Optimal sampling design for household finance surveys using administrative income data

Giulio Barcaroli [1], Giuseppe Ilardi [2], Andrea Neri [2], Tiziana Tuoto [3]

Abstract

Household finance surveys, which collect detailed information on household income and wealth, are increasingly used for policy-making. They should provide an accurate picture of the economic situation of all households. Unfortunately, the upper parts of the wealth distribution are often missing in household surveys. Since rich households concentrate a large share of total income and wealth, survey-based estimators may be biased. The ideal situation would be to have access to auxiliary information on household finances at the design stage. This is rarely the case. In this paper we present an application that uses tax records in the design of a major survey on household finances. We discuss the methodological challenges of using administrative information for designing the sample. We propose a method for an optimal stratification and sample allocation.

Keywords: Household finance surveys, Household Finance and Consumption Survey - HFCS, register data, tax records, income, sampling design, optimal stratification, calibration.

1 Independent expert, formerly at the Italian National Institute of Statistics - Istat (gbarcaroli@gmail.com).
2 Central Bank of Italy/*Banca d'Italia* (Giuseppe.Ilardi@bancaditalia.it; Andrea.Neri@bancaditalia.it).
3 Italian National Institute of Statistics - Istat (tuoto@istat.it).

The views and opinions expressed are those of the authors and do not necessarily reflect the official policy or position of the Italian National Institute of Statistics - Istat.

The authors would like to thank the anonymous reviewers for their comments and suggestions, which enhanced the quality of this article.

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 recognised to interact with their functions (Eurosystem Household Finance and Consumption Network, 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 and Rampazzi, 2016). Moreover, Central Banks are also increasingly interested in understanding the effects of their policies on the household's economic conditions and in particular on income and wealth inequality (Casiraghi *et al.*, 2018; Colciago *et al.*, 2019; Dobbs *et al.*, 2013; Dossche *et al.*, 2021).

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 harmonised 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; Neri and Ranalli, 2011; D'Alessio and 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 availability of auxiliary information at the design stage (such as administrative records relating to household finances) would prove extremely effective in addressing these issues. Such information would enable survey agencies to identify correctly this rare population, also making it possible to oversampling it to compensate for the difficulties in enrolling it in the survey. Unfortunately, such 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* (the Central Bank of Italy) which runs the survey, and the Italian National Statistical Institute – Istat which 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 HFCS survey is a two-stage sample with municipalities selected as primary sampling units (PSUs) and households selected as second-stage units (SSUs). In this paper, we discuss the first time that the information from the personal income register is used to optimise the sample design, focussing on the second stage, while treating the first-stage sample as fixed. A more general and even complex optimal sampling design, which also considers first stage units, is possible and desirable. However, the impact of such a design would have on the organisational procedures that support the survey would certainly be heavy. Therefore, to introduce and manage innovations gradually, it was a survey requirement to deal with the second stage treating the first stage as fixed.

The paper is organised 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 Sections 5 and 6 describe 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, 2017).

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 2017 wave of the HFCS, seventeen out of twenty-two countries used different strategies to oversample richer households (Household Finance and Consumption Network, 2020). Italy was one of the five countries in the HFCS which had no access to auxiliary information that could be used in the sampling design. The oversampling 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 of 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 not 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. They may also 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 focussed 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 behaviours. 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 specialised 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 such a 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 household 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-year 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 taxable income of people afferent to the Italian Tax System. It is worthwhile noting that in Italy, people with an income below certain thresholds do not have to provide a tax declaration. Yet, the TR is based on multiple sources which enable to recover the information also for those who are below these thresholds. The main limit of the TR is that it 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 (according to national accounts, interests and dividends account for about 15 percent of household disposable income). This data gap may limit the utility of the personal tax data to target wealthy households, which usually concentrate a large share of financial wealth.

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 included in a separate self-representing stratum. It accounts for 0.01 percent of the total population and 0.6 percent of total income. Since the households in this stratum represent a very hard-to-reach population which may require different *ad hoc* strategies, we exclude them from the present analysis. The final sampling list 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. SHIW and surveys in general use "economic household" concept, *i.e.* those actually living together and sharing the essentials of living (Jäntti *et al.* 2013).

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, that is the amount of income a property owner would get by renting her/his own house: in SHIW is adopted the self-assessment method which consists in asking directly the respondents to provide their best estimate, 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 and Zizza, 2010; Fiorio and 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.*, 2018). 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.

In SHIW/HFCS, the particular aim of the stratification should be to increase precision in the top of the wealth distribution. So far, this has not been implemented in Italy.

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 and 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 minimises the total cost of the sample required to satisfy the precision constraints. This algorithm is implemented in the *R* package *SamplingStrata* (Barcaroli *et al.*, 2020). This package, of current use in the Italian National Statistical Institute for various sampling surveys, has been used in the New Zealand Statistical Institute, tested at Statistics Denmark, and considered for evaluation at Statistics Canada. Eurostat used *SamplingStrata* for designing its 2018 *LUCAS* survey (Ballin *et al.*, 2018). In addition, the World Bank adopted *Sampling Strata* and embedded it in its *Survey Solutions Sampling Tools* integrated application.

Unlike other similar packages (as the package *stratification* Baillargeon and 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 and Barcaroli, 2013; Barcaroli, 2014; Ballin and Barcaroli, 2016.

In the following, we recall its fundamentals before illustrating the application to the SHIW sampling design. It is worth recalling that the optimal stratification proposed in this paper is only related to the second stage units (the households) of the overall sampling design, since the first stage units (the municipalities) are treated as fixed due to survey requirements related to organisational and fieldwork aspects.

As the aim of the optimisation performed through the genetic algorithm is to find a stratification that minimises the variance inside the strata with respect to all the survey target variables, an important step of the method is to estimate consistently the population variance in all the strata. 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).

5.1 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 sampling frame. 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 optimisation problem is said to be multivariate; otherwise it is univariate. For a given stratification, in the univariate case the optimisation 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 *Y*, their sampling variance is:

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

where:

$\hat{Y}_g = \sum_{h=1}^{H} \sum_{i=1}^{n_h} y_{g,i} w_i$: target estimate

H : number of strata

N_h : population in stratum h

n_h : sampling units in stratum h

$S_{h,g}^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{g,i} - \bar{y}_{g,h})^2$: sampling estimate of the variance of the *g-th* variable in stratum h

It should be noted that $Var(\hat{Y}_g)$ is the design-variance of the estimator of the population total for the g-th target variable when the sample design is stratified simple random sampling. A more general variance formula would be needed if the optimisation were to find an allocation of PSUs to their strata and families within PSUs to their strata. The general formula would include both a stratified component due to first-stage sampling and another stratified component due to second-stage sampling; component variance formulas are presented, for instance, in Cochran (1977, pp. 308-310), Hansen, Hurwitz, and Madow (1953, ch. 6 and 7), or Valliant, Dever, and Kreuter (2018, ch. 9).

If we introduce the following cost function:

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

where C_0 indicates a fixed cost (not dependent on the sample size) and C_h

represents the average cost of collecting and processing data for a sampling unit in stratum h, then the optimisation problem can be formalised in this way:

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

under the constraints

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

where the Vg (g=1,..,G) are the upper bounds for the expected sampling variance for $\hat{Y}_1,\ldots,\hat{Y}_G$.

Bethel (1989) suggested that the problem can be more easily solved by considering the following function of n_h:

$$x_h = \begin{cases} 1/n_h & \text{if } n_h \geq 1 \\ \infty & \text{otherwise} \end{cases}$$

Using x_h the cost function can be written as

$$C(x_1,..,x_H) = C_0 + \sum_{h=1}^{H} \frac{C_h}{x_h}$$

and the variances as

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

Consequently, the multivariate allocation problem can be defined as the search for the minimum (with respect to x_h) of the convex function under a set of linear constraints

$$\sum_{h=1}^{H} N_h^2 S_{h,g}^2 x_h - N_h S_{h,g}^2 \leq V_g \quad g = 1,\ldots,G$$

A numerical optimisation algorithm, that is proved to converge to the solution (if it exists), was provided by Bethel by applying the Lagrangian multipliers method to this problem.

It should be noted that there are also other algorithms for solving nonlinear programming problems in sample allocation, for instance, the *proc optmodel* in SAS offers alternatives like the trust region method, Newton-Raphson method with line search, conjugate gradient method, and a quasi-Newton method; the R packages *alabama* and *nloptr* use the augmented Lagrangian

algorithm and the method of moving asymptotes, respectively. In this paper, we focus on the use of Bethel's algorithm and genetic algorithm, through *SamplingStrata*, which allows performing the optimisation steps in two different ways, depending on the nature of the stratification variables Xs.

5.2 Optimisation 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 average cost C of allocating a sampling unit in the stratum.

Starting from the initial atomic stratification, it is possible to generate, by differently aggregating the atomic strata, all the combinations that belong to the universe of stratifications. The number of possible different stratifications is exponential with respect to the number of the atomic strata. In concrete cases, it is therefore impossible to examine all the different possible alternative stratifications in order to individuate the best, *i.e.* the one of minimal cost. The genetic algorithm allows to explore the universe of stratifications in an efficient way, thus finding a solution not far from the optimal, by performing the following steps:

1. an initial set (*generation* in the terminology of the genetic algorithm) of stratifications (*individuals*) is randomly generated by aggregating the atomic strata: a given *individual* is a stratification where each atomic stratum is randomly attributed to one aggregate stratum identified by a combination of values of the stratification variables; each generated individual is characterised by a *genome*, *i.e.* a vector of integer numbers (*chromosomes*) indicating for each atomic stratum to which aggregate stratum it belongs;

2. for each aggregate stratum the information required (population, means and standard deviations of Ys, cost) is calculated and its *fitness* (total cost of the sample required to satisfy precision constraints) is determined by applying the Bethel algorithm;
3. the next set of individuals is generated by applying the usual operators of the genetic algorithm, *i.e. mutation*, *selection* and *crossover*.

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.

To clarify the above, let us consider a very simple example: a sampling frame with two stratification variables X1 and X2, and related domains respectively ("A", "B") and ("1", "2", "3"). Considering the Cartesian product of the two domains, there will be six atomic strata: a1=("A","1"), a2=("A","2"), a3=("A","3"), a4=("B","1"), a5=("B","2"), a6=("B","3").

The initial step consists in randomly generating, say, 20 individuals (first generation), each one characterised by a genome.

For instance, the first individual could be I1=(1,3,2,2,1,3), that is a stratification characterised by three aggregated strata: according to the position of the elements in the genome, the first stratum is the aggregation of a1 and a5, the second stratum is the aggregation of a3 and a4, the third stratum is the aggregation of a2 and a6.

For each one of these 20 stratifications, the corresponding fitness is calculated (applying the Bethel algorithm), as the cost of the sample necessary to be compliant with the precision constraints. The one with the best fitness (minimum cost) is retained as the optimal solution.

In order to generate the next generation of individuals, the 20 individuals are ordered by their fitness: if we set the *elitism rate* to 20%, the best 4 individuals will be retained as they are, with no change. Then, the remaining 80% individuals (16) will be generated in this way:

1. First, to each individual will be applied a *mutation* operator. Consider again the individual I1. If we set the mutation chance equal to 5%, we scan the elements of its genome, each time generating a random number between 0 and 1: if it is less than 0.05, the element is changed

by assigning a random number, otherwise it is left unchanged. Suppose a mutation happens for the third element, that is changed from 2 to 1: now the genome of I1 is (1,3,1,2,1,3).

2. From the 16 elements, 16 couples are selected, with a selection probability proportional to their fitness (*selection* operator).
3. From each couple, a new individual is generated by applying the *crossover* operator. Consider a selected couple I1=(1,3,1,2,1,3) and I5=(3,2,1,1,1,1). A number p is generated between 1 and the number of elements in the genome (6), for instance 2: the genome of the new individual will be given by the first 2 elements of I1 and the last 4 elements of I5, that is (1,3,1,1,1,1).

The new generation, composed by the 4 best individuals from the previous, and the new 16 obtained by mutation, selection and crossover, is now available. For each new individual will be calculated its fitness; if one of them has a better fitness than the current optimal solution, it will replace it.

The process continues until reaching the desired number of iterations.

5.3 Optimisation with continuous stratification variables

When all the stratification variables are continuous (or even categorical, but of the ordinal type), a variant of the above optimisation 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 categorisation of the variable;
- aggregate strata are consequently determined by cross-classifying units in the sampling frame according to their values belonging to the segments previously defined.

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

5.4 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-employment 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 optimised solution, it is possible to carry out the optimisation 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 and 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 $\epsilon \sim N(0, \sigma 2X\gamma)$ (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 optimisation function of *SamplingStrata*. The optimisation 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, since the PSU and the panel households are considered fixed.

As already mentioned, register data use different concepts and definitions from the survey and they have also 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.

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

Target variable	R2	Beta	Sigma
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

Source: Authors' own processing, 2018

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".

Using these models, we assign to each unit in the sampling frame the predicted values for each one of the variables of interest.

One may ask why we use data from the refresh sample for the 2016 wave, linked to Tax Register, only to fit the models, and we do not directly use it

to estimate the means, stratum variances, and the other quantities needed in the optimisation step. The answer is that this is necessary for two reasons: because the optimisation step with continuous stratification variables requires that their values are available for each unit in the sampling frame; and because, optimal strata values must be assigned to each unit in the frame when we select the final sample, and this can be done only knowing the values of the stratification variables.

As a second step, we chose the precision constraints in terms of the maximum expected coefficient of variation for the target mean estimates in the different domains (NUTS1 level Italian territorial units). The precision constraints are set equal to 5% in every domain and for all estimates.

We then run the optimisation step to define the stratification, the sample size, and its allocation. We use the sampling frame described in Section 4, containing 12,351,950 units (households). After removing the households with a source of income above 1 million euros (for the operational reasons previously explained), the resulting final population size is 12,334,342.

Numerous executions of this step have been attempted, varying the kind of optimisation (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 categorisation (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 favour of the latter.

Another important decision is to fix the number of optimised strata to be expected in each one of the 5 territorial domains (NUTS1). This parameter is quite important in terms of the final results of the optimisation: in general, increasing the desired number of final strata determines a decrease of the sample size necessary to be compliant with the precision constraints, until a certain point, from which on, this number increases. Hints on which this point could be are given by using a particular function available in *SamplingStrata*, which performs a sequential application of a k-means algorithm, varying the number of the clusters (in this case coincident with the number of final strata) from a minimum (usually 2) to a maximum, for instance 20. The indication was to set this value to 10.

Another important parameter is the minimum number of units per stratum: too low, and the risk in case of high non-response is to have strata without respondents; too high, and the constraint may have a negative impact on the optimality of the solution. In our case, it was set to 50 households.

The optimisation 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.

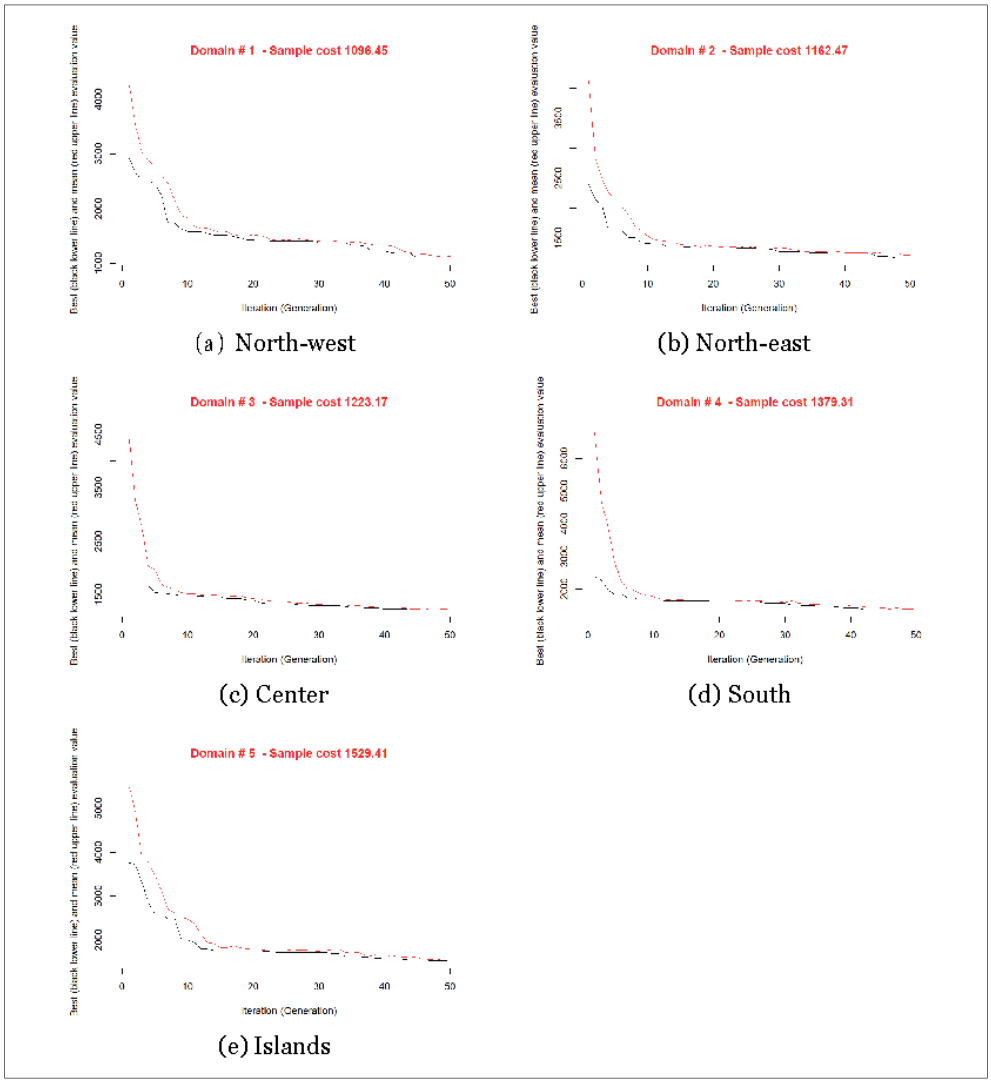
Figure 1 contains a graphical representation of the search for the optimal solution in the different domains. Each plot in this Figure can be interpreted in this way:

- in the x-axis are reported the different iterations (from 1 to 50): in correspondence to a given iteration, a set of 20 individuals have been generated, for each of them the Bethel solution in terms of sample size has been calculated;
- in the y-axis is reported the cost of the solution (in our case, the sample size);
- the red line represents the mean of the 20 Bethel solutions for each generation;
- the black line represents the cost of the best solution found so far.

Analysing these plots, a common situation for the different domains can be found: there is a smooth convergence towards the final solution of both the red and the black lines, and, more important, the lines towards the end are almost parallel to the x-axis, thus implying that adding more iterations should not increase substantially the optimality of the solution.

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

Figure 1 - Optimisation in the different domains



Source: Authors' own processing, 2018

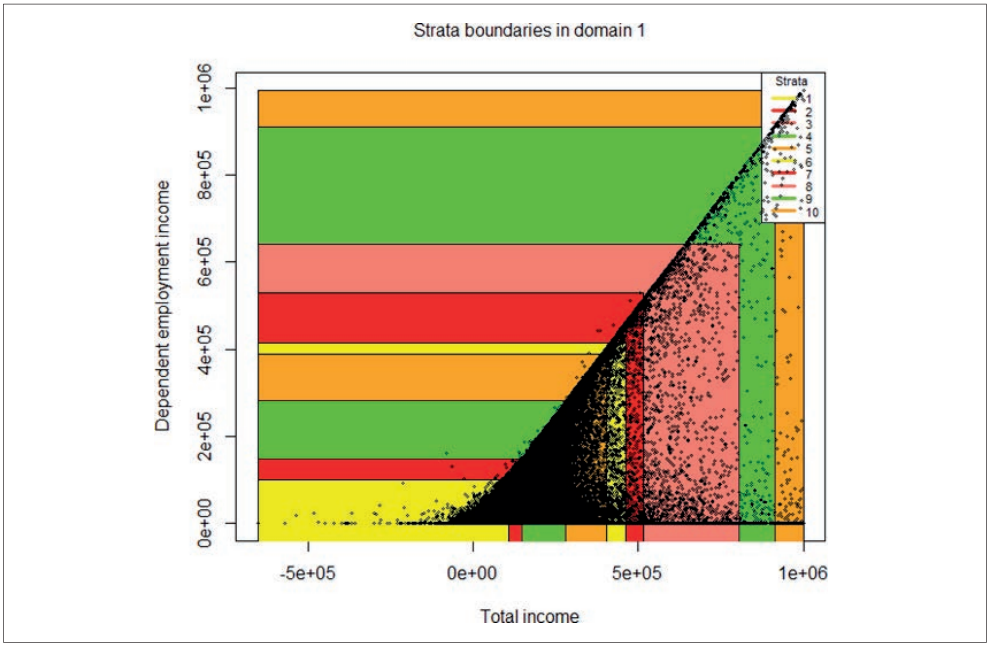
The package allows visualising in a two-dimensional graph the obtained strata, each time choosing a couple of variables. For instance, Figure 2 shows the characterisation of the strata in the first domain, by considering "Total income" and "Dependent employment income". The points in the plot represent households in the sampling frame. Colours identify the different strata.

In Figure 3, optimised strata with population, sampling allocation, and sampling rates are reported together with the range of the two stratification variables. The intensity of the green is proportional to the values of Population and Allocation in strata, while the length of the red bar is proportional to the sampling rate.

Considering the two figures together, we can better understand this graphical representation. For instance, the first stratum (the yellow one) at the bottom left of the plot in Figure 2 includes all the households whose Total Income is less than -105,895 and Dependent Employment Income is less than 99,369. The second stratum (the red one) includes all the households whose Total Income is in the interval (-282,649; -147,433) and Dependent Employment Income is less than 162,801. The same interpretation for the other strata.

We do not report all the possible combinations of couples of variables because of their number (11 per each domain), we just report this one to show how strata appear, in their characteristic "7-shaped" format.

Figure 2 - Strata resulting from the execution of the genetic algorithm (North-west, by *Dependent employment income* and *Total income*)



Source: Authors' own processing, 2018

Figure 3 - Strata population, allocation and range of stratification variables (North-west)

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

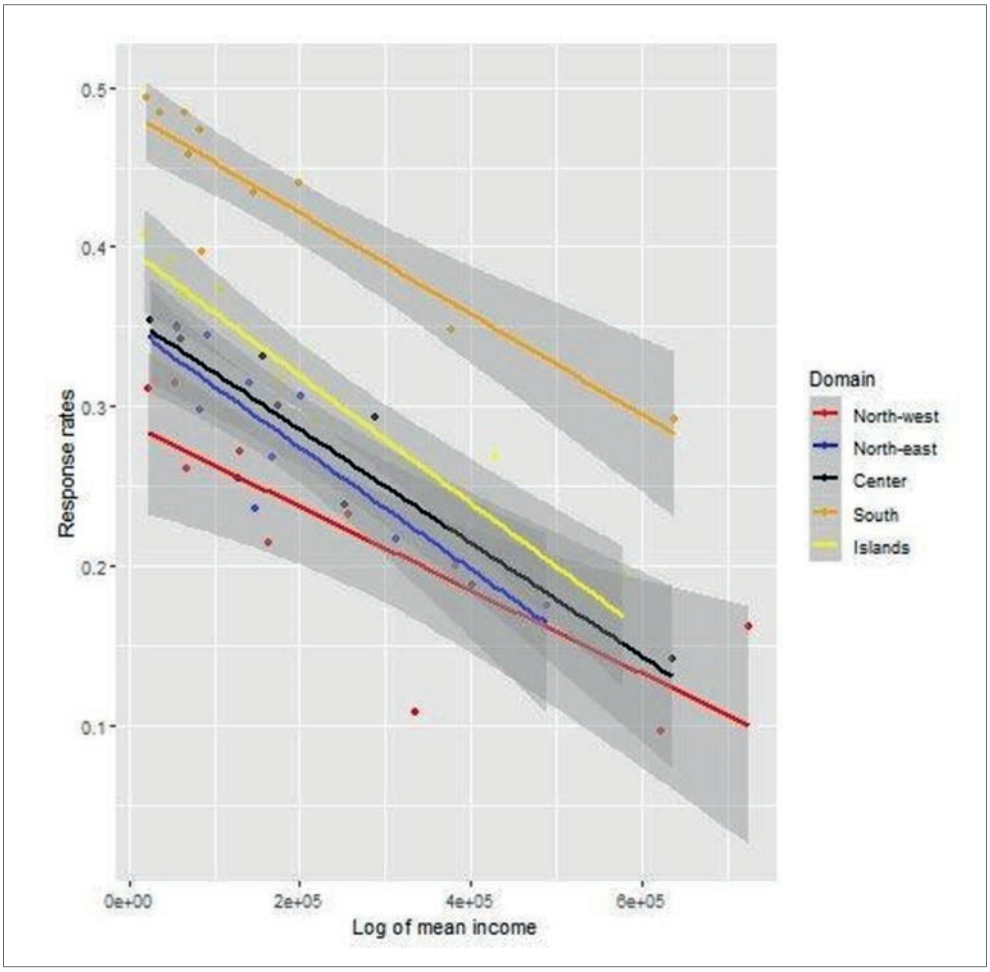
Source: Authors' own processing, 2018

The solution is characterised 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.

In order to take the expected non-response rate into account, 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 non-respondents 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. In Figure 4 we also report confidence bands around the lines, based on model standard errors.

The sample of units to be interviewed has been redefined by taking into account the propensity to non-response 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 optimised strata taking into account the initial allocation and the average propensity to the response in strata.

Figure 4 - Response rate and mean income in strata



Source: Authors' own processing, 2018

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	Initial Allocation	Final allocation	Sampling rate
1	1	2,384,770	365	1064	0.000446
1	2	523,170	272	784	0.001499
1	3	59,346	55	192	0.003235
1	4	35,528	50	211	0.005939
1	5	58,824	105	350	0.005950
1	6	3,846	50	195	0.050702
1	7	2,203	50	420	0.190649
1	8	5,028	50	226	0.044948
1	9	938	50	469	0.500000
1	10	1,137	50	280	0.246262

Source: Authors' own processing, 2018

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 on estimates of the mean (%)

Domain	Total income	Dependent emp. income	Self-employment 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. Centre	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

Source: Authors' own processing, 2018

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.

Of course, none of these assumptions hold in reality. In particular, assuming (1) and (3) leads to an under-estimation of the real values of the coefficients of variation, while the (2) might over-estimate them. For this reason, in the next Section we present the results of a simulation exercise that takes this issue into account.

7. Evaluation of the new sample design

In this Section, 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 in the 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 non-response;
3. sample size has been inflated to 17,608 households to take into account the expected non-response.

7.1 Simulations using Population Register for calibration

The first simulation consists of the following steps.

First, we use the models introduced in Section 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) are selected from the sampling frame. For each household, we simulate the non-response mechanism using the model described in Section 6. The decision to participate is then taken by drawing a value from a Bernoulli variable with the probability of success (the propensity to respond) equal to the propensity estimated by the non-response model.

For each sample of respondents, initial weights are computed considering the probabilities of inclusion of both first and second stage, and the final weights are obtained by calibrating using the total number of households in the strata in the Population Register, as defined by the new design.

In the end, for each target estimate (means of total income and of the four components), 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 samples.

Table 4 - Estimated coefficients of variation of the new sample design (%)

Domain	Total income	Dependent emp. income	Self-employment 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. Centre	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

Source: Authors' own processing, 2018

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. 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.

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

Domain	Total income	Dependent emp. income	Self-employment income	Pension income	Rents
1. North-west	-3.8	-12.8	5.6	8.3	-1.5
2. North-east	-2.6	-8.8	3.0	6.4	-2.0
3. Centre	-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

Source: Authors' own processing, 2018

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 number 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 6.

Table 6 - Expected coefficients of variation for the old sample design (%)

Domain	Total income	Dependent emp. income	Self-employment 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. Centre	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

Source: Authors' own processing, 2018

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 ones for two main reasons. First, the sample size is larger (about 8,000 households for each wave). Second, the sampling weights are calibrated in a way 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-employment 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. Centre	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

Source: Authors' own processing, 2018

Table 8 - Coefficients of variation estimated in the 2014 Italian HFCS wave (%)

Domain	Total income	Dependent emp. income	Self-employment 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. Centre	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

Source: Authors' own processing, 2018

Following the same approach previously used, 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 non-response has been simulated accordingly to the predicted non-response 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 by strata in the Population Register as defined in the old design.

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. income	Self-employment income	Pension income	Rents
1. North-west (Rip1)	6.18	9.50	31.73	10.34	6.61
2. North-east (Rip2)	4.96	8.17	25.77	9.34	5.48
3. Centre (Rip3)	5.51	8.68	22.22	10.00	5.68
4. South (Rip4)	4.85	7.54	19.60	8.18	5.30
5. Islands (Rip5)	7.42	11.33	29.30	14.26	7.79

Source: Authors' own processing, 2018

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

Domain	Total income	Dependent emp. income	Self-employment income	Pension income	Rents
1. North-west (Rip1)	-5.92	-14.55	-6.78	8.12	-2.71
2. North-east (Rip2)	-4.20	-10.15	-6.26	5.97	-2.03
3. Centre (Rip3)	-4.60	-11.39	-7.39	7.14	-2.55
4. South (Rip4)	-2.85	-7.19	-3.8	4.03	-2.12
5. Islands (Rip5)	-3.18	-8.36	-4.40	4.71	-2.22

Source: Authors' own processing, 2018

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

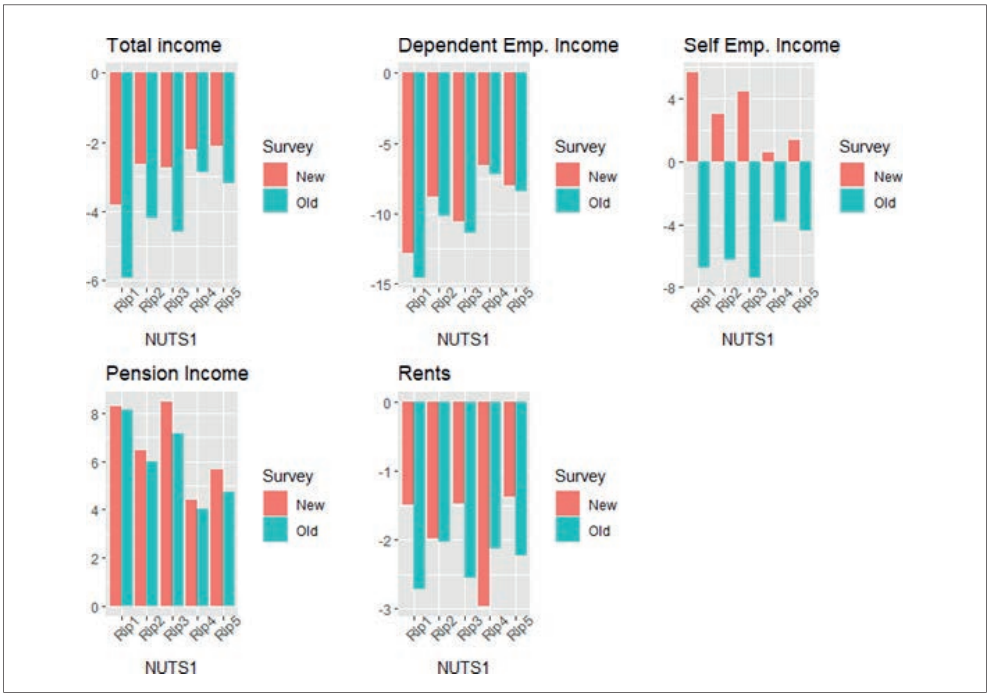
Source: Authors' own processing, 2018

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 the Mean Squared Error (MSE).

As for the bias, the new sample design is still better, but there are 6 cases out of 25 in which the old design performs better.

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



Source: Authors' own processing, 2018

7.2 Simulations using Tax Register for calibration

In the previous simulations, we did 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 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

Domain	Total income		Dependent emp. income		Self-employment income		Pension income		Rents	
	New	Old	New	Old	New	Old	New	Old	New	Old
1. North-west (Rip1)	0.54	0.99	0.23	0.41	1.93	3.10	0.53	1.06	3.34	6.17
2. North-east (Rip2)	0.46	0.72	0.21	0.39	1.80	3.26	0.48	0.72	2.73	3.70
3. Centre (Rip3)	0.51	0.76	0.24	0.31	2.02	3.42	0.53	0.79	2.70	4.20
4. South (Rip4)	0.55	0.71	0.24	0.48	2.16	2.89	0.56	0.84	2.89	3.80
5. Islands (Rip5)	0.52	1.21	0.24	0.63	2.14	4.33	0.54	1.16	2.61	5.76

Source: Authors' own processing, 2018

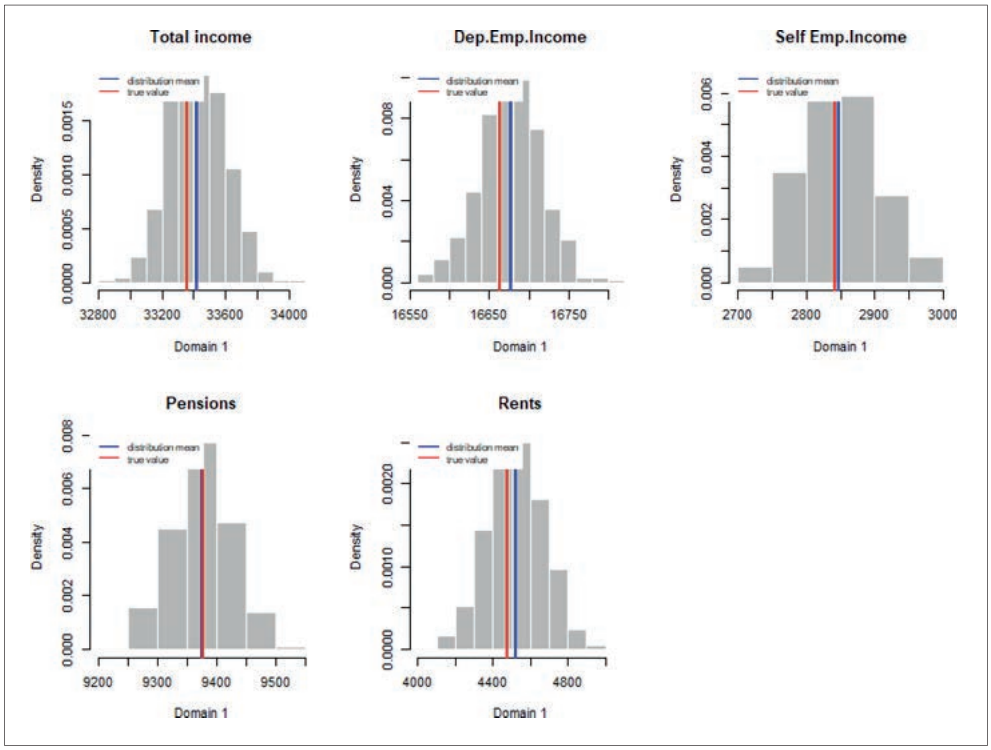
Table 12 - Estimated relative bias of the new and old sample designs (%) with calibration using Tax Register variables

Domain	Total income		Dependent emp. income		Self-employment income		Pension income		Rents	
	New	Old	New	Old	New	Old	New	Old	New	Old
1. North-west (Rip1)	0.19	0.32	0.08	0.16	0.20	0.63	-0.02	-0.28	1.00	1.93
2. North-east (Rip2)	0.00	-1.77	0.08	-2.09	0.17	-1.20	-0.04	-2.48	-0.26	0.44
3. Centre (Rip3)	0.22	0.59	0.08	0.00	-0.02	0.94	-0.01	0.17	1.25	3.21
4. South (Rip4)	-0.21	-1.43	0.02	-2.02	-0.57	-0.58	-0.07	-1.88	-0.99	0.78
5. Islands (Rip5)	0.09	0.44	0.05	-0.17	0.22	2.26	0.03	-0.53	0.27	3.06

Source: Authors' own processing, 2018

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

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



Source: Authors' own processing, 2018

There is an evident reduction of CVs and bias for both new and old sample design, with a comparison always in favour 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 Tax Register variables as explanatory variables. 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 also includes 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 also include these variables, results in terms of an overall reduction of MSE should be even greater.

References

Albarea, A., M. Bernasconi, A. Marenzi, and D. Rizzi (*eds.*). 2018. "Income under reporting and tax evasion in Italy. Estimates and distributive effects" *Documento di Valutazione*, N. 8. Roma: Senato della Repubblica, Ufficio Valutazione di Impatto/*Impact Assessment Office*.

Baillargeon, S., and L.-P. Rivest. 2012. "Univariate Stratification of Survey Populations". *R Package*, Version 2.2-3. The Comprehensive R Archive Network – CRAN.

Ballin, M., and G. Barcaroli. 2016. "Optimization of Stratified Sampling with the R Package SamplingStrata: Applications to Network Data". In Dehmer, M., Y. Shi, and F. Emmert-Streib. *Computational Network Analysis with R: Applications in Biology, Medicine, and Chemistry. Volume 7*. Hoboken, NJ, U.S.: John Wiley & Sons.

Ballin, M., and G. Barcaroli. 2013. "Joint determination of optimal stratification and sample allocation using genetic algorithm". *Survey Methodology*, Volume 39, N. 2: 369-393.

Ballin, M., G. Barcaroli, M. Masselli, and M. Scarnò. 2018. "Redesign sample for Land Use/Cover Area frame Survey (LUCAS) 2018". *Statistical Working Papers*, Eurostat. Luxembourg: Publications Office of the European Union.

Barcaroli, G. 2014. "SamplingStrata: An R Package for the Optimization of Stratified Sampling". *Journal of Statistical Software*, Volume 61, Issue 4: 1–24.

Barcaroli, G., M. Ballin, H. Odendaal, D. Pagliuca, E. Willighagen, and D. Zardetto. 2020. "SamplingStrata: Optimal Stratification of Sampling Frames for Multipurpose Sampling Surveys". *R Package*, Version 1.5-1. The Comprehensive R Archive Network – CRAN.

Bethel, J. 1989. "Sample allocation in multivariate surveys". *Survey Methodology*, Volume 15, N. 1: 47–57.

Bover, O. 2010. "Wealth Inequality And Household Structure: U.S. vs. Spain". *Review of Income and Wealth*, Volume 56, Issue 2: 259–290.

Bover, O., E. Coronado, and P. Velilla. 2014. "The Spanish Survey of Household Finances (EFF): Description and Methods of the 2011 Wave". *Documentos Ocasionales/Occasional Papers*, N. 1407. Madrid, Spain: *Banco de España*.

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

Chakraborty, R., I.K. Kavonius, S. Pérez-Duarte, and P. Vermeulen. 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 - JOS*, Volume 35, Issue 1: 31–65.

Cochran, W.G. 1977. *Sampling Techniques. Third edition*. New York, NY, U.S.: John Wiley & Sons.

Colciago, A., A. Samarina, and J. de Haan. 2019. "Central Bank Policies and Income and Wealth Inequality: A Survey". *Journal of Economic Surveys*, Volume 33, N. 4: 1199–1231.

D'Alessio, G., and A. Neri. 2015. "Income and wealth sample estimates consistent with macro aggregates: some experiments". *Questioni di Economia e Finanza, Occasional Papers*, N. 272. Roma, Italy: Banca d'Italia.

Dobbs, R., S. Lund, T. Koller, and A. Shwayder. 2013. "QE and ultra-low interest rates: Distributional effects and risks". *Discussion Paper*, McKinsey Global Institute. New York, NY, U.S.: McKinsey & Company.

Dossche, M., J. Slačálek and G. Wolswijk. 2021. "Monetary policy and inequality". In *ECB Economic Bulletin*, Issue 2/2021. Frankfurt am Main, Germany: European Central Bank – ECB.

Eckerstorfer, P., J. Halak, J. Kapeller, B. Schütz, F. Springholz, and R. Wildauer. 2016. "Correcting for the Missing Rich: An Application to Wealth Survey Data". *The Review of Income and Wealth*, Volume 62, Issue 4: 605–627.

Eurosystem Household Finance and Consumption Network. 2009. "Survey Data on Household Finance and Consumption. Research Summary and Policy Use". *Occasional Paper Series*, N. 100. Frankfurt am Main, Germany: European Central Bank – ECB.

Fiorio, C.V., and F. D'Amuri. 2006. "Tax Evasion In Italy: An Analysis Using A Tax-Benefit Microsimulation Model". *The IUP Journal of Public Finance*, Volume IV, Issue 2: 19–37.

Hansen, M.H., W.N. Hurwitz, and W.G. Madow. 1953. *Sample Survey Methods and Theory: Volumes I-II.* Hoboken, NJ, U.S.: John Wiley & Sons.

Horgan, J.M. 2006. "Stratification of Skewed Populations: A review". *International Statistical Review*, Volume 74, N. 1: 67–76.

Household Finance and Consumption Network - HFCN. 2020. "The Household Finance and Consumption Survey: methodological report for the 2017 wave". *Statistics Paper Series*, N. 35. Frankfurt am Main, Germany: European Central Bank – ECB.

Jäntti, M., V.-M. Törmälehto, and E. Marlier (*eds.*). 2013. "The use of registers in the context of EU–SILC: challenges and opportunities". Eurostat, *Statistical working papers*. Luxembourg: Publications Office of the European Union.

Kareem, A.O., I.O. Oshungade, G.M. Oyeyemi, and A.O. Adejumo. 2015. "Moving Average Stratification Algorithm for Strata Boundary Determination in Skewed Populations". *CBN Journal of Applied Statistics*, Volume 6, N. 1: 205–217.

Kennickell, A.B. 2019. "The tail that wags: differences in effective right tail coverage and estimates of wealth inequality". *The Journal of Economic Inequality*, Volume 17, Issue 4, N. 1: 443-459.

Kennickell, A.B. 2017. "Modeling Wealth with Multiple Observations of Income: Redesign of the Sample for the 2001 Survey of Consumer Finances". *Statistical Journal of the IAOS*, Volume 33, Issue 1: 51-58.

Kennickell, A.B. 2008. "The Role of Over-sampling of the Wealthy in the Survey of Consumer Finances". In Bank for International Settlements (*ed.*). *The IFC's contribution to the 56th ISI Session*, Volume 28: 403-408. Lisbon, August 2007.

Khan, M.G.M., N. Nand, and N. Ahmad. 2008. "Determining the optimum strata boundary points using dynamic programming". *Survey Methodology*, 34, N. 2: 205–214.

Michelangeli, V., and C. Rampazzi. 2016. "Indicators of financial vulnerability: a household level study". *Questioni di Economia e Finanza, Occasional Papers*, N. 369. Roma, Italy: Banca d'Italia.

Neri, A., and M.G. Ranalli. 2011. "To Misreport or not to report? The Case of the Italian Survey on Household Income and Wealth". *Statistics in Transition - new series*, Volume 12, N. 2: 281–300.

Neri, A., and R. Zizza. 2010. "Income reporting behaviour in sample surveys". *Temi di discussion, Working papers*, N. 777. Roma, Italy: Banca d'Italia.

Schmitt, L.M. 2001. "Fundamental Study. Theory of genetic algorithms". *Theoretical Computer Science*, Volume 259, Issues 1-2: 1–61.

Valliant, R., J.A. Dever, and F. Kreuter. 2018. *Practical Tools for Designing and Weighting Survey Samples. Second Edition*. New York, NY, U.S.: Springer.

Vermeulen, P. 2018. "How fat is the top tail of the wealth distribution?". *The Review of Income and Wealth*, Volume 64, Issue 2: 357–387.