Optimal sampling design with administrative sources for household finance surveys

Giulio Barcaroli[*], Giuseppe Ilardi[†], Andrea Neri [‡], Tiziana Tuoto [§]

Abstract

Household finance surveys are increasingly used for policy-making. It is therefore essential that they provide an accurate picture of the economic situation of all households. Unfortunately, research has shown that the upper parts of the wealth distribution are often missing in household finance surveys. Since rich households generally concentrate a large share of total income and wealth, survey-based estimators may be biased or affected by low precision. The ideal situation to cope with this issues would be to have access to auxiliary information on household finances at the design stage. In practice, however, this is rarely the case. In this paper we present an empirical application that uses personal tax records in the design of a major survey on household finances. We first discuss the methodological challenges to be dealt with when using administrative information for designing the sample. We then propose a method for an optimal stratification and sample allocation. Finally, we estimate the benefits of the methodology in terms of precision and bias-reduction of the estimators.

Key Words: Household finance surveys, optimal stratification, register data.

1. Introduction

The measurement of households' economic conditions is high on the political and economic research agenda. In recent years, this topic is becoming increasingly important also for National Central Banks, as it has been recognized to interact with their functions (ECB, 2009).

One of their main targets is to guarantee price stability through monetary policy. To this purpose, they need to have a good knowledge of how households make their spending decisions and how they respond to changes in their finances. Central banks also have to supervise the risks for financial stability arising from the household sector. For this reason, they need to monitor the household's ability to face their levels of indebtedness if some shock occurs (such as the loss of a job of some member of the household) (Michelangeli & Rampazzi, 2016). Moreover, Central banks are also increasingly interested in understanding

[*]Italian National Statistical Institute, Directorate for Methodologies
[†]Banca d'Italia, Directorate General for Economics, Statistics and Research.
[‡]Banca d'Italia, Directorate General for Economics, Statistics and Research.
[§]Italian National Statistical Institute, Directorate for Methodologies

the effects of their policies on the household's economic conditions and in particular on income and wealth inequality (Colciago et al., 2019; Dobbs et al., 2013; Casiraghi et al., 2018)).

Sample surveys are the main tool used to collect granular information on these aspects. In the Euro area, the European Central Bank has established a network of survey specialists, statisticians and economists to collect harmonized microdata on household income and wealth through the Household Finance and Consumption Survey (HFCS). Because of the range of purposes for which these data are used, it is particularly important that the survey adequately represent the full distribution of income and wealth. In practice, the greatest difficulties are in obtaining a sufficient number of observations in the two extremes of the distributions. Households with very poor finances may see little relevance in participating in a survey about finances. Moreover, they could live in areas that could be dangerous for the interviewers. Under-representation of these households is likely to have little impact on estimates of mean, but it would affect many other statistics such as those related to the income distribution or poverty.

At the other end of the spectrum, research has shown that very affluent households are likely to be under-represented: see for example, Eckerstorfer et al. (2016); Ranalli & Neri (2011); D'Alessio & Neri (2015); Kennickell (2019); Vermeulen (2018); Chakraborty et al. (2019). Indeed, wealthy respondents are generally a hard-to-reach population since they may live in multiple locations, which, also, may have security measures that make it difficult for the interviewer to contact the household to negotiate the interview. Moreover, rich persons may be difficult to persuade to participate since they are generally busy or less willing to declare their finances. Although such households are small in number, they own a large share of total income or wealth. Thus, the under-representation of these households would have negative effects on many estimates.

The ideal situation to cope with these issues would be to have access to auxiliary information such as administrative records relating household finances at the design stage. Such information would enable survey agencies to identify correctly this rare population making also possible to oversampling it to compensate for the difficulties in enrolling it in the survey. Unfortunately, such auxiliary information is rarely available, mainly because of confidentiality issues that prevent the exchange of personal data among the owner and other institutions. Moreover, even if this information is available, generally it is not consistent with the definitions and the concepts used in the survey.

This study discusses the use of register data on personal income in the sampling design of the Italian HFCS survey. It draws on a collaboration between Banca d'Italia, who runs the survey, and the Italian National Statistical Institute (Istat) who has access to the administrative records. Thanks to this collaboration, we have been able to create two unique archives that are essential for our strategy.

The paper is organized as follows. The following section will provide a brief overview of the different use of administrative records in the main household finance surveys and the main contributions of our article. Sections 3 and 4 will introduce the survey and

register data we use for our application, while Section 5 and 6 describes the methods used in our sample design. The results are presented in Section 7. The article concludes with a summary and discussion of the main results in Section 8.

2. The use of register data in household finance surveys

Administrative records are increasingly used for statistical purposes. Some countries already used them in the design of their household finance surveys.

The US survey of Consumer Finances employs a dual-frame design, including an area-probability (AP) and a list component. The list sample is used to oversample households that are likely to be relatively wealthy. The basis of the sample is a set of specially edited individual income tax returns developed by the Statistics of Income Division (SOI) of the Internal Revenue Service (Kennickell, 2008). The list sample is stratified using a "wealth index" computed using income data to predict a rank ordering of people by wealth. After defining the stratifying variable in terms of the whole population, the list is reduced for the actual selection to include only cases that filed returns from a municipality included in the PSUs underlying the AP sample. Within each stratum, cases are oversampled by a progressively larger proportion in richer strata (Kennickell, 2001).

In Canada, the design of the survey of Financial security foresees that each province is stratified into rural and urban areas and different design is used in each. In rural areas, a multi-stage sample is selected using the Labour Force Survey area frame. In urban areas, information from the administrative records at the family level, such as age and income, is used to stratify the Address Register into groups of dwellings having similar well-being.

In the second wave of the HFCS, fifteen out of twenty countries used different strategies to oversample richer households (Household Finance and Consumption Network, 2016). The strategies varied significantly between countries, and are heavily dependent on the available data.

The Spanish Survey of Household Finances (EFF) has used, at least for some waves, individual wealth tax files. The sampling is achieved thanks to the collaboration of the INE (Spain's statistical institute) and the Tax Authorities (TA), through a complex coordination mechanism (for confidentiality reasons). The population frame contains information on fiscal wealth and income for each household. The choice of defining the wealth strata is based on the households' percentile distribution of the wealth tax for Spain. Cases in richer strata are over-sampled progressively at higher rates (Bover et al., 2014).

The French Wealth survey uses tax registers on personal wealth data to identify four strata: wealthy city dwellers, equity-based wealth, real estate-based wealth, lower wealth. Richer strata are sampled at higher rates.

Tax registers on personal income are used in Estonia, Finland, Latvia, and Luxembourg, while in Cyprus the sampling is based on the Customer register of the electricity authority.

The main limitation to the use of administrative records is the legal restrictions to protect the privacy of households. Depending on the country, the limitations may relate

to the use of the data (for instance, restricting the use to detect tax-evasion purposes) or the transfer the microdata to any institution outside the producing agency.

Other countries adopt different sampling strategies to compensate for the unavailability of register data at the individual level. Greece, Ireland, Hungary, Poland, and Slovenia use the information at area level (such as average income and real estate) as proxies of households' economic conditions).

Despite the use of register data is not a novelty, to the best of our knowledge, there are no many studies in the literature discussing the benefits and the challenges in the use of register data in the design of a household finance survey. Indeed, administrative records are not built for statistical use and therefore they generally adopt different concepts and definitions from the ones used in the survey. Also, they may suffer from quality issues such as under-coverage, lack of timeliness, and errors. These issues should be taken into account when using them for sampling purposes. Still, in the literature or the methodological notes of the surveys, many choices are not documented. For example, it is not always clear how the strata boundaries are chosen, how the allocation is defined, or how the above-mentioned differences are taken into account.

The few studies available are mainly focused on the benefits of using register data. For the US survey on consumer finances, Kennickell (2008) shows that the availability of a list of individuals based on income tax returns produces far more precise estimates of wealth than would be possible with a less-structured sample of the same size, and it provides a framework for correcting for non-response, which is higher among the wealthy. Similar results are found by Bover (2010) as far as the Spanish survey on household finances is concerned. Other research evaluates the effectiveness of the different strategies in obtaining samples that represent adequately the whole distributions of income and wealth (see for instance Household Finance and Consumption Network (2016)).

We contribute to the existing literature in two ways. The first one is that we present a discussion on the challenges and the (expected) benefits of using personal income tax data, drawing on the data of a real survey. In particular, we present a way to address the issue of biased variance estimates based on administrative records. The second contribution of our paper is to present an optimal stratification and sample allocation strategy to be used for multivariate populations. This solution enables us to jointly identify the optimal stratification based on the tax data and the optimal sample size in each stratum. The method presented in the paper has been applied in the 2020 Italian HFCS. Hopefully, our application may contribute to give insights for other data producers.

3. The Italian Survey on Household Income and Wealth

Banca d'Italia conducts the Survey on Household Income and Wealth (SHIW) since the 1960s. Starting from 2010, the survey is part of the Eurosystem's Household Finance and Consumption Survey (HFCS), coordinated by the European Central Bank.

The target population of the survey is all individuals that are officially resident in

Italy. People living in institutions (convents, hospitals, prisons, etc.) or those who are in the country illegally are out of the scope of the survey. The survey is used to collect granular information on many aspects ranging from the socio-demographic characteristics of the household and of its members, to the different sources of income, to the household's assets and liabilities to the consumption and saving behaviors. A household is defined as a person living alone or a group of people who live together in the same private dwelling and share expenditures, including the joint provision of the essentials of living. Persons usually resident, but temporarily absent from the dwelling for less than six months (for reasons of holiday travel, work, education, or similar) are included as household members. On the contrary, possible other persons with usual residence in the dwelling but not sharing expenditures (e.g. lodgers, tenants, etc.) are treated as separate households.

The sample consists of about 8,000 households. The sample size is chosen to produce estimates at the national level. Since 1989 about half of the sample has included households interviewed in previous surveys (panel households). Data collection is entrusted to a specialized company using professional interviewers and CAPI methodology.

The sample is drawn in two stages, with municipalities and households as, respectively, the primary and secondary sampling units. In the first stage, a stratified sample of about 400 municipalities is selected. The variables used for stratification are the region and population size. In the second stage, a simple random sample of households to be interviewed is then selected from the population registers. Participation in the survey is not mandatory. In case a household refuses to participate in the survey, it is replaced by another one living in the same municipality, randomly selected from population registers.

At present, no auxiliary information relating to the household's finances is available at the design stage. This implies that in the final sample only a few rich households are selected. For instance, just by chance, only 80 households belonging to the top 1 percent will be selected. Moreover, once one this household refuses to participate, the available information does not allow replacing them with another with similar finances. Starting from the 2014 wave Banca d'Italia has progressively taken all the legal steps necessary to have access to the fiscal ids of the persons in the sample to make data linkage with register data possible.

4. Register data

In Italy, several public administrations (including the Tax authority) are committed by law to provide their administrative data to the Italian National Statistical Institute (Istat) to reduce the cost of data collection and the burden on the citizens. The two registers (held by Istat) exploited in this work are the Italian Population Register (PR) and the Italian Tax Register (TR).

The PR contains individual records for citizens enrolled in the Italian municipality registers, grouped in their administrative declared households. These registers are regularly updated by municipalities based on the declarations they receive from citizens. Whenever

there is a change in the household composition, such as people getting married or moving to another city, individuals are supposed to communicate this change to the offices in charge of the population register. In most instances some incentives bring people to keep their official records updated: for example, some taxes are lower for houses that are officially primary residences, so in case of purchase of the main residence people immediately update the official records. The PR is used as a sampling frame of all the households surveys in Italy. It is also used to draw the sample of the Italian HFCS for a long time. In this study, we use the version available at the end of 2018.

The second register we use is the Italian Tax Register held by the tax authority. The latest available version of this register has a 2-years time lag, so, the reference time of the TR is 2016 when writing this paper. The TR contains all the records corresponding to the yearly tax declaration of people afferent to the Italian Tax System. It is worthwhile noting that in Italy, people with an income below a certain threshold do not have to provide a tax declaration. Moreover, the TR does not include the income for financial assets (interest and dividends) that generally are taxed with a different system and that are not reported in fiscal declarations. The income variables used in this study are "Total income", "Dependent employment income", "Self-employment income", "Pension income" and "Rent". This information is available at the individual level.

In Italy, the tax agency provides individuals from birth with a unique code, foreigners are provided with the code when they enter the country and ask for permission to stay. The two registers have been linked using these identifiers.

The final data frame contains both demographic information (including household composition) and fiscal incomes at the individual level. The new archive has been created only for the persons living in the municipalities selected as primary sampling units in the survey (around 27.5 million individuals). Individual incomes have then been aggregated at the household level using the official PR definition of household. Households with members with an income higher than a given threshold (1 million euros) have been excluded from the target population for practical reasons: these households are extremely rich and therefore require different/special contact strategies compared to those applied to the rest of the population. They account for 0.01 percent of the total population and 0.6 percent of total income. The final sampling frame consists of about 12 million households.

Register data are not built for statistical use and therefore they adopt concepts and definitions that may be different from those used in the survey. The first one relates to the definition of household composition. Population registers collect information on all the individuals that are officially resident in the same household, while the target of the survey is the "de facto" household composition in the reference year (irrespective of the official residency). The two concepts may differ because of changes that may occur between the selection of data from the registry (September of the reference year) and the time of the interview (from January and June of the year following the year of reference). Moreover, in some instances, people may not have an incentive to update their official status, such as immigrants coming back to their native countries for good. Finally, the official composition

of the household may be affected by the taxation system. For example, a household could be fictitiously divided into two groups for saving taxes linked to the different taxation of the main residence compared to secondary dwellings.

The second difference between register and survey data relates to the definitions of the income sources. In the survey, incomes are collected net of taxes and social contributions, while in the TR each income source is recorded gross and only the total amount of taxes paid by each person is available. Moreover, in the case of self-employed taxable incomes are affected by fiscal rules (such as the possibility of deducting operating losses or investments made in previous years) that do not apply in the survey. Another important incoherence is due to the difference in the methodology for assessing the incomes from non-rented dwellings: in SHIW is adopted the self-assessment methods of the imputed rents, while in the TR the cadastral income (*rendite catastali*) is used for evaluating the stream of these incomes. The cadastral income is a figurative income that can be obtained by multiplying the surface of the property by a specific coefficient, calculated by the Italian Tax Agency according to the municipality, the census zone, the type of dwellings, and its quality. Given that the coefficients are not regularly updated, these incomes significantly underestimate the true value of market rents.

Besides the two differences above mentioned, it worth noting that tax data have quality issues due for instance to tax evasion (Neri & Zizza, 2010; Fiorio & D'Amuri, 2006) and depending on the method used to estimate under-reporting, the magnitude of the problem varies between 7 and 14 percent (Albarea et al., 2017). Moreover, tax data are available with a two-year time lag and therefore may no longer reflect the real situation of the household (especially in the case of self-employed).

One of the main consequences of the above-mentioned issues is that using administrative records for variance estimation in the sample design stage is likely to produce biased results which, in turn, may lead to a sub-optimal selection of the sample.

5. Optimal stratification and sample allocation methodology

Stratification is one of the most widely used techniques in sample survey design, serving the twofold purpose of providing samples that are representative of major subgroups of the population and of improving the precision of estimators.

The design of stratification involves a sequence of decisions relating the choice of the stratification variables, the choice of the number of strata to be formed, the mode in which strata boundaries are determined, the choice of sample size be taken from each stratum (allocation of the sample) and the choice of sampling design within strata.

Studies have provided procedures for the determination of the strata boundaries under a given sample allocation, which are mainly applicable to univariate cases (see for instance Kareem A. O. & Adejumo (2015); Horgan (2006)). On the other hand, there are studies proposing methods to solve the problem of optimum allocation for multivariate populations when the strata are already decided (see for instance Khan (2008)). To the best of our

knowledge, in the literature, there are no studies proposing methods to deal simultaneously with the issue of strata boundaries definition and sample allocation for multivariate populations.

In this paper, we propose the use of a genetic algorithm (Schmitt (2001)) that can explore the universe of all the possible stratifications looking for the one that minimizes the total cost of the sample required to satisfy the precision constraints. This algorithm is implemented in the R package *SamplingStrata* (Barcaroli et al. (2019)). This package, of current use in the Italian National Statistical Institute for various sampling surveys, has been used in the NewZealand Statistical Institute, tested at Statistics Denmark, and considered for evaluation at StatisticsCanada. Eurostat used SamplingStrata for designing its 2018 LUCAS survey (Ballin et al. (2018)). Also, World Bank adopted *SamplingStrata* and embedded it in its SurveySolutions SamplingTools integrated application.

Differently from other similar packages (as the package *stratification* Baillargeon & Rivest (2012)), *SamplingStrata* is applicable to the *multivariate* (more than a target variable) and *multidomain* (more than a domain of estimation) case, that is exactly the Italian HFCS case. The methodology is fully described in Ballin & Barcaroli (2013), Barcaroli (2014) and Ballin & Barcaroli (2016). In the following, we recall its fundamentals before illustrating the application to the SHIW sampling design.

An important step of the method is to estimate consistently the population variance in all the stratum. As already mentioned, register data use different concepts and measures compared to survey data. Moreover, they are likely to suffer from quality issues such as tax evasion and tax elusion and delays. As a consequence, they should not be used as such for the allocation of the sample. In our study, we consider the variables from tax records as proxies of the variables we want to measure. We then estimate measures of goodness-of-fit of these proxies. Finally, we use such measures to inflate our population estimates of the variance in the strata (the higher the goodness-of-fit the lower the inflating factor).

Optimal stratification with the R package SamplingStrata

In a stratified sampling design with one or more stages, a sample is selected from a frame containing the units of the population of interest, stratified according to the values of one or more auxiliary variables (X) available for all units in the population. For a given stratification, the overall size of the sample and the allocation in the different strata can be determined on the basis of constraints placed on the expected accuracy of the various estimates regarding the survey target variables (Y). If the target survey variables are more than one the optimization problem is said to be multivariate; otherwise it is univariate. For a given stratification, in the univariate case the optimization of the allocation is in general based on the Neyman allocation (Cochran (1977)). In the multivariate case it is possible to make use of the Bethel algorithm (Bethel (1989)). The criteria according to which stratification is defined are crucial for the efficiency of the sample. With the same precision constraints, the overall size of the sample required to satisfy them may be

significantly affected by the particular stratification chosen for the population of interest. Given G survey target variables, their sampling variance is:

$$Var(\hat{Y}_g) = \sum_{h=1}^{H} N_h^2 (1 - \frac{n_h}{N_h}) \frac{S_{h,g}^2}{n_h} \quad g = 1, ..., G$$

If we introduce the following cost function:

$$C(n_1, ..., n_H) = C_0 + \sum_{h=1}^{H} C_h n_h$$

the optimization problem can be formalized in this way:

$$min = C_0 + \sum_{h=1}^{H} C_h n_h$$

under the constraints

$$\begin{cases} CV(\hat{Y}_1) < U_1 \\ CV(\hat{Y}_2) < U_2 \\ ... \\ CV(\hat{Y}_G) < U_G \end{cases}$$

where

$$CV(\hat{Y}_g) = \frac{\sqrt{Var(\hat{Y}_g)}}{mean(\hat{Y}_g)}$$

SamplingStrata allows performing the optimization steps in two different ways, depending on the nature of the stratification variables Xs.

Optimization with categorical stratification variables

Given a population frame with m auxiliary variables $X_1, ..., X_M$we define as *atomic stratification* the one that can be obtained considering the cartesian product of the definition domains of the m variables. To each atomic stratum relevant information is attached:

- the values assumed by the stratification variables Xs;
- the population N (number of units in the sampling frame belonging to the stratum);
- values of means and standard deviations associated to each target variable Y;
- the cost C to observe a unit in the stratum.

Starting from the atomic stratification, it is possible to generate all the different combinations that belong to the universe of stratifications. The number of feasible strata is exponential with respect to the number of initial atomic strata. In concrete cases, it is therefore impossible to examine all the different possible alternative stratifications. The genetic algorithm explores the universe of stratifications by performing the following steps:

1. an initial set (*generation*) of strata (*individuals*) is randomly generated: a given individual is characterized by a stratification where each atomic stratum is attributed to one aggregate stratum identified by a combination of values of the stratification variables;

2. for each aggregate stratum the information required is calculated (population, means and standard deviations of Ys, cost) and its (*fitness*) (total cost of the sample required to satisfy precision constraints) is determined by applying the Bethel algorithm;

3. next set of individuals is generated by applying the usual operators of the genetic algorithm: *selection*, *crossover* and *mutation*.

Step 3 is repeated a given number of times. At the end, the individual with the best fitness (i.e. the stratification with the minimum cost of the associated sample) is retained as the best solution.

Optimization with continuous stratification variables

When all the stratification variables are continuous (or even categorical, but of the ordinal type), a variant of the above optimization step is applicable. Instead of generating the atomic strata as a preliminary step, the algorithm provides to generate aggregate strata for each individual by operating in this way:

- for each continuous stratification variable, a predetermined number of values internal to its definition domain are randomly generated: these values (*cuts*) determine a segmentation of the domain that is equivalent to a categorization of the variable;

- aggregate strata are consequently determined by cross-classifying units in the sampling frame accordingly to their values belonging to the segments previously defined.

After this, the sequence of optimization is identical to the one seen in the case of categorical stratification variables.

Anticipated variance

In real situations, the information contained in the sampling frame is not directly regarding the target variables of the survey, but *proxy* variables, i.e. variables that are

correlated to the variables of interest. In our application, we know that income from self-employement collected in tax records is based on fiscal rules In order to take into account this problem, and to limit the risk of overestimating the expected precision levels of the optimized solution, it is possible to carry out the optimization by considering, instead of the expected coefficients of variation related to proxy variables, the anticipated coefficients of variation (ACV) that depend on the model that is possible to fit on couples of real target variables and proxy ones. In the current implementation, only models linking continuous variables can be considered. The definition and the use of these models is the same that has been implemented in the package *stratification* (Baillargeon & Rivest (2012)). In particular, the reference here is to two different models (applicable only to continuous variables):

1. the linear model with heteroscedasticity: $Y = \beta \times X + \epsilon$, with $\epsilon \sim N(0, \sigma^2 X^\gamma)$ (where γ indicates the heteroscedasticity)
2. the log-linear model: $Y = \exp(\beta \times log(X) + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$

After fitting one model for each couple target / proxy variables, their parameters are given as an additional input to the optimization function of *SamplingStrata*. The optimization step will be then performed by calculating correctly the distributional values (means and standard deviations).

6. Application to the Italian HFCS

The method described in the previous sections has been applied to the 2020 wave of the Italian HFCS survey. In particular, it has been used in the second stage of the design to select non-panel households.
As already mentioned, register data use different concepts and definitions from the survey and, also, they have several quality issues. As a result, the information on household income coming from tax records is only a proxy of the actual economic situation. As a first step, we estimate the goodness of these proxies. To this purpose, we use the refresh sample selected for the 2016 wave. These data have been linked to the Tax Register via individual ids. Considering respondents only, the link was successful for 4,328 households. For these units, we have information on the reported values for the five target variables ("Total income", "Dependent employment income", "Self employment income", "Pension income", "Rents") and the corresponding fiscal values. The associations between the two types of information are reported in table 1.

There is an evident variability in the goodness of fitting: from a 68% in the case of "Dependent employment income" to a 13% in the case of "Rents".
As a second step, we chose the precision constraints in terms of the maximum expected coefficient of variation for the target variables in the different domains. The precision constraints are set equal to 5% in every domain and for all domains.

Table 1: Linear regression models between observed variables and Tax Register variables (Italian HFCS, 2016 wave).

Target variable	R^2	$Beta$	$Sigma^2$
Total income	0.5771541	0.8417096	11945.78
Dependent employment income	0.6835152	0.8229064	12547.71
Self employment income	0.2304688	0.5571044	18639.69
Pension income	0.6364706	0.7665643	5834.692
Rents	0.1366157	0.1653843	0.5436948

We then run the optimization step to define the strata, the sample size, and its allocation. We use the sampling frame described in section 4, containing 12,351,950 units (households). For operative reasons, we remove from the population of interest all the households with a source of income above 1 million euros. Since these households are extremely rich, they require a different contacting strategy that would probably result in an excessive increase in the survey costs. So, the resulting final population size is 12,334,342. The excluded households hold about 0.6 percent of total fiscal income.

Numerous executions of this step have been attempted, varying the kind of optimization (with categorical or continuous variables) and the maximum number of final strata. Even if stratification variables are continuous, we try the first algorithm after their categorization (obtained by applying the univariate k-means clustering method). The comparison with the results obtained with the second algorithm (directly applied to stratification variables as they are) is in favor of the latter.

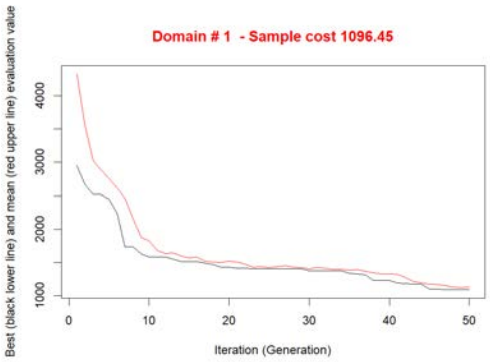
Another important decision is to fix the number of optimized strata to be expected in each one of the 5 territorial domains (NUTS1). The determination has been done by using a particular function available in *SamplingStrata*, which is the sequential application of k-means algorithm, varying the number of the cluster from a minimum (usually 2) to a maximum. The indication was to set this value to 10. The minimum number of units per stratum is set to 50 households (this choice is based on operative considerations).

The optimization has been carried out distinctly for the various domains. The number of iterations was set to 50, for each iteration 20 different solutions were generated, for a total of 1,000 solutions evaluated by applying the Bethel algorithm. The search for an optimal solution shows a common trend in each domain (see figure 1).

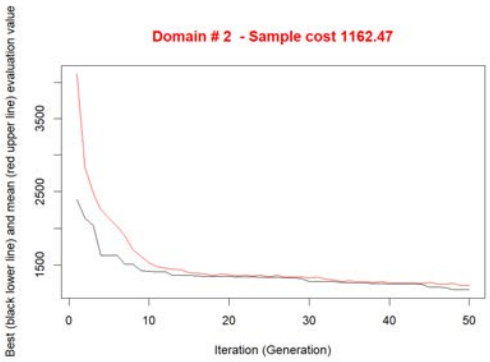
The overall sample size required to satisfy the precision constraints under the optimal solution is equal to 6,400.

The package allows visualizing in a two-dimensional graph the obtained strata, each time choosing a couple of variables. For instance, figures 2 is reported the characterization of the strata in the first domain by considering "Total income" and "Dependent employment income".

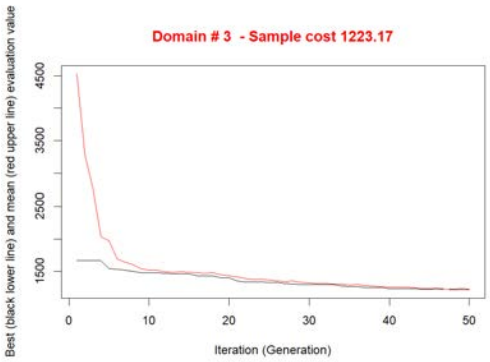
Figure 1: Optimization in the different domains



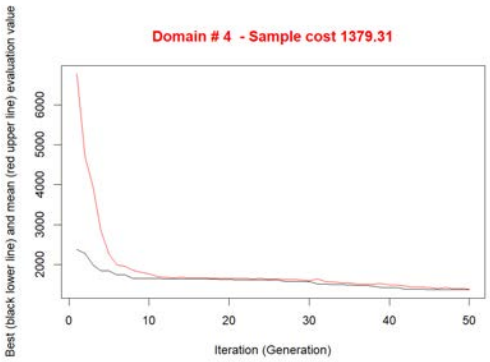
(a) North-west



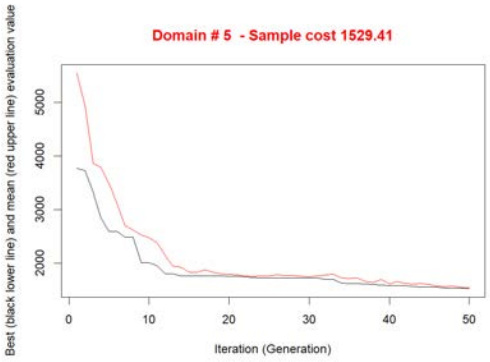
(b) North-east



(c) Center

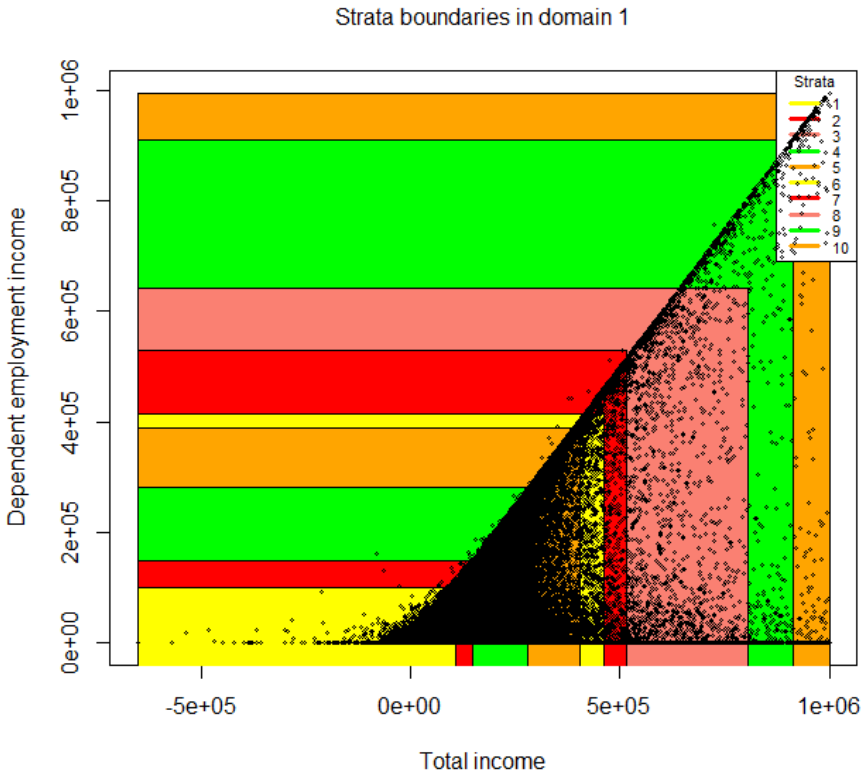


(d) South



(e) Islands

Figure 2: Strata resulting from the execution of the genetic algorithm (domain 1).



In figure 3 optimized strata with population, sampling allocation, and sampling rates are reported together with the range of the two stratification variables.

Figure 3: Strata population, allocation and range of stratification variables (domain 1).

Stratum	Population	Allocation	SamplingRate	Bounds Total income	Bounds Dependent employment income
1	2384770	365	0.0001529532	-652064-105895	0-99369
2	523170	272	0.0005193699	-282649-147433	0-162801
3	59346	55	0.0009336701	-421027-149039	0-149002
4	35528	50	0.0014073407	25566-280219	0-281464
5	58824	105	0.0017776218	12513-405510	0-387762
6	3846	50	0.0130005200	93801-462718	0-413450
7	2203	50	0.0226963232	161895-515268	0-530787
8	5028	50	0.0099443119	59136-804789	0-641535
9	938	50	0.0533049041	392890-912168	0-909600
10	1137	50	0.0439753738	511877-999682	0-995589

The solution is characterized by a sample size equal to 6,400, and the expected coefficients of variations have been calculated assuming that all sampled units will respond to the interviewers. This assumption is far from reality.

As a final step, we need to estimate the total sample that is required to get a final sample of around 6,400 households. Using the sample selected for the 2016 survey linked to tax records, we link both respondents and nonrespondents to the Tax Register. We then estimate a model for the probability of participating in the survey using as predictors the four components of income (with the exclusion of the "Total income") and the twenty NUTS2 Italian regions.

Considering the plot in figure 4, there is clear evidence of a linear direct inverse relationship between the log of the mean income in a stratum, and the propensity to respond. The sample of units to be interviewed has been redefined by taking into account the propensity to nonresponse calculated for each unit in the sampling frame using the above model. The total number of households to be interviewed is 17,608, units that have been allocated in the optimized strata taking into account the initial allocation and the average propensity to the response.

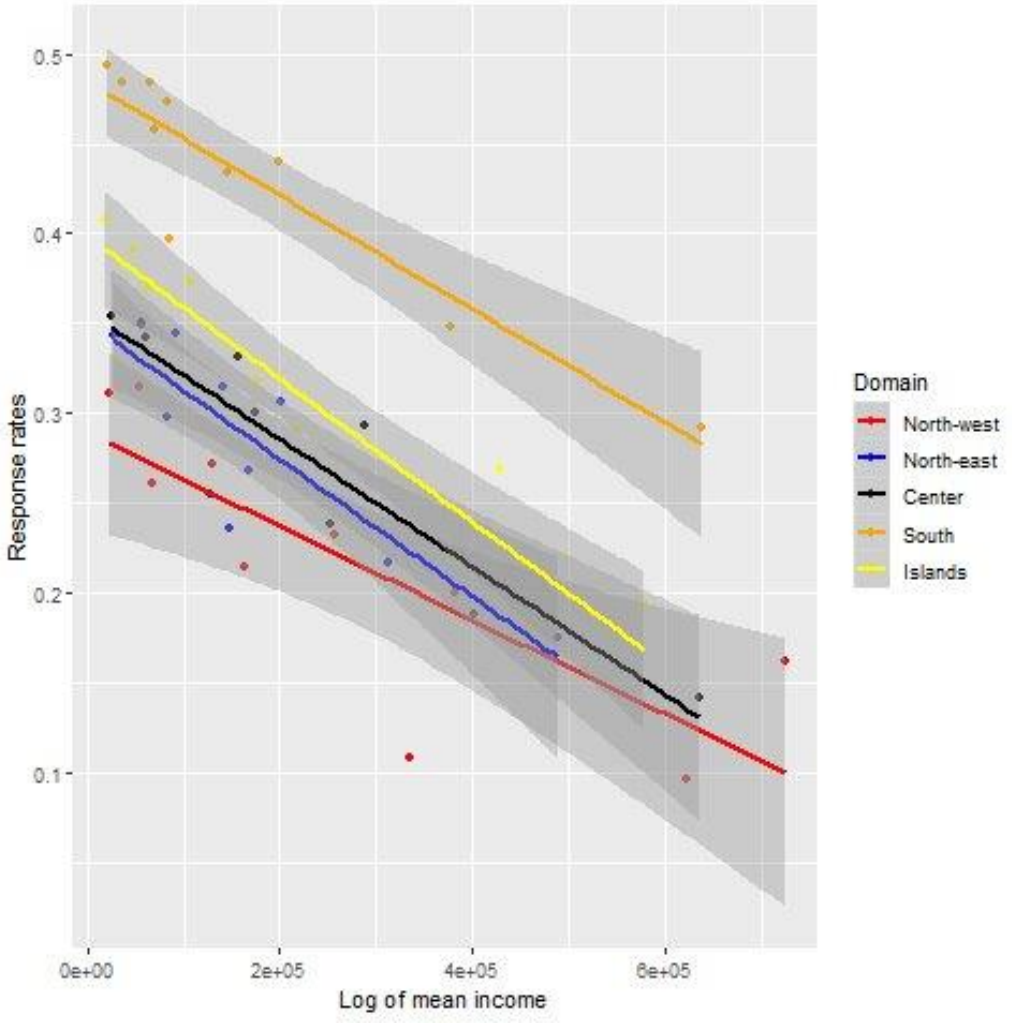


Figure 4: Response rate and mean income in strata.

For example, in table 2 has been reported the final solution, with the initial and final allocation, for the first domain.

Table 2: Optimal Stratification, initial and final allocation.

Domain	Stratum	Population	Allocation	New allocation	Sampling rate
1	1	2384770	365	1064	0.000446
1	2	523170	272	784	0.001499
1	3	59346	55	192	0.003235
1	4	35528	50	211	0.005939
1	5	58824	105	350	0.005950
1	6	3846	50	195	0.050702
1	7	2203	50	420	0.190649
1	8	5028	50	226	0.044948
1	9	938	50	469	0.500000
1	10	1137	50	280	0.246262

Table 3 reports the coefficients of variation achievable with the selected sample (6,400 units). The solution allows meeting all the precision requirements. It can be seen that for the first variable ("Total income") the precision is about double than prescribed.

Table 3: Expected coefficients of variation (%).

Domain	Total income	Dependent emp.income	Self emp.income	Pension income	Rents
1. North west	2.5	5.0	4.8	4.9	4.8
2. North east	2.4	4.7	4.9	4.6	4.8
3. Center	2.6	4.8	5.0	5.0	4.7
4. South	2.3	4.3	4.8	4.8	4.9
5. Islands	2.3	4.0	5.0	5.0	4.9

These estimates of the expected CVs have been calculated (using a specific function in the package *SamplingStrata*) assuming that:

1. the survey adopts a single stage sampling process;
2. estimates are obtained by Horvitz-Thompson estimator;
3. all 6,400 units in the sample respond to the survey.

7. Evaluation of the new sample design

In this paragraph, we run several simulations to have a more robust evaluation of the new design. Each simulation is based on the archive created by linking the Population and the Tax registers and on the information coming from the 2016 SHIW survey integrated with tax records. In the simulations, we extract 500 samples using both the new and the old design and we compute measures of precision and bias of the five income estimators. The difference between the two types of simulation is the following. In the first set, we only use the information on Population Register for the calibration of final weights, in line with what is currently done in the SHIW survey. In the second set of simulations, we also use tax records in the weighting stage.

Each simulation is based on the following assumptions:

1. the survey uses a two-stage sampling design, so when evaluating variance of estimates, weights associated with Primary Sampling Units (the municipalities selected at the first stage) have to be taken into account;
2. estimates are obtained by calibration estimators, to handle total nonresponse;
3. the final sample size has been inflated to 17,608 households to take into account the expected nonresponse.

7.1 Simulations using Population Register for calibration

The first simulation consists of the following steps.
First, we use the models introduced in par. 6 to predict, for each unit in the sampling frame, the values of target variables.
Then, 500 samples of the required size (17,608 households) have been selected from the sampling frame. For each household, we simulate the nonresponse mechanism using the model described in the previous section. The decision to participated is then taken by drawing a value from a Bernoulli variable with the probability of success (the propensity to respond) equals the propensity estimated by the nonresponse model.
For each sample of respondents, the final weights are calibrated using the total number of households in the Population Register.
In the end, coefficients of variation and relative bias have been calculated, averaging over the 500 replicated samples. Bias is measured as the difference between the mean value of the 500 survey-based estimates and the population means coming from administrative records.
Results are reported in tables 4 and 5. The precision of the estimators is in line with one of the selected sample.

The simulation shows the presence of a negative bias for incomes from employment and rents. The opposite situation holds for incomes from self-employment and pensions.

Table 4: Estimated coefficients of variation of the new design (%).

Domain	Total income	Dependent emp.income	Self emp.income	Pension income	Rents
1. North west	2.6	5.4	4.3	4.8	3.5
2. North east	2.4	4.8	4.5	4.5	3.3
3. Center	2.4	4.8	3.7	4.5	3.1
4. South	2.3	4.4	3.5	4.6	3.3
5. Islands	2.3	4.1	3.5	4.8	3.1

Table 5: Estimated relative bias of the new sample design (%).

Domain	Total income	Dependent emp.income	Self emp.income	Pension income	Rents
1. North west	-3.8	-12.80	5.64	8.25	-1.49
2. North east	-2.6	-8.8	3.0	6.4	-2.0
3. Center	-2.7	-10.5	4.5	8.5	-1.5
4. South	-2.2	-6.5	0.6	4.4	-3.0
5. Islands	-2.1	-8.0	1.4	5.7	-1.4

The presence of bias depends on our response probability model, which is estimated using household-specific administrative information. In some strata, this model generates a high (within) variability of response propensities. Therefore, a simple calibration of the weights of respondents to the total number of households in the population is not enough to compensate for missing households.

The old sample design is a two-stage process where the first stage is identical to the new one, with the selection of the same 454 municipalities (via PPS). The allocation of SSU units is based on the following rule: if the total population in the selected municipality is higher than 500,000 then 200 households are assigned, otherwise only 32. The total amount of SSU units is 14,864. Based on this SSU stratification and allocation, we run a sample of 6,400 units for the frame. This sample represents therefore the one we have selected using the old design. The expected CVs for the selected sample are reported in table 11.

This table has been computed using the same assumptions made for table 3. By comparing the two, it is clear that the expected CVs for the old design are higher than those calculated for the new one. In particular they are much higher for *Self employment income* and *Rents*.

For comparison, we report in tables 7 and 8 the observed CVs of the target variables computed using the 2014 and 2016 Italian HFCS. These tables are not directly comparable with the previous one for two main reasons. First, the sample size if is larger (about 8,000 households for each wave). Second, the sampling weights are calibrated in a way

Table 6: Expected coefficients of variation for the old sample (%).

Domain	Total income	Dependent emp.income	Self emp.income	Pension income	Rents
1. North west	4.6	6.3	23.1	6.4	17.2
2. North east	3.7	5.0	20.4	5.6	15.4
3. Center	4.3	5.8	23.4	6.8	17.3
4. South	4.6	5.8	25.4	6.7	21.3
5. Islands	6.0	7.8	32.5	9.8	26.1

that is not possible for the 2020 survey since we miss some demographic information on respondents. The possibility to calibrate using other information (such as the job status) contributes to reducing the final variability of the estimators. Still, two important points can be drawn from these tables. First, the expected CVs shown in this paper are probably upper bounds for the actual ones that will be observed for the 2020 wave. Second, the advantage of the new design is also in reducing the instability of the estimators across surveys. This is particularly the case for incomes from self-employment and rents, which show significant changes in the precision from one wave to another. This is because the available information does not allow us to have full control of the final sample composition. This situation will change thanks to the new design.

Table 7: coefficients of variation estimated in the 2016 Italian HFCS wave(%).

Domain	Total income	Dependent emp.income	Self emp.income	Pension income	Rents
1. North west	2.1	3.8	11.0	4.1	25.4
2. North east	3.4	4.4	12.5	4.1	14.2
3. Center	2.3	5.3	11.5	4.8	18.9
4. South	2.5	4.6	14.1	4.5	28.5
5. Islands	3.1	4.9	22.4	5.2	34.4

Table 8: coefficients of variation estimated in the 2014 Italian HFCS wave(%).

Domain	Total income	Dependent emp.income	Self emp.income	Pension income	Rents
1. North west	2.2	2.9	8.2	3.8	11.3
2. North east	1.9	2.7	9.6	4.4	14.4
3. Center	2.4	3.4	18.3	4.0	10.4
4. South	2.8	4.6	21.4	3.7	22.2
5. Islands	2.7	4.1	12.7	7.3	44.4

Following the approach used in the previous section, we then run a simulation based on the old design. In particular, we perform the following steps:

1. 500 samples have been drawn from the same sampling frame, i.e. the one enriched by predicted target variables;
2. for each sample, the mechanism of nonresponse has been simulated accordingly to the predicted nonresponse propensity associated with each unit in the frame;
3. for each resulting sample of respondents, calibrated estimates of interest have been calculated, where known totals are given by the number of households in the different strata.

In other words, the simulation has been carried out with the same setting used for the new sample design.

In the end, coefficients of variation and relative bias for the old sample design have been calculated, averaging over the 500 replicated samples. Results are reported in tables 9 and 10.

Table 9: Estimated coefficients of variation of the old sample designs (%).

Domain	Total income	Dependent emp.	Self-emp.	Pensions	Rents
1. North west	6.18	9.50	31.73	10.34	6.61
2. North east	4.96	8.17	25.77	9.34	5.48
3. Center	5.51	8.68	22.22	10.00	5.68
4. South	4.85	7.54	19.60	8.18	5.30
5. Islands	7.42	11.33	29.30	14.26	7.79

Table 10: Estimated relative bias of the old sample designs (%).

Domain	Total income	Dependent emp.	Self-emp.	Pensions	Rents
1. North west	-5.92	-14.55	-6.78	8.12	-2.71
2. North east	-4.20	-10.15	-6.26	5.97	-2.03
3. Center	-4.60	-11.39	-7.39	7.14	-2.55
4. South	-2.85	-7.19	-3.8	4.03	-2.12
5. Islands	-3.18	-8.36	-4.40	4.71	-2.22

Figure 5: Comparison of coefficients of variation obtained for the new and old sample designs.

Figures 5 and 6 summarise the over-performance of the new sample compared to the old sample in terms of both coefficients of variation and bias, respectively.

It can be seen that as for the CVs, there is a clear indication of the superiority of the new design compared to the old one in terms of the sampling variance component of MSE.

As for the bias, there is a slight prevalence of the new sample design, with only 6 cases out of 25 in which the old design performs better.

7.2 Simulations using Tax Register for calibration

In the previous simulations, we do not use the known totals available from the Tax Register, i.e. the sum of the components of the income (Dependent Employment, Self Employment, Pensions, Rents) by the different domains of interest (the five Italian NUTS1 geographical zones).

To fully exploit the information achievable in the administrative sources, we carried out the same simulations described before but using a different calibration model: instead of the known totals of households in the strata defined by the old and new sampling designs, we made use of both totals of households at NUTS1 level and the Tax Register incomes at

Figure 6: Comparison of relative bias obtained for the new and old sample designs.

stratum level.

Results in terms of CVs and bias are reported in tables 11 and 12.

Table 11: Estimated coefficients of variation of the new and old sample designs (%) with calibration using Tax Register variables.

	Total income		Dependent emp.		Self-emp.		Pensions		Rents	
Domain	New	Old	New	Old	New	Old	New	Old	New	Old
1. North west	0.54	0.99	0.23	0.41	1.93	3.10	0.53	1.06	3.34	6.17
2. North east	0.46	0.72	0.21	0.39	1.80	3.26	0.48	0.72	2.73	3.70
3. Center	0.51	0.76	0.24	0.31	2.02	3.42	0.53	0.79	2.70	4.20
4. South	0.55	0.71	0.24	0.48	2.16	2.89	0.56	0.84	2.89	3.80
5. Islands	0.52	1.21	0.24	0.63	2.14	4.33	0.54	1.16	2.61	5.76

The distribution of the 500 replicated estimates is reported in figure 7, only for the first domain and only for the new sample design.

Table 12: Estimated relative bias of the new and old sample designs (%) with calibration using Tax Register variables.

	Total income		Dependent emp.		Self-emp.		Pensions		Rents	
Domain	New	Old	New	Old	New	Old	New	Old	New	Old
1. North west	0.19	0.32	0.08	0.16	0.20	0.63	-0.02	-0.28	1.00	1.93
2. North east	0.00	-1.77	0.08	-2.09	0.17	-1.20	-0.04	-2.48	-0.26	0.44
3. Center	0.22	0.59	0.08	0.00	-0.02	0.94	-0.01	0.17	1.25	3.21
4. South	-0.21	-1.43	0.02	-2.02	-0.57	-0.58	-0.07	-1.88	-0.99	0.78
5. Islands	0.09	0.44	0.05	-0.17	0.22	2.26	0.03	-0.53	0.27	3.06

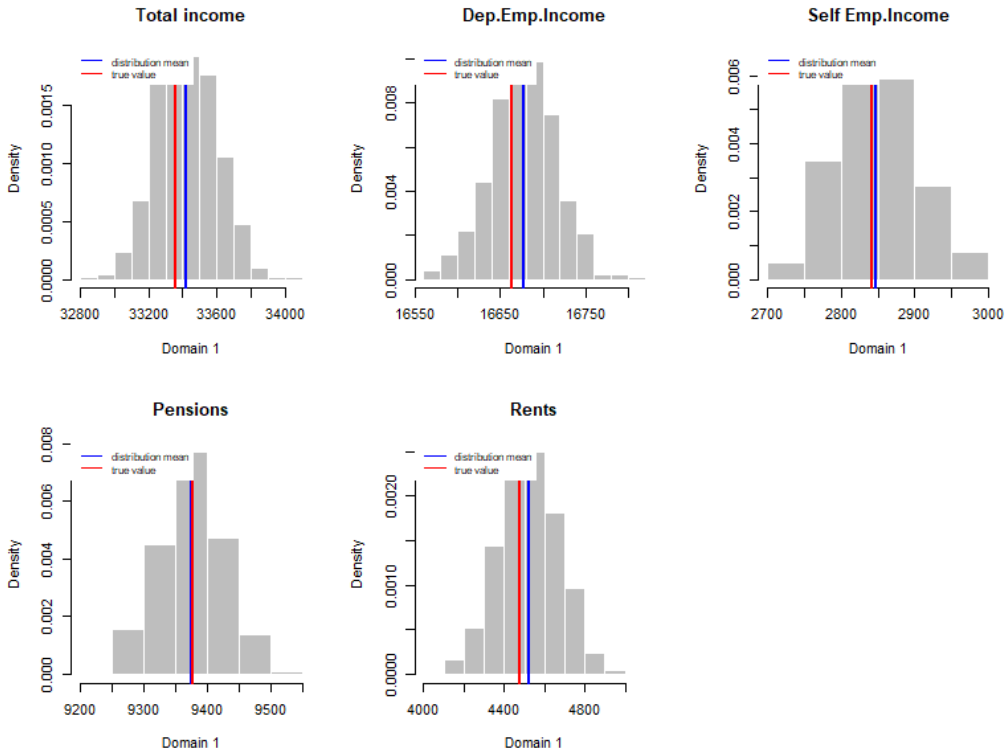


Figure 7: Distribution of the 500 replicated estimates in the first domain (new design, calibration adding Tax Register totals).

There is an evident reduction of CVs and bias for both new and old sample design, with a comparison always in favor of the new design.

This simulation is only indicative of the potential of this calibration because results so positive depend on the fact that the target values in the frame have been generated

by models that make use of the same Tax Register variables as explanatory variables and using the same Tax Register variables as known totals in the calibration model introduces a great simplification of the real situation that may somehow compromise the full validity of these results. Nonetheless, it is expected that a model-assisted approach which includes also Tax Register variables would substantially improve the accuracy of the estimates.

8. Conclusions

The paper presents an empirical application of tax personal income data in the sampling design of finance surveys. Tax data are not collected for statistical purposes and therefore they use definitions and measures different from those adopted in the survey. Furthermore, they are subject to various quality problems (such as tax avoidance or evasion, the presence of thresholds below which the declaration is not necessary, and time delays before becoming available).

As a consequence, their use for statistical purposes is not straightforward. Nonetheless, this application has shown that one possible solution is to consider them as proxies for the variables of interest and to inflate the estimators of variance used for determining sample size accordingly. We are able to estimate the goodness of these proxies by linking survey data to administrative records. Our simulations show that their use enables us to take under control the expected accuracy of income estimators, despite all the limits of tax data. A second (and strictly related) advantage is that the availability of register data enables us to keep under control the fieldwork of the survey. This implies, for instance, specific households can be oversampled and those refusing to participate could be replaced with others belonging to the same stratum. This should guarantee to obtain a final sample, which is very close to the selected one, i.e. the most efficient one. Consequently, the expected benefits in terms of variance reduction should turn into effective advantages.

Another potential advantage is linked to the possibility of reducing bias due to non-response. Our simulation has shown that the new sample design allows not only greatly reducing the sampling variance, but also the bias component of the Mean Square Error of estimates even if we do not include Tax Register variables in the calibration model. If we include also these variables, results in terms of an overall reduction of MSE should be even greater.

References

Albarea, A., Bernasconi, M., Marenzi, A., & Rizzi, D. (2017). *Income under reporting and tax evasion in Italy*. Documento di Valutazione, 8 Senato della Repubblica.

Baillargeon, S., & Rivest, L.-P. (2012). *stratification: Univariate Stratification of Survey Populations*. R package version 2.2-3.

Ballin, M., & Barcaroli, G. (2013). Joint determination of optimal stratification and sample allocation using genetic algorithm. *Survey Methodology*, *39*, 369–393.

Ballin, M., & Barcaroli, G. (2016). Optimization of stratified sampling with the r package samplingstrata: applications to network data. In M. Dehmer, Y. Shi, & F. Emmert-Streib (Eds.), *Computational Network Analysis with R: Applications in Biology, Medicine and Chemistry* chapter 5. (pp. 125–150). Wiley.

Ballin, M., Barcaroli, G., Masselli, M., & Scarnò, M. (2018). *Redesign sample for Land Use/Cover Area frame Survey (LUCAS) 2018*. Statistical Working Papers Eurostat.

Barcaroli, G. (2014). SamplingStrata: An R package for the optimization of stratified sampling. *Journal of Statistical Software*, *61*, 1–24.

Barcaroli, G., Ballin, M., Odendaal, H., Pagliuca, D., Willighagen, E., & Zardetto, D. (2019). *SamplingStrata: Optimal Stratification of Sampling Frames for Multipurpose Sampling Surveys*. R package version 1.5-1.

Bethel, J. (1989). Sample allocation in multivariate surveys. *Survey Methodology*, *15*, 47–57.

Bover, O. (2010). Wealth Inequality And Household Structure: U.S. Vs. Spain. *Review of Income and Wealth*, *56*, 259–290.

Bover, O., Coronado, E., & Velilla, P. (2014). The spanish survey of household finances (eff): Description and methods of the 2011 wave. *Banco de Espana Occasional Paper*, *1407*.

Casiraghi, M., Gaiotti, E., Rodano, L., & Secchi, A. (2018). A "reverse Robin Hood"? The distributional implications of non-standard monetary policy for Italian households. *Journal of International Money and Finance*, *85*, 215–235.

Chakraborty, R., Kavonius, I., Perez-Duarte, S., & Vermeulen, P. (2019). Is the top tail of the wealth distribution the missing link between the household finance and consumption survey and national accounts? *Journal of official Statistics*, *35*, 31–65.

Cochran, W. (1977). *Sampling Techniques*. (Third edition ed.). New York: John Wiley Sons.

Colciago, A., Samarina, A., & de Haan, J. (2019). Central bank policies and income and wealth inequality: A survey. *Journal of Economic Surveys*, *33*, 1199–1231.

D'Alessio, G., & Neri, A. (2015). *Income and wealth sample estimates consistent with macro aggregates: some experiments*. Questioni di Economia e Finanza (Occasional Papers) 272 Bank of Italy, Economic Research and International Relations Area.

Dobbs, R., Lund, S., Koller, T., & Shwayder, A. (2013). *QE and ultra-low interest rates: Distributional effects and risks*. Report McKinsey Global Institute.

ECB (2009). *Survey data on household finance and consumption: research summary and policy use* . Technical Report 100 Eurosystem Household Finance and Consumption Network.

Eckerstorfer, P., Halak, J., Kapeller, J., Schütz, B., Springholz, F., & Wildauer, R. (2016). Correcting for the missing rich: An application to wealth survey data. *Review of Income and Wealth*, *62*, 605–627.

Fiorio, C. V., & D'Amuri, F. (2006). Tax Evasion In Italy: An Analysis Using A Tax-Benefit Microsimulation Model. *The IUP Journal of Public Finance*, (pp. 19–37).

Horgan, J. M. (2006). Stratification of skewed populations: A review. *International Statistical Review*, *74 (1)*, 67–76.

Household Finance and Consumption Network (2016). *The Household Finance and Consumption Survey: Methodological report for the second wave*. ECB Statistics Paper Series 17, European Central Bank.

Kareem A. O., O. O. G. M. O., I, & Adejumo, A. O. (2015). Moving average stratification algorithm for data boundary determination in skewed populations. *Journal of Applied Statistics*, *6 (1)*, 205–217.

Kennickell, A. (2001). Modeling wealth with multiple observations of income: Redesign of the sample for the 2001 survey of consumer finances. *Statistical Journal of the IAOS*, *33*.

Kennickell, A. (2008). The role of over-sampling of the wealthy in the survey of consumer finances. *Irving Fisher Committee Bulletin*, *28*.

Kennickell, A. (2019). The tail that wags: differences in effective right tail coverage and estimates of wealth inequality. *The Journal of Economic Inequality*, .

Khan, N. N. A. N., M.G.M. (2008). Determining the optimum strata boundary points using dynamic programming. *Survey Methodology*, *34* , 205–214.

Michelangeli, V., & Rampazzi, C. (2016). Indicators of financial vulnerability: a household level study. *Questioni di Economia e Finanza (Occasional Papers),Banca d'Italia, Economic Research and International Relations Area.*, *369*.

Neri, A., & Zizza, R. (2010). *Income reporting behaviour in sample surveys*. Temi di discussione (Economic working papers) 777 Bank of Italy, Economic Research and International Relations Area.

Ranalli, M. G., & Neri, A. (2011). To misreport or not to report?, The case of the Italian survey on household income and wealth. *Statistics in Transition new series*, *12*, 281–300.

Schmitt, L. (2001). Theory of genetic algorithms. *Theoretical Computer Science*, *259*, 1–61.

Vermeulen, P. (2018). How fat is the top tail of the wealth distribution? *Review of Income and Wealth*, *64* , 357–387.