DEMB Working Paper Series

N. 49

Imputation of missing expenditure information in standard household income surveys

Massimo Baldini*, Daniele Pacifico**
and Federica Termini***

January 2015

* University of Modena and Reggio Emilia
CAPP (Centre for the Analysis of Public Policies)
Address: Viale Berengario 51, 41121 Modena, Italy,
email: massimo.baldini@unimore.it
**Organization for Economic Cooperation and Development and CAPP
email: daniele.pacifico@oecd.org
*** University of Bologna
email: federica.termini@studio.unibo.it

ISSN: 2281-440X online

Imputation of missing expenditure information in standard household income surveys

Massimo Baldini[†], Daniele Pacifico[ǂ], Federica Termini[*]

January 2015

Abstract

The aim of this paper is to present a new methodology for dealing with missing expenditure information in standard income surveys. Under given conditions, typical imputation procedures, such as statistical matching or regression-based models, can replicate well in the income survey both the unconditional density of household expenditure and its joint density with a set of socio-demographic variables that the two surveys have in common. However, standard imputation procedures may fail in capturing the overall relation between income and expenditure, especially if the common control variables used for the imputation have a weak correlation with the missing information. The paper suggests a two-step imputation procedure that allows reproducing the joint relation between income and expenditure observed from external sources, while maintaining the advantages of traditional imputation methods. The proposed methodology suits well for any empirical analysis that needs to relate income and consumption, such as the estimation of Engel curves or the evaluation of consumption taxes through micro-simulation models. An empirical application shows the makings of such a technique for the evaluation of the distributive effects of consumption taxes and proves that common imputation methods may produce significantly biased results in terms of policy recommendations when the control variables used for the imputation procedure are weakly correlated with the missing variable.

Introduction

Lack of household expenditure information in standard income surveys has been always an issue in several social studies. Typical examples are the analysis related to living standards and to the policy initiatives to improve them. Indeed, the evidence indicates

[†] University of Modena and Reggio Emilia (Modena, Italy) and Centre for the Analysis of Public Policies (Modena, Italy); email: massimo.baldini@unimore.it.

[ǂ] Organization for Economic Cooperation and Development (Paris, France) and Centre for the Analysis of Public Policies (Modena, Italy); email: daniele.pacifico@oecd.org

[*] University of Bologna (Bologna, Italy); email: federica.termini@studio.unibo.it

that while income can be a good proxy for living standards, it is better when supplemented with a wider range of measures, especially expenditure.[1] However, typical income surveys do not collect information on household consumption while common household-budget surveys do not have information on household income.[2]

To overcome this inconvenience it is possible to impute the missing expenditures to the income dataset through several methods.[3] A common procedure is to estimate Engel curves for each category of household expenditure (e.g. food, clothing, health, etc.) with parametric or non-parametric functions that relate the budget shares to the common variables of both surveys, and then to apply the estimated function to the income dataset (Decoster et al., 2011; Sekhon, 2011). An alternative method is to rely on distance functions rather than on statistical models. The idea is to base the imputation on the similarities between households' characteristics of both datasets, through the use of a distance function to be minimized (Zhao, 2004).

Unfortunately, there is no clear evidence in the literature about which method performs better than the other, even though a recent literature has show that imputation based on distance functions tends to provide more reliable estimates.[4] The main reason is that consumption imputation through Engel curves suffers from a number of weaknesses. Firstly, the imputed variance of the expenditure vector is typically lower than the observed one due to an imputation technique that heavily relies on regression methods; secondly, sample selection in consumption items needs to be accounted for when imputing consumption vectors and, thirdly, estimating Engel curves for several commodities can be computationally demanding – especially when imputing goods not consumed by a significant share of the population – so the need to create large consumption aggregates that may generate loss of information. On the other hand, methods based on distance functions allow reproducing not only the mean but also the variance of the consumption vector and do not require the aggregation of consumption items into macro categories. However, their ability to adequately reproduce the consumption information crucially depends on the number and similarity of the common socio-demographic characteristics between the two surveys. Nevertheless, only a few common variables typically exist in the two types of surveys and their relationship with consumption behaviour may not be that strong.

The aim of this paper is to propose a mixture of the two approaches described above – i.e. a combination of regression and distance-function methods – that relies on information from external sources that have aggregate information on both income and

[1] This is consistent with the recommendations of the Report on the Measurement of Economic Performance and Social Progress (Stiglitz, Sen, & Fitoussi, 2009).
[2] In some European countries household budget surveys may contain some broad information on household income. However, this is normally not enough for the empirical analysis, especially when it requires detailed information on different income sources.
[3] An in-depth overview of these methods is done in Decoster et al. (2007), who also assess the goodness of these techniques to integrate household income and budget surveys.
[4] See Decoster et al. (2007) for a review.

consumption. Indeed, even if highly detailed income (household budget) surveys do not typically have comprehensive information on consumption (income), it is common to find other less-detailed surveys where *broad* information about consumption and income are jointly collected. As an example, in most European countries the Survey on Income and Living Conditions (EU-SILC) collects highly-detailed information on income but not on consumption; at the same time, homogenous household-budget surveys exist in almost all European countries and have highly-detailed information on household expenditures but most of them have no information about income. However, the Eurosystem's Household Finance and Consumption Survey (HFCS), coordinated by the European Central Bank, collects for most of the Euro-Area countries household-level data on household income and consumption. Typically, surveys like the HFCS cannot be used directly for the purpose of the analysis because of the small sample size and of the generic manner in which information on income and consumption are collected, which typically results in only two vectors of total household expenditure and income. Whereas such information is typically not enough for the purpose of an in-depth analysis, it can be extremely useful for improving the quality of imputation through the standard imputation methods described above. The idea is the following: use surveys like the HFCS to estimate a regression model of household expenditure on disposable income plus a series of covariates that the survey shares with the income survey. Then, use the estimated parameters to reproduce a vector of consumption expenditures in the income dataset and create percentiles of imputed household consumption. Now, depending on the model's ability to rank households in terms of their overall consumption – which is typically very high, given the presence of household income among the predictors – households from the same percentiles will have similar consumption-to-income elasticities. Thus, an imputation of consumption items through standard distance-function methods by percentiles of household expenditure would significantly improve the imputation quality, due to the high predictive power that consumption-to-income elasticity has in explaining total consumption behaviour. Moreover, the use of a third survey is limited to the creation of consumption percentiles in the income dataset; this, on the one hand, reduces the bias associated to the use of a third survey and, on the other hand, minimizes the error in sorting households from the income survey, due to the discretization of the predicted consumption vector into percentiles. To synthetize, our method is composed of the following stages:

1) In the survey with broad information on both income and expenditure, regress total household expenditure on disposable income and a set of socio-demographic variables that the Household Budget Survey (HBS) shares with the income survey.
2) Obtain a predicted value of total expenditure on the basis of point 1) regression in the income survey, and sort households by percentiles of imputed total expenditure.
3) After sorting also households of the HBS by (the original) overall expenditure, perform a distance-function matching of the vector of consumption items from

each percentile of the HBS to the corresponding percentile of the income dataset.

As will be shown, under given conditions, typical imputation methods based on distance-function matching allow replicating well the frequency distribution of the donor survey, even for sub-categories of household expenditures. However, without the imputation by percentiles described above the imputed vector of household expenditures may have a relatively low correlation with the observed vector of household income, especially when it is compared with the correlation observed in surveys where both vectors are jointly collected. This result crucially depends on the (often) low correlations between the consumption (and income) vector(s) and the covariates that the two surveys have in common, which can significantly affect the derived relationship between the imputed consumption and the observed income vectors, causing therefore biased results in the subsequent analysis.

The rest of the paper proceeds as follows: section 2 shows the empirical implementation and the results of standard matching methods between two standard income and budget surveys, section 3 shows the performance of the proposed technique and section 4 an application on the distributive effects of the value added tax through micro-simulation analysis.

Section 2: Standard imputation methods in practice

In this section we show the results of a typical imputation exercise of the vector of household expenditures from a standard household budget survey to a household income survey. The datasets used for the empirical application are the European Union Survey on Income and Living Conditions (SILC), referred to Italy, as recipient dataset and the Italian Household Budget Survey (HBS) as donor. SILC provides extremely detailed information on income variables, since it aims at collecting timely and comparable micro-data on income poverty, social exclusion and living conditions for all European countries. The Italian sample for the year 2012 contains data on 19,579 households and has limited data on expenditures. This lack of information will be compensated by using the 2012 HBS, which contains detailed data on expenses and consumption habits of 22,933 Italian households, besides other socio-demographic variables, such as professional conditions, education, age, etc.

The preferred imputation method for our application is via standard distance-function algorithms because, as discussed in the introduction, they are more common in the empirical literature than regression-based methods.[5] Any imputation method relies on pre-requisites of harmonization and coherence of data from both datasets, thus the need of a reconciliation of the socio-economic variables used to implement the procedure.[6] The next table shows the common variables between the two surveys.

[5] For recent applications see: Eurostat Working Paper, 2013, Conti et al., 2014, Pisano et al., 2014.

[6] See Ballin et al. (2009) and Donatiello et al. (2014) for a similar exercise.

Table 1: Common socio-demographic characteristics between the two surveys

Demographic variables referred to the household reference person	Sex, Age, Education, Region, Marital Status, Professional Status, Classification of economic activity
Household structure	Number of members, number of dependent children, number of adults, number of elderly people, number of women, number of workers
House	Tenure status, Rent, Bathroom
Durable goods owned	Washing machine, car, telephone, TV, computer

Thus, the first step is to analyse the similarities of the distributions of the set of common variables between the two datasets. However, the comparison requires homogenizing such variables, recoding or renaming them, or creating new ones from a combination of those present in the data. The tables in the annex show the comparisons among the frequency distributions of some of these recoded variables. It can be easily noticed that the variables have large similarities in their frequency distributions; differences clearly exist but, still, they are in the class of 5 percentage points, at the most. Once a set of common variables has been identified and harmonised, the next step is to select those to include in the matching algorithm; here, we follow the common procedure to include only the variables with a similarly-enough structure between the two surveys, according to standard statistical tests such as the chi-squared or the Kolmogorov-Smirnov.[7]

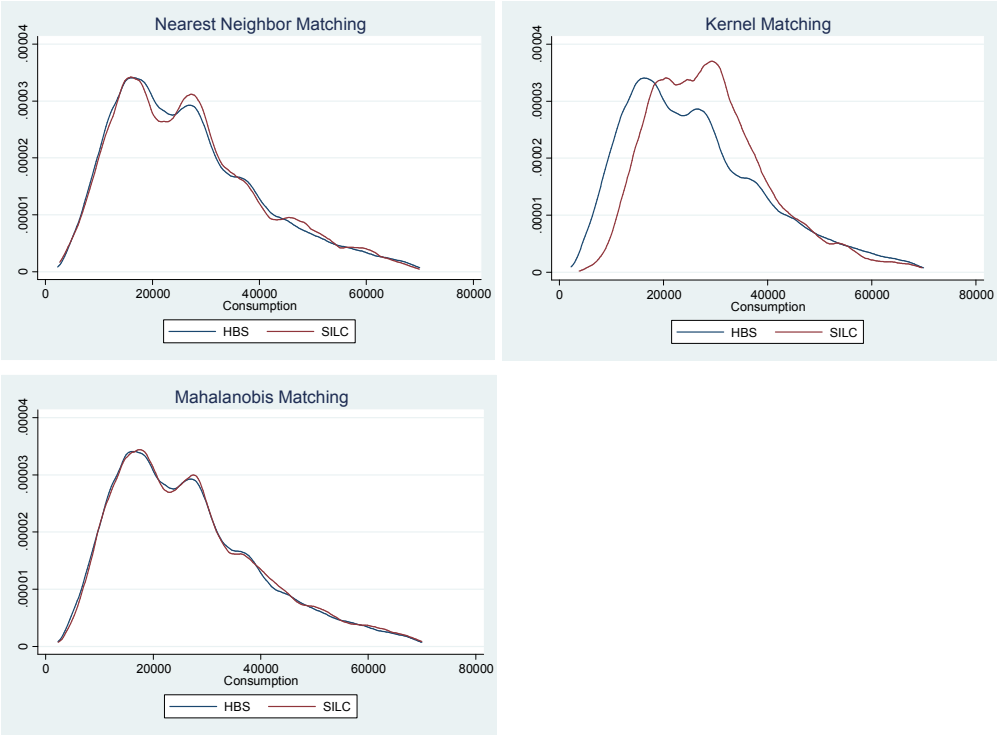
The final step is to choose the matching algorithm to pair households from the two surveys. Here several options are available: nearest neighbour matching (NNM), kernel matching (KM), radius matching, stratified matching and Mahalanobis metric matching (MM). For our purposes the most appropriate methods are the NNM with replacement, KM and the MM because the remaining algorithms may produce loss of observations in the recipient dataset depending on the distance between donor and recipient households (Becker et al., 2002). Since there is no rule to determine which algorithm is the most appropriate, the three methods have been all implemented.

As other empirical analyses have shown (see Zhao, 2004 and Diamond et al., 2005), Mahalanobis matching outperforms the other two methods. This can be seen from the comparison of the densities of the imputed and observed expenditure vectors herein below and also from Table A-2 in the annex, which clearly shows that Mahalanobis matching minimizes the percentage deviation of the average consumption between the donor and the recipient surveys by a large set of socio-demographic categories.[8]

[7] See also Caliendo et al. (2008) for this point. A vector containing most of the variables of Table 1 was eventually selected, with the only exception of those variables related with durable goods and the classification of economic activity of the household head. Results about the chi-squared and the Kolmogorov-Smirnov tests can be provided upon request.

[8] Also a series of parametric and nonparametric tests such as the chi-squared and the Kolmogorov-Smirnov tests have been conducted to check the performance of each matching algorithm in terms of both the variability of the target variable and its joint distributions with the set of control variables. Results can be provided upon request.

Figure 1: Estimated kernel densities for three standard matching algorithms



Note: our computation based on SILC 2012 after imputation of consumption items from HBS 2012.

Kernel matching clearly produces the worst results. The reason depends on this type of matching algorithm, which imputes to each household of the income survey a weighted average of the whole consumption vector from the donor survey, with weights given by the following equation:

$$W_{ij} = \frac{K(\frac{p_i - p_j}{h})}{\sum_{j \in C} K(\frac{p_i - p_j}{h})}$$

Where $K(,)$ is a typical Kernel function, p_i represents the propensity score of the recipient household i, p_j represents the propensity score of the donor household j and h is the bandwidth parameter, used to control the degree of smoothing. Thus, given that each household in the income survey receives an imputed consumption based on a weighted average of the whole consumption vector from the donor survey, differences between the imputed vector and the original one become significant when either the variance of the observed consumption vector is high or when the variance of the propensity scores is low (Abadie et al., 2006; Imbens, 2004).

Thus, it can be argued that the imputation via Mahalanobis matching is successful in reproducing in the income survey both the conditional and the unconditional density of household expenditure; this is why such a methodology is so common in empirical studies. However, so far nothing has been said about the estimated relationship between income and expenditure. The next table shows the results of a simple regression of the logarithm of imputed consumption on the logarithm of disposable

income. The main result of this regression is probably the value of the R-squared, which indicates that only 6 per cent of the consumption variability is explained by the variation of disposable income, an extremely low value if compared to common economic beliefs.

Table 2: Regressing total imputed consumption (log) on disposable income (log)

Log Tot. Expenditure	Coeff.	Std. Err.	t	95% Conf. Interval	
Log Disp. Income	-0.34	0.03	-12.01	-0.40	-0.29
Log Disp. Income [square]	0.03	0.00	22.07	0.03	0.04
Constant	9.98	0.14	72.42	9.71	10.25
R-squared	**0.06**				

Note: our computation based on SILC 2012 after imputation of consumption items from HBS 2012

These findings can be explained by the relatively poor statistical relationship between household consumption and the socio-demographic variables that have been used in the matching algorithm. Thus, as long as the assessment of the matching quality is made with respect to the variables that enter the matching procedure, we do not find significant differences in the average consumption by the categories of these variables (provided their distribution is similar in the two datasets). However, as soon as we evaluate the performance of the imputation with respect to other dimensions – and especially household income – differences start to appear. In other words, the socio-demographic characteristics used to match the two surveys are able to capture only part of the actual living standards (measured either in terms of consumption or income); thus common matching procedures that aim at imputing a consumption vector in income surveys may not be appropriate as such because they would produce biased results.

Section 3: an alternative methodology

A possible way out is to make use of a third survey that collects joint information about income and expenditures. In Italy, the Survey on Household Income and Wealth (SHIW) – which actually provides data for Italy for the Eurosystem's HFCS described in the introduction – can be of use, as it collects detailed information on disposable household income and (only broad) information about household expenditures. Such survey cannot be used directly for the analysis of interest because the sample size is significantly lower than the other two surveys (8,151 households in 2012), and especially because consumption is only approximately collected (with a recall question on total and durable yearly expenditure), which may not be enough for the subsequent empirical analysis.[9]

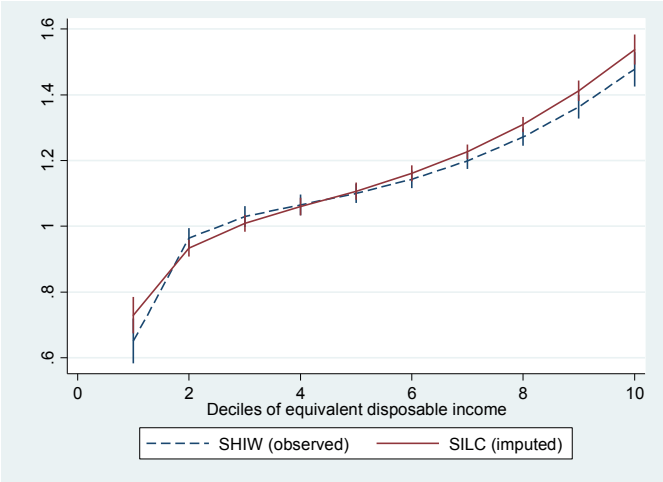
However, SHIW can be useful for estimating a consumption model with a strong predictive power when income is included among the covariates.[10] Thus, the estimated

[9] For instance, the distributive analysis of consumption taxes typically requires the observation of each consumption item in the income survey because of the different tax rates that are applied to different good and services.

[10] The regression coefficients and their standard errors are reported in the appendix (Table A-3). It is worth noting that SHIW variables needed to be made homogeneous with respect to SILC variables. In this case the differences in frequency distributions are slightly larger, as table A-1 shows. These greater dissimilarities largely depend on differences in the sample scheme and in the sample size.

coefficients can be used to impute a value of total consumption in the SILC survey with a high degree of precision.[11] Figure 2 shows the observed income-consumption ratio in SHIW by deciles of household disposable income (red line) and compares it with the predicted one in the income survey (blue line).

Figure 2: Income-consumption ratio by deciles of disposable income



Note: our computation based on SILC and SHIW data. Vertical bars represent 95-percent confidence intervals.

As the figure shows, the regression model is able to reproduce well the average income to consumption ratio by deciles of disposable income in SILC. Moreover, the ratios above are never statistically different between the two surveys according to standard 95 per cent confidence intervals. In our procedure the predicted consumption vector in the income survey is used only to create consumption percentiles, so as to minimize the bias from the use of a third survey and the errors in sorting SILC households. Now, given the model's ability to reproduce the overall relationship between income and consumption, households in the n^{th} percentile will have a similar consumption-to-income elasticity in both SILC and SHIW. Thus, a propensity score imputation of expenditures between SILC and HBS by consumption percentiles (i.e. separate for each percentile) would significantly improve the imputation quality because the procedure now accounts for a very strong predictor of household income, i.e. the class of consumption-to-income elasticity the household belongs to. Furthermore it is demonstrated that our imputation fulfils the main levels of validation in survey-data integration stated by Rassler in Rassler (2004).[12]

The next table shows again the results of a simple regression of (log) consumption on household (log) disposable income in both SILC and SHIW. Results are clearly different

[11] Baldini, Giarda and Olivieri (2015) apply this procedure to evaluate the effects of the recent increase in the ordinary Vat rate in Italy.

[12] In her paper, Rassler refers to four level of validity that a procedure may achieve to be considered effective, namely *Preserving Individual Values* (unfeasible most of the times), *Preserving Joint Distributions, Preserving Correlation Structures, Preserving Marginal Distributions.*

from those reported in the previous section. Now the R-squared is at 60 per cent, and also the slope parameters are not statistically different between the two surveys, meaning that the imputation method reproduces well the relationship between the two vectors that is observed in SHIW. Moreover, as Table A-2 in the annex shows, the proposed methodology outperforms the other imputation methods described in Section 2 also in terms of minimizing the mean square error between the observed and the imputed average expenditures by a large set of socio-demographic variables.

Table 3: Total imputed consumption on observed disposable income – SILC

Log Tot. Expenditure	Coeff.	Std. Err.	t	95% Conf.	Interval
Log Disposable Income	-0.68	0.02	-41.28	-0.71	-0.65
Log Disposable Income [square]	0.07	0.00	72.86	0.07	0.07
Constant	9.80	0.08	124.62	9.65	9.96
R-squared:	**0.60**				

Note: our computation based on SILC 2012 after imputation of consumption items from HBS 2012 using the proposed methodology.

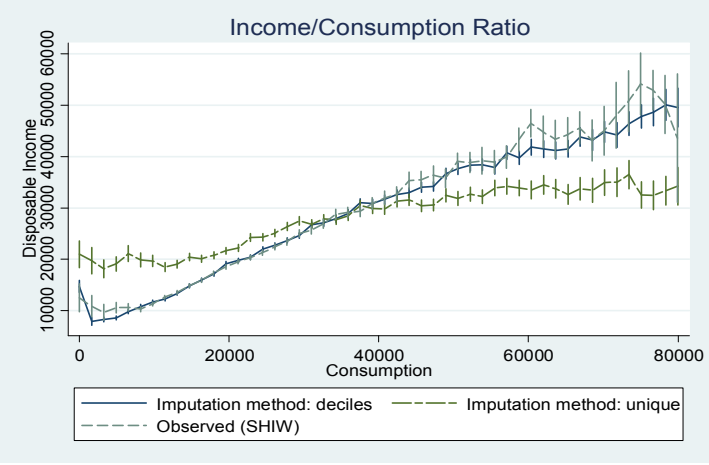
Table 4: Total observed consumption on observed disposable income – SHIW

Log Tot. Expenditure	Coeff.	Std. Err.	t	95% Conf.	Interval
Log Disposable Income	-0.68	0.02	-32.70	-0.72	-0.64
Log Disposable Income [square]	0.07	0.00	57.90	0.06	0.07
Constant	9.95	0.10	102.29	9.76	10.14
R-squared:	**0.60**				

Note: our computation based on SHIW 2012 (observed variables).

The next figure shows the results of a set of local-polynomial regressions of the income-to-consumption ratios as observed in SHIW and as estimated in SILC with standard Mahalanobis matching and with the proposed methodology.

Figure 3: Income-consumption ratio, observed and imputed



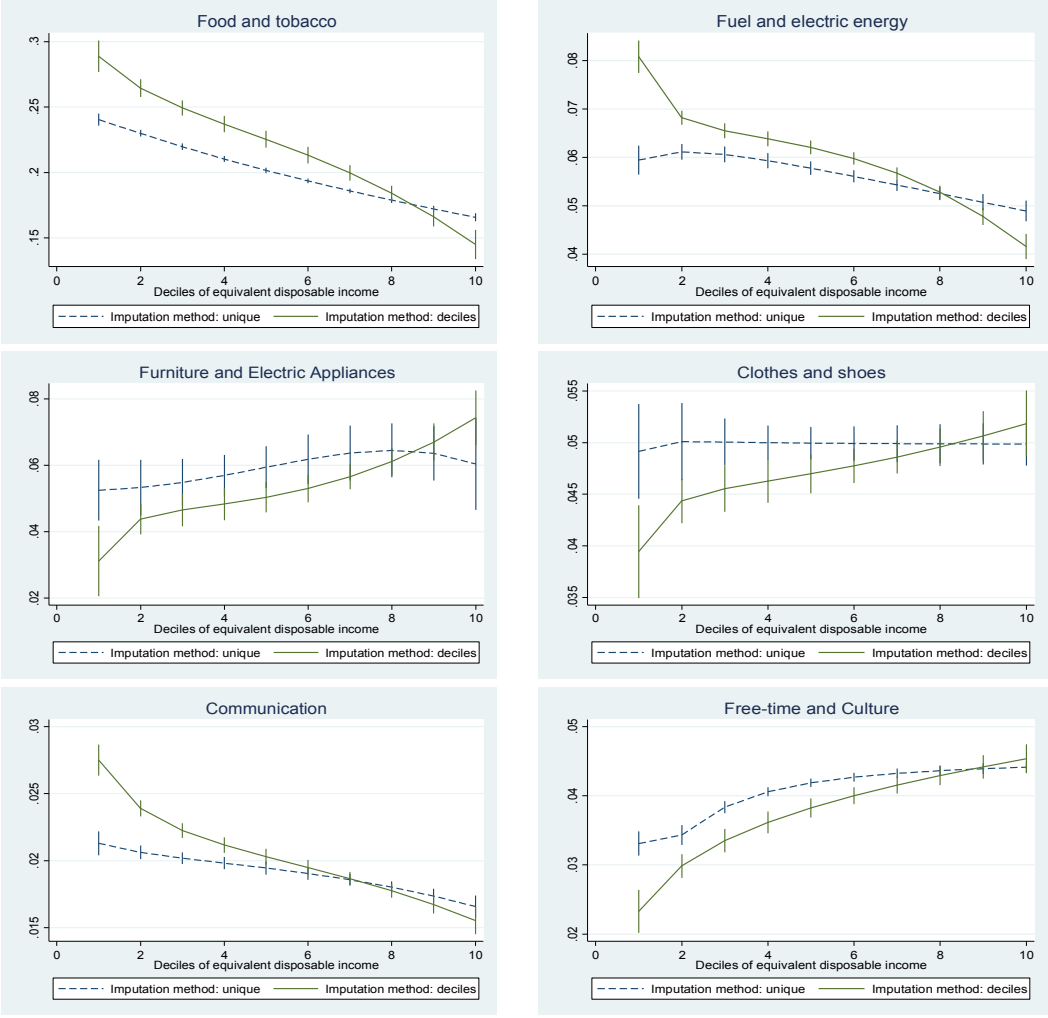
Note: Our computation based on SILC and SHIW data after the imputation of consumption data from HBS. Vertical bars represent 95-percent confidence intervals.

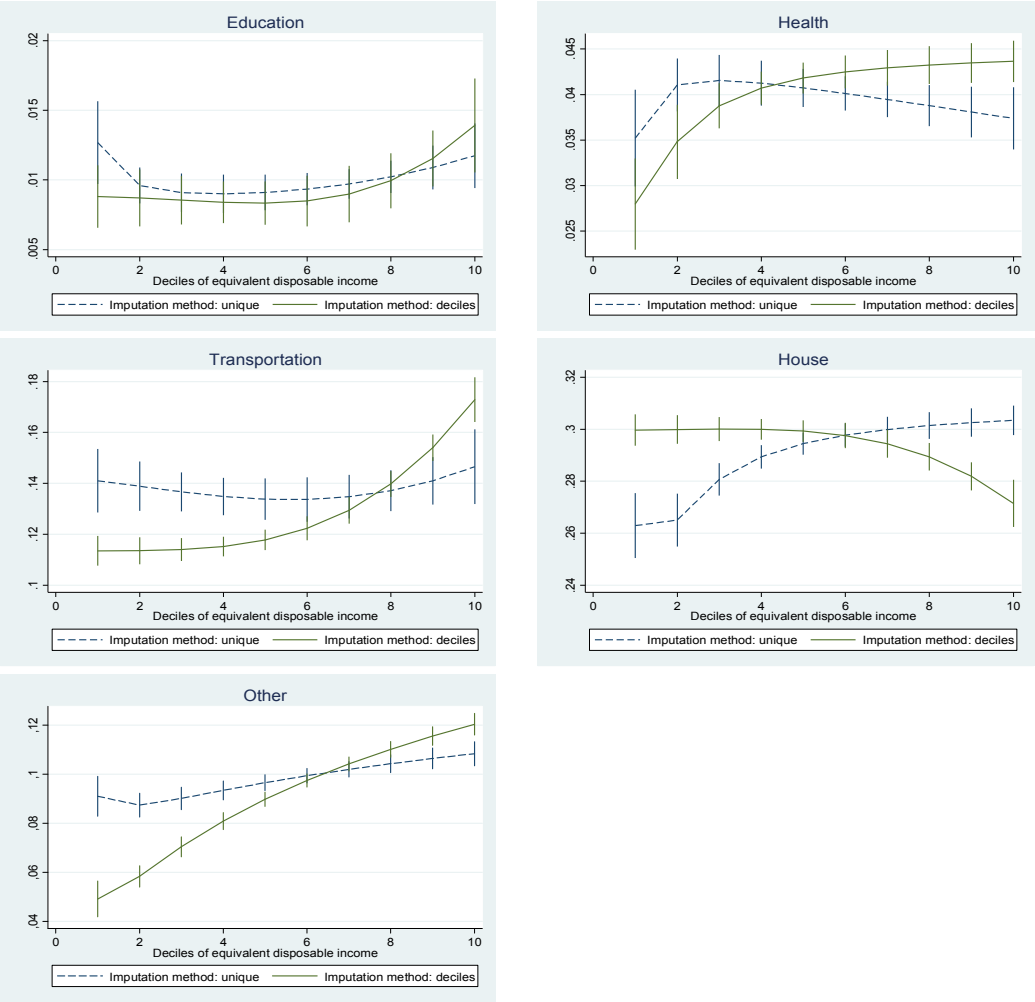
As the figures shows, the gradient of the income-to-consumption ratio simulated in SILC is statistically identical to that observed in SHIW when the imputation is based on the proposed methodology. On the other hand, the gradient of the relationship between

income and consumption obtained with a standard imputation method is statistically different from the others and, importantly, it is relatively flat with respect to income.

A simulated flat relationship between income and expenditure with respect to income may have several consequences in any analysis that requires relating income and consumption. An example can be seen herein below, where Engel curves are estimated for both imputation methods, Mahalanobis matching and Mahalanobis matching by percentiles of household consumption. The dashed blue line is the Engel curve imputed without percentiles, thus not controlling for the relationship between income and expenditure, and the solid green curve is produced by the proposed methodology, which makes use of imputed consumption percentiles.

Figure 4: Estimated Engel curves with different matching procedures





Note: our computation with SILC 2012 after imputation of consumption items by different imputation techniques. Vertical bars represent 95 percent confidence intervals.

As the figures above show, the two imputation methods present significantly different results in reproducing the share of consumption for different subcategories of expenditure by deciles of household income. In fact, Mahalanobis matching produces flatter Engel curves, due to the fact that also the general imputation of consumption exhibits a horizontal path in relation to income; therefore we observe fewer differences in consumption habits for increasing levels of income, which can be counterintuitive with respect to common economic believes.

On the other hand, the proposed method seems to impose a strongest gradient in the relationship between sub-groups of household expenditure and levels of disposable income. More specifically, The Engel curves estimated with the proposed methodology decrease faster than the ones obtained with the classical method, especially for expenditures related to necessary goods such as food, housing and energy or communication. Thus, the results above confirm that, for poorer household, the share of consumption for normal goods is higher than for those in top income level, and the new methodology makes the gap between the bottom and the top decile of income even more pronounced. As for the other items, an increasing path in expenditures related to health, education and culture is observed for both imputation methods, confirming that

households with higher income spend relatively more on luxury goods. However, as for normal goods, the curves obtained with the standard imputation method tend to be significantly flatter.

Section 4: An application for policy evaluation

In this section we describe a possible use of the integrated data for policy evaluation. In particular, the imputed vectors of consumption items are used together with the Italian Treasury fiscal micro-simulation model – ITaxSIM – to analyse the distributive effects of the Italian Value Added Tax.[13] ITaxSIM allows replicating accurately the most important taxes and benefits of the Italian fiscal system and its main source of information is SILC 2012, which – as explained in the previous sections – has little data on household expenditures. To supplement this lack of information, a Mahalanobis statistical matching between SILC and the Italian HBS for the year 2012 has been implemented following both the standard and the proposed methodologies outlined above. The fiscal simulator is then used to compute, for each imputed consumption item, the household's VAT liabilities by taking into account the whole legislation on the Italian VAT, which taxes different good and services at different tax rates.

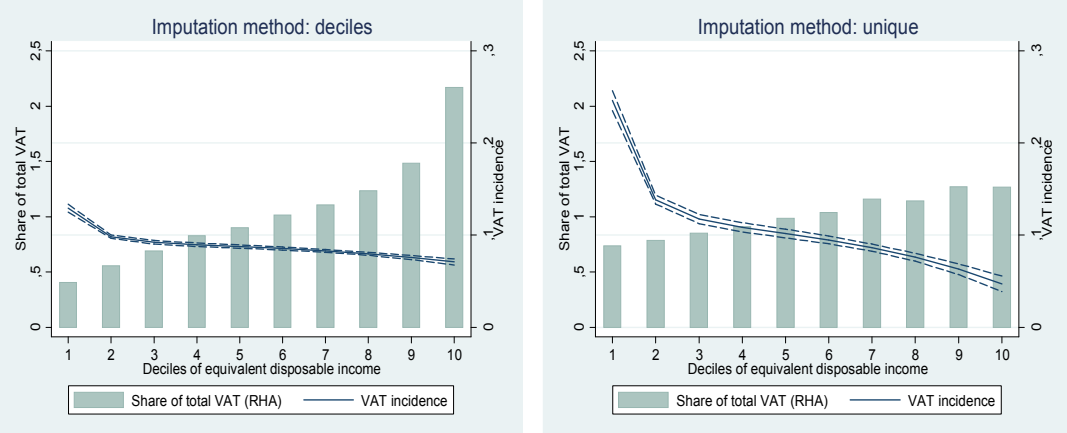
The next figure shows the main results of the analysis, which are the incidence curve of the VAT with respect to the disposable household income and the distribution of the imputed VAT share on the national total VAT by deciles of equivalent disposable income.[14] Importantly, we report the results for both imputation methods, with and without the use of consumption percentiles in the matching procedure, so as to show that results can change significantly depending on the imputation method.

In general, both figures show that the incidence of VAT with respect to household income decreases with the income deciles. Thus, the distributive effect of VAT is regressive with respect to income, as also shown by the positive variation of the Gini index before and after this tax.

[13] See http://www.dt.tesoro.it/en/analisi_programmazione_economico_finanziaria/modellistica/itaxim.html.

[14] The equivalent household income is obtained by dividing the household disposable income by the OECD modified equivalence scale.

Figure 5: Incidence of Vat in the SILC survey after two imputation methods



Gini index variation: + 1.47 percentage points **Gini index variation: +3.56 percentage points**

Note: ITaxSim microsimulation model elaboration based on SILC 2012 after imputation of consumption items by different imputation techniques.

However, as the figure shows, the matching by percentiles of household consumption significantly changes the magnitude of the distributive analysis. Specifically, the variation of the Gini index when consumption deciles are not considered is significantly higher (3.56 p.p. versus 1.47 p.p.). This result depends on the poor predictive power that the matching variables have in explaining the variation of consumption across households. Thus, if no constraints are set to the matching algorithm (so as to better account for the relationship between consumption and income), then the association of households between the two surveys will be scarcely related with the household income. This can be seen in the right-hand graph and especially with the relative uniform distribution of total expenditure across income deciles (bars). Consequently, if expenditure is more uniformly distributed across deciles, so are VAT liabilities and, therefore, the VAT incidence with respect to average household income will be higher in the first deciles.

On the other hand, when the matching procedure takes into account the different income-to-consumption elasticities across deciles, the distribution of household expenditure is significantly shifted towards rich households, causing lower tax liabilities for households in the first deciles of household income (and therefore a smaller increase of the Gini Index).

Conclusions

Economic well being of individuals can be measured using a wide variety of possible indicators. Two of the most frequently used ones are income and consumption, which provide complementary but not identical pictures of the level and distribution of living standards across the population. Despite the importance of data on these domains, rarely a single survey contains high quality information on both income and consumption. This lack of joint information on both income and consumption is often dealt with imputation techniques that, independently from their types, are strongly sensitive to the vector of common control variables that enter the imputation

procedure. Thus, if the relationship between these variables and the target variables is poor – i.e. income and consumption – then also the relationship between the target variables in the integrated dataset will be poor. This paper suggests a strategy to fill in this gap, by imputing to the income survey a complete vector of consumption items for each household, taking also into account the relationship between income and expenditure. The proposed procedure can be of help in building fiscal micro-simulation models that consider also indirect taxation, an increasing source of revenue in advanced countries. Moreover, the proposed procedure enriches the information provided by income surveys like EU-SILC in several directions. Firstly, it is possible to study poverty not only in terms of income but also of consumption; secondly, it is also possible to analyse the differences between these two dimensions of poverty, a theme made more urgent by the increasing diffusion of absolute poverty, usually measured in terms of lack of consumption, also in rich countries.

References

Abadie, Alberto, and Guido W. Imbens. "Large sample properties of matching estimators for average treatment effects." *Econometrica* 74.1 (2006): 235-267.

Baldini, Massimo, Giarda Elena and Arianna Olivieri. "A tax-benefit microsimulation model for Italy: A partial evaluation of fiscal consolidation in the period 2011-14", Prometeia, Nota di Lavoro n. 2015-01, www.prometeia.com (2015).

Ballin, Marco, et al. "Statistical Matching of Two Surveys with a Common Subset." *Tec. Report* 124 (2009): 68-79. Department of Economic and Statistics - University of Trieste.

Becker, Sascha, and Andrea Ichino. "Estimation of average treatment effects based on propensity scores." *The stata journal* 2.4 (2002): 358-377.

Caliendo, Marco, and Sabine Kopeinig. "Some practical guidance for the implementation of propensity score matching." *Journal of economic surveys* 22.1 (2008): 31-72.

Decoster, Andrè, et al. "Comparative analysis of different techniques to impute expenditures into an income dataset." *AIM-AP deliverable* (2007).

Decoster, André, et al. "Microsimulation of indirect taxes." *International journal of microsimulation* 4.2 (2011): 41-56.

Diamond, Alexis, and Jasjeet S. Sekhon. "Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies." *Review of Economics and Statistics* 95.3 (2013): 932-945.

Donatiello, Gabriella et al. "Statistical Matching of Income and Consumption Expenditures." *Report from 9th International Academic Conference, Istanbul* (2014)

Eurostat, "Statistical matching of EU-SILC and the Household Budget Survey to compare poverty estimates using income. expenditures and material deprivation." *Methodologies and Working Papers* (2013)

Guo, Shenyang, and Mark W. Fraser. "Propensity score analysis: Statistical methods and applications." Vol. 12. *Sage Publications*, (2009).

Imbens, Guido W. "Nonparametric estimation of average treatment effects under exogeneity: A review." *Review of Economics and statistics* 86.1 (2004): 4-29.

Pisano, Elena, and Simone Tedeschi. "Micro Data Fusion of Italian Expenditures and Incomes Surveys." *Working Paper No. 164* (2014). Department of Public Economics - University of Rome, La Sapienza,.

Rässler, Susanne. "Data fusion: identification problems, validity, and multiple imputation." *Austrian Journal of Statistics* 33.1-2 (2004): 153-171.

Rubin, Donald B. "Bias reduction using Mahalanobis-metric matching."*Biometrics* (1980): 293-298.

Sekhon, Jasjeet S. "Multivariate and propensity score matching software with automated balance optimization: the matching package for R." *Journal of Statistical Software,* 42.7 (2011): 1-52.

Stiglitz, Joseph, “Report on the Measurement of Economic Performance and Social Progress.” *Commission on the Measurement of Economic Performance and Social Progress, Paris* (2009)

Zhao, Zhong. "Using matching to estimate treatment effects: data requirements, matching metrics, and Monte Carlo evidence." *The review of economics and statistics. MIT Press* 86.1 (2004): 91-107.

ANNEX

Table A.1: Frequency distributions of common variables

	HBS	SILC	SHIW	% Deviation (HBS-SILC)	% Deviation (SILC-SHIW)
Profession					
Manager	5.39	3.77	4.69	1.62	-0.92
White collar	15.86	14.13	14.44	1.73	-0.31
Blue collar	18.8	18.77	21.4	0.03	-2.63
Self-employed	11.56	13.89	9.33	-2.33	4.56
Unemployed	3.97	4.58	3.77	-0.61	0.81
Housewife. Retired. Other	44.42	44.85	46.38	-0.43	-1.53
Age					
0-39	15.17	18.73	15.18	-3.56	3.55
40-49	20.88	20.04	20.28	0.84	-0.24
50-59	19.56	17.83	17.23	1.73	0.6
60-69	16.69	16.41	17.85	0.28	-1.44
70+	27.7	27	29.47	0.70	-2.47
Education					
PhD	1.08	1.96	1.12	-0.88	0.84
Bachelor/Master Degree	10.93	10.21	10.58	0.72	-0.37
High School Diploma	26.21	26.72	26.74	-0.51	-0.02
Junior High School diploma	29.26	28.04	27.12	1.22	0.92
Primary School Certificate	22.12	21.69	21.09	0.43	0.6
Region					
Piemonte/Val d'Aosta	8.19	8.18	12.68	0.01	-4.5
Lombardia	17.19	17.21	12.29	-0.02	4.92
Trentino Alto Adige	1.72	1.73	4.1	-0.01	-2.37
Veneto	8.06	8.06	6.76	0.00	1.3
Friuli-Venezia Giulia	2.21	2.21	3.39	0.00	-1.18
Liguria	3.11	3.11	2.6	0.00	0.51
Emilia-Romagna	7.84	7.84	8.89	0.00	-1.05
Toscana	6.42	6.42	5.26	0.00	1.16
Umbria	1.51	1.51	1.34	0.00	0.17
Marche	2.52	2.52	2.33	0.00	0.19
Lazio	9.39	9.4	9.41	-0.01	-0.01
Abruzzo	2.16	2.16	1.55	0.00	0.61
Molise	0.51	0.51	0.96	0.00	-0.45
Campania	8.36	8.37	8.39	-0.01	-0.02
Puglia	6.1	6.11	5.6	-0.01	0.51
Basilicata	0.91	0.91	3.13	0.00	-2.22

Calabria	3.1	3.1	3.81	0.00	-0.71
Sicilia	7.94	7.92	5.19	0.02	2.73
Sardegna	2.75	2.75	2.33	0.00	0.42
Rent					
No	83.12	81.6	77.23	1.52	4.37
Yes	16.88	18.4	22.77	-1.52	-4.37
Number of children					
0	73.7	73.1	76.13	0.60	-3.03
1	12.78	14.84	12.16	-2.06	2.68
2+	13.52	12.06	11.71	1.46	0.35
Number of hh. members					
1	31.66	31.1	38.54	0.56	-7.44
2	27.37	26.87	26.53	0.50	0.34
3	18.59	20.34	15.71	-1.75	4.63
4	17.17	16.79	13.97	0.38	2.82
5+	4.31	3.77	3.97	0.54	-0.2
Sex					
Female	32.94	33.09	37.65	-0.15	-4.56
Male	67.06	66.91	62.35	0.15	4.56
Married					
No	43.58	44.48	48.84	-0.90	-4.36
Yes	56.42	55.52	51.16	0.90	4.36
Single					
No	83.38	82.16	83.15	1.22	-0.99
Yes	16.62	17.84	16.85	-1.22	0.99

Table A.2: Difference in the distribution of total expenditure by to socio-demographic variables, all matching algorithms

	NNM	KM	MM	By deciles
	% Deviation from HBS	% Deviation from HBS	% Deviation from HBS	% Deviation from HBS
Age				
0-39	-0.12	-0.11	-0.06	0.09
40-49	-0.01	-0.01	-0.01	0.05
50-59	0.02	0.04	-0.02	-0.04
60-69	-0.02	-0.02	0.01	-0.07
70+	-0.04	-0.03	-0.06	-0.06
Gender				
Female	-0.08	-0.07	-0.02	0.04
Male	-0.01	0.00	-0.03	-0.02
Region				
Piemonte/Val d'Aosta	0.03	0.05	-0.04	0.07
Lombardia	0.05	0.06	-0.01	0.02
Trentino Alto Adige	0.04	0.10	-0.05	0.06
Veneto	-0.04	0.06	-0.07	0.09
Friuli-Venezia Giulia	-0.09	-0.08	-0.06	0.00
Liguria	-0.14	-0.10	-0.09	-0.06
Emilia-Romagna	-0.02	0.05	0.04	0.07

Toscana	0.02	0.00	-0.07	0.03
Umbria	-0.08	-0.10	0.00	0.01
Marche	-0.05	-0.05	-0.08	0.00
Lazio	-0.07	-0.05	-0.02	-0.12
Abruzzo	-0.03	-0.04	0.03	0.01
Molise	-0.01	0.02	-0.01	0.01
Campania	-0.06	-0.10	0.03	-0.04
Puglia	-0.15	-0.14	-0.03	-0.06
Basilicata	-0.04	-0.11	0.01	0.08
Calabria	-0.06	-0.14	0.06	-0.07
Sicilia	-0.12	-0.18	-0.06	-0.05
Sardegna	-0.14	-0.14	-0.05	-0.08
Education				
PhD	0.25	0.34	-0.04	0.02
Bachelor/Master	0.09	0.16	0.04	0.04
High School	0.01	0.01	-0.02	0.08
Junior High School	-0.08	-0.08	-0.02	-0.04
Primary School	-0.11	-0.08	-0.09	-0.03
# hh. members				
1	-0.16	-0.16	-0.04	0.01
2	-0.03	-0.02	-0.02	0.00
3	0.01	0.02	-0.02	-0.05
4	0.04	0.08	-0.01	-0.05
5+	0.16	0.20	-0.03	-0.04
Married				
No	-0.10	-0.10	-0.02	0.07
Yes	0.01	0.03	-0.03	-0.05
Single				
No	-0.01	0.00	-0.02	-0.03
Yes	-0.12	-0.14	-0.05	0.15
# children				
0	-0.05	-0.05	-0.03	-0.01
1	0.03	0.03	0.00	0.04
2+	0.01	0.03	-0.05	-0.05
Rent				
No	0.00	0.02	-0.02	-0.01
Yes	-0.20	-0.23	-0.12	-0.01
Profession				
Manager	0.14	0.23	-0.10	-0.01
White collar	0.00	0.01	-0.04	-0.03
Blue collar	-0.08	-0.09	-0.04	0.04
Self-employed	0.02	0.05	-0.01	0.06
Unemployed	-0.23	-0.24	0.05	0.01
Housewife, Retired, Other	-0.06	-0.05	-0.03	-0.05
Total	-0.03	-0.02	-0.03	-0.01

Table A.3: Regression of consumption on income in SHIW

Number of observation	5933
R-squared	0.76

Log Consumption	**Coef.**	**Robust Std. Err**	**t**	**P>t**	**95% Conf. Interval**	
Log Income	-0.53	0.03	20.97	0.00	0.58	-0.48
Log Income (square)	0.05	0.00	33.13	0.00	0.05	0.06
Region						
Lombardia	0.12	0.02	5.43	0.00	0.08	0.16
Trentino Alto Adige	0.12	0.03	4.13	0.00	0.06	0.18
Veneto	0.07	0.02	2.76	0.01	0.02	0.12
Friuli-Venezia Giulia	0.08	0.04	2.25	0.03	0.01	0.15
Liguria	0.09	0.03	3.44	0.00	0.04	0.15
Emilia-Romagna	0.06	0.02	2.88	0.00	0.02	0.10
Toscana	0.08	0.02	3.70	0.00	0.04	0.12
Umbria	0.06	0.03	2.16	0.03	0.01	0.11
Marche	0.07	0.03	2.83	0.01	0.02	0.12
Lazio	0.13	0.02	5.29	0.00	0.08	0.18
Abruzzo	0.04	0.03	1.26	0.21	0.02	0.11
Molise	0.01	0.05	0.32	0.75	0.07	0.10
Campania	0.05	0.03	1.95	0.05	0.00	0.10
Puglia	-0.02	0.03	-0.62	0.54	0.07	0.03
Basilicata	-0.16	0.04	-4.21	0.00	0.23	-0.09
Calabria	-0.03	0.03	-1.14	0.25	0.09	0.02
Sicilia	0.01	0.02	0.52	0.60	0.03	0.05
Sardegna	-0.01	0.03	-0.34	0.74	0.06	0.05
Profession						
Employee	-0.03	0.02	-1.60	0.11	0.06	0.01
Worker	-0.08	0.02	-3.96	0.00	0.12	-0.04
Self-employed	-0.16	0.02	-7.16	0.00	0.20	-0.11
Education						
High School	-0.02	0.03	-0.74	0.46	0.07	0.03
Junior High School	-0.05	0.03	-2.05	0.04	0.11	0.00
Elementary School	0.00	0.03	0.14	0.89	0.05	0.05
Men	0.00	0.01	0.04	0.97	0.03	0.03
Rent	0.00	0.01	-0.02	0.99	0.03	0.03
Single	-0.03	0.02	-1.50	0.13	0.06	0.01
Married	0.04	0.02	2.31	0.02	0.01	0.08
Age						
40-49	-0.01	0.02	-0.25	0.80	0.04	0.03
50-59	0.03	0.02	1.41	0.16	0.01	0.07
60-69	-0.02	0.02	-0.72	0.47	0.06	0.03
70+	-0.02	0.02	-0.86	0.39	0.07	0.03
# children						
1	0.06	0.02	2.77	0.01	0.02	0.10
2+	0.09	0.03	3.17	0.00	0.03	0.14
# hh. members						
2	0.06	0.02	3.64	0.00	0.03	0.09
3	0.08	0.02	3.59	0.00	0.04	0.13
4	0.08	0.03	2.44	0.02	0.02	0.14

5	0.07	0.04	1.96	0.05	0.00	0.14
6+	0.15	0.05	2.90	0.00	0.05	0.26
Const	9.71	0.14	67.07	0.00	9.42	9.99

Note: own computation on SHIW 2012