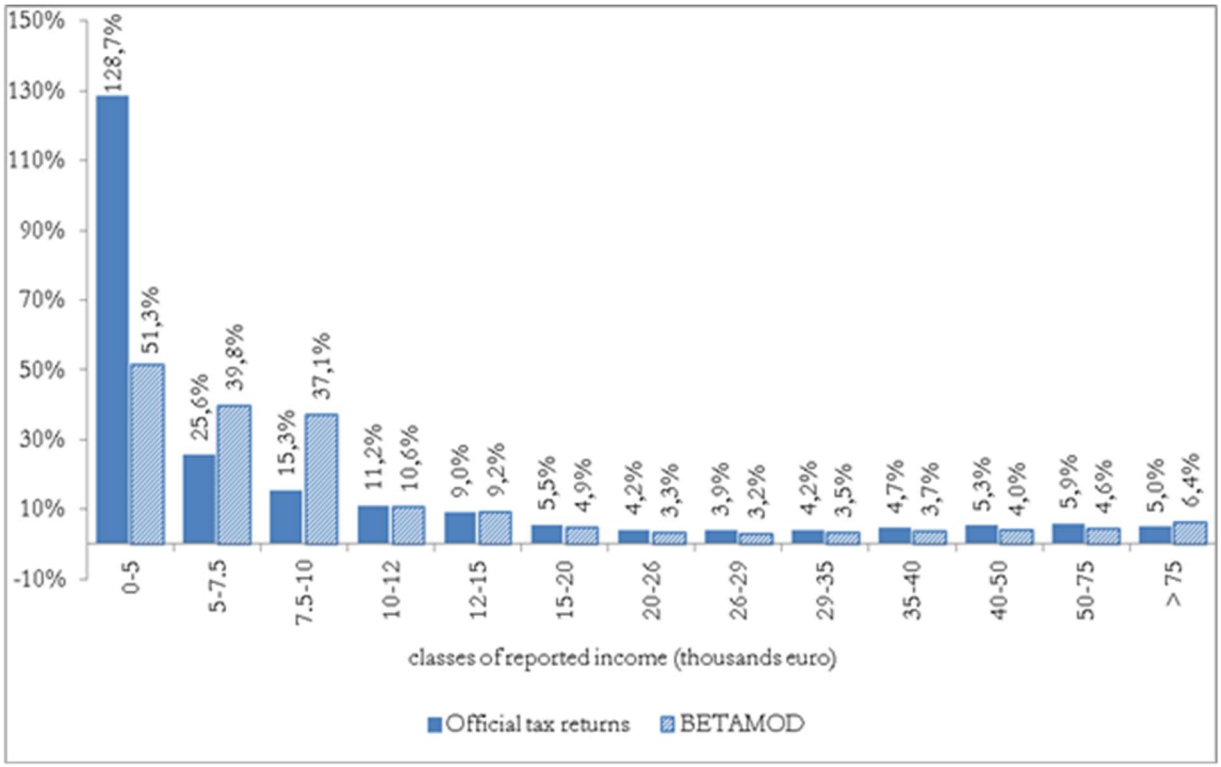

Figure C.2 The 19% tax credits as a proportion of reported income (%)

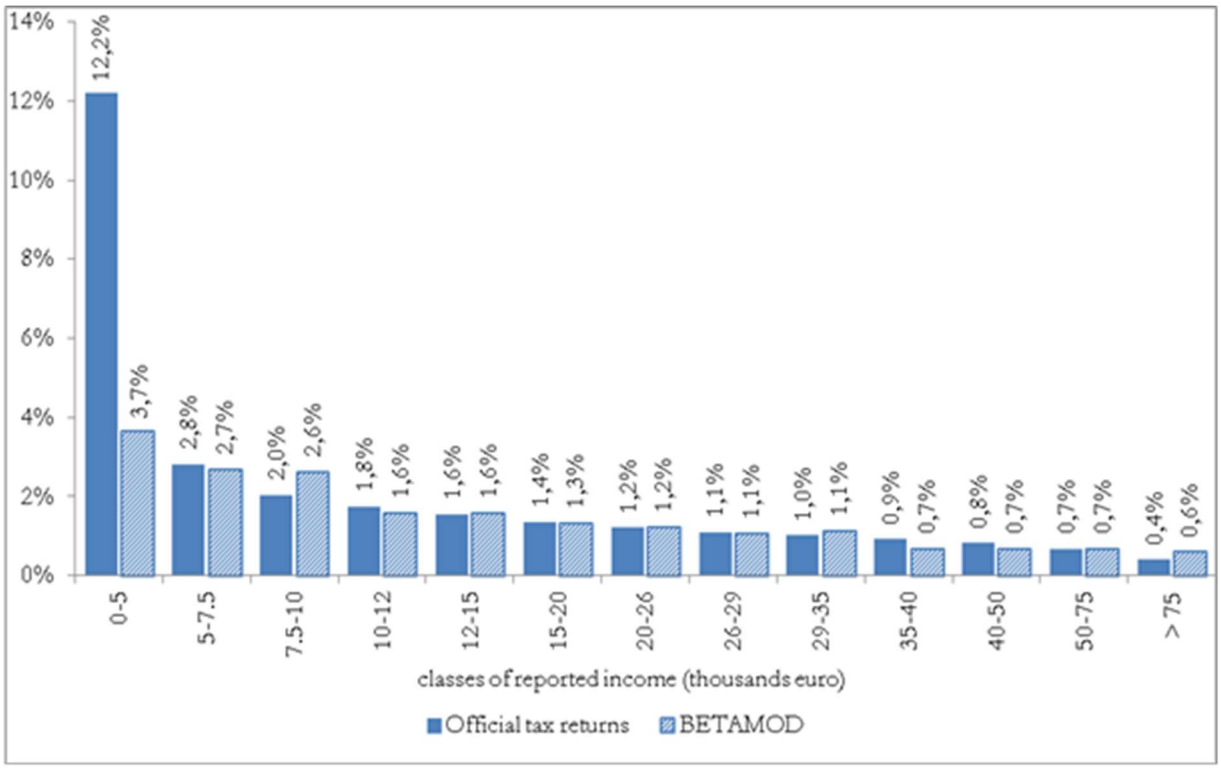

Figure C.3 Social insurance contributions paid by self-employed individuals as a proportion of reported income (%)

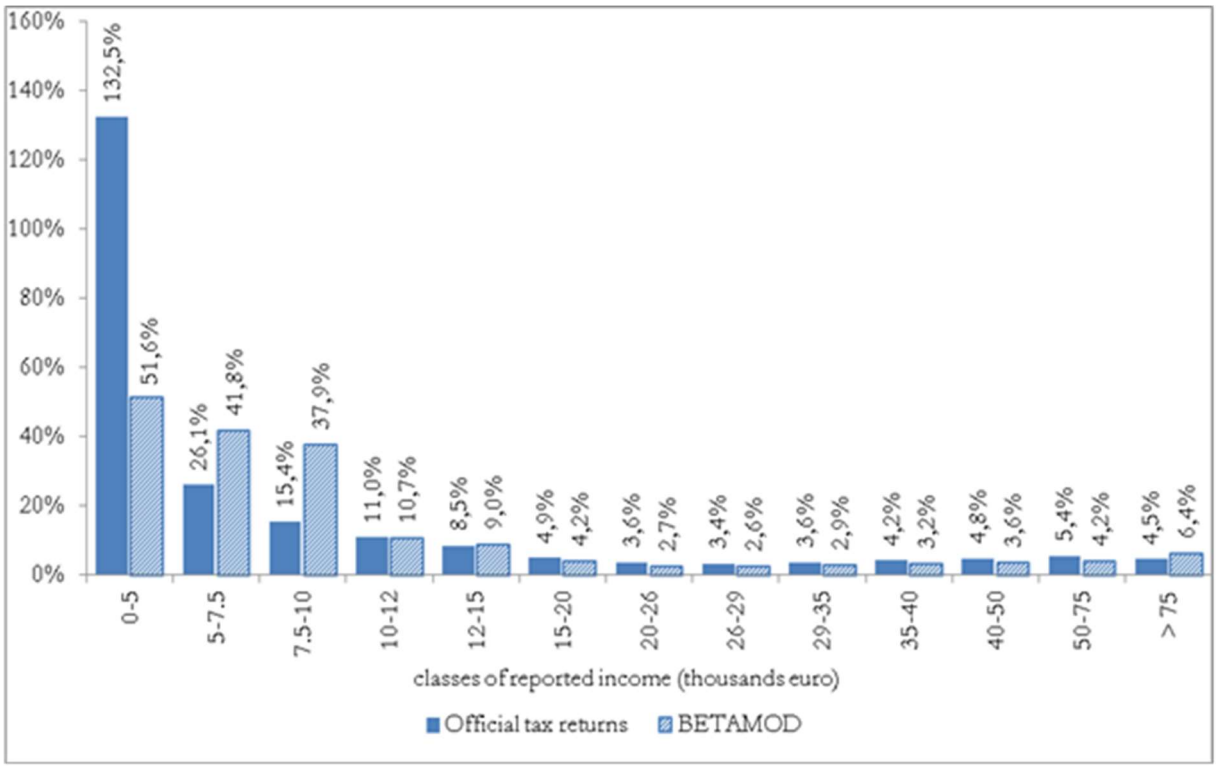

Figure C.4 The 19% tax credit for healthcare expenses as a proportion of reported income (%)