BANCA D'ITALIA

EUROSISTEMA

Questioni di Economia e Finanza

(Occasional Papers)

Improving survey information on household debt using granular credit databases

by Antonietta di Salvatore and Mirko Moscatelli

March 2024

Number 839

BANCA D'ITALIA
EUROSISTEMA

Questioni di Economia e Finanza

(Occasional Papers)

Improving survey information on household debt using granular credit databases

by Antonietta di Salvatore and Mirko Moscatelli

Number 839 – March 2024

The series Occasional Papers *presents studies and documents on issues pertaining to the institutional tasks of the Bank of Italy and the Eurosystem. The* Occasional Papers *appear alongside the* Working Papers *series which are specifically aimed at providing original contributions to economic research.*

The Occasional Papers *include studies conducted within the Bank of Italy, sometimes in cooperation with the Eurosystem or other institutions. The views expressed in the studies are those of the authors and do not involve the responsibility of the institutions to which they belong.*

The series is available online at www.bancaditalia.it .

ISSN 1972-6643 (online)

Designed by the Printing and Publishing Division of the Bank of Italy

IMPROVING SURVEY INFORMATION ON HOUSEHOLD DEBT USING GRANULAR CREDIT DATABASES

by Antonietta di Salvatore and Mirko Moscatelli*

Abstract

Distributional information on the debt held by households and on the characteristics of debtors is fundamental for creating and updating policy-relevant indicators and models. The primary source for this information in Italy is the Survey on Household Income and Wealth (SHIW), held periodically by the Bank of Italy. Its estimates, however, are affected by several types of non-sampling errors inevitably present in surveys. In this work, we use granular credit registers to improve debt estimates and determine which households are more likely to have measurement errors on their debts. The results show that integrating SHIW with information derived from the credit registers increases household debt participation and the amount of debt households hold. Moreover, we find that households belonging to the wealthiest quintiles of the population, residing in the South and Islands, and for which the reference person has low financial education are more likely not to report in SHIW their loans for property purchases.

JEL Classification: C83, D31, G51.
Keywords: survey, administrative data, residential mortgages, consumer credit.
DOI: 10.32057/0.QEF.2023.0839

* Bank of Italy – Directorate General for Economics, Statistics and Research

1 Introduction[1]

The financial crisis of 2008, the Covid-19 outbreak, and other recent changes in the economic environment have increased the demand for consistent distributional information for the household sector (ECB, 2020). Distributional indicators on the debt held by households and on the characteristics of the debtors, in particular, is fundamental to estimating the number of vulnerable households and the share of debt they have (Michelangeli and Pietrunti, 2014; Attinà et al, 2019), analyzing the evolution of inequalities, and understanding how policy measures affect borrowers based on their characteristics.

The primary source for this type of information in Italy is the Survey on Household Income and Wealth (SHIW) of the Bank of Italy that comprises, in the recent period, on average, about 7,000 households distributed over approximately 300 Italian municipalities (see for example Baffigi et al, 2017, and Gambacorta and Porreca, 2022). The survey contains information on household demographic characteristics (age, education, household composition, etc.) and its consumption, income, wealth, and liabilities. Concerning the latter item, households are asked to distinguish between loans for property purchases, consumer credit, and other debts. Since 2010, the survey has provided the data for Italy to the Household Finance and Consumption Survey, a harmonized dataset coordinated by the European Central Bank[2].

SHIW estimates, however, are affected by several types of non-sampling errors inevitably present in surveys. The two main factors are unit nonresponse (caused by some households refusing to participate in the survey) and measurement error, which includes non-reporting (caused by the fact that some participating households avoid reporting information) and misreporting (caused by the fact that some households fail to report the correct information). Di Salvatore et al (2022) showed that the decision of households to participate in the survey is not significantly correlated with debt ownership or with its amount. Less is known on the distorsions of estimates of household liabilities caused by measurement error. Households who respond to the survey may avoid reporting all or part of their debts for reluctance to declare them, e.g. because they are linked to properties that they do not want to declare or because they have problems paying them, or to shorten the interview, or they could have difficulty in retrieving correct information for poor memory or knowledge (even if the SHIW respondent is the most knowledgeable person in the household, he or she may not know the exact situation of all the other components, especially for small debts). In all these cases, estimates based on the answers of the households are different from the actual values. Previous studies showed that the effect of the errors is typically to underestimate the values; this is true not only for debt statistics but also, and often to a greater extent, for wealth and income statistics (D'Alessio and Faiella, 2002; Biancotti et al, 2008; Neri and Ranalli, 2012).

[1] The authors wish to thank David Loschiavo, Andrea Neri and Alfonso Rosolia for their valuable comments, and a special thanks to Francesco Vercelli for his insightful suggestions.

[2] For a general description of the survey, see https://www.ecb.europa.eu/stats/ecb_surveys/hfcs/html/index.en.html.

The aim of this work is to use granular records present in credit databases available to the Bank of Italy, namely the Italian Credit Register (CR) and a database on consumer credit belonging to Consorzio per la Tutela del Credito (CTC), to mitigate these errors in the 2020 edition of the SHIW, in order to obtain more accurate estimates of Italian households' debt participation and of the total amount of debt they hold, as well as of the values of specific loans for property purchases and of consumer credit.

The new estimates significantly increase both the share of households participating in the credit market and the amount of debt they hold with respect to the unadjusted SHIW estimates. Moreover, from an analysis of the households that have a loan for property purchase in CR but do not report it in SHIW, it emerges that non-reporting is significantly correlated with several household characteristics.

The new estimates can be used to obtain more accurate policy relevant indicators of financial vulnerability, inequality, and characteristics of debtors. Moreover, they can be used as new starting data for the construction of the Distributional Wealth Accounts (DWA), the ECB project that has the aim of comparing and bridging micro data from the Household Finance and Consumption Survey with macro data from National Accounts and Financial Accounts, and of developing quarterly distributional results for household macro balance sheets (Ahnert et al, 2020; ECB, 2020; Engel et al, 2022; Neri et al, forthcoming).

2 Data

In the paper, we use five sources: the SHIW survey; two granular credit databases (the Italian Credit Register and the CTC consumer credit dataset); the Financial Accounts; and the MFI balance sheet items.

The SHIW, our primary source, is a survey conducted by the Bank of Italy since 1960s[3]. The target population refers to private households that are officially living in Italy. People living in institutions (convents, hospitals, prisons, etc.) or those in the country illegally are out of the scope of the survey. The sample is drawn in two stages, with municipalities and households as, respectively, the primary and secondary sampling units. Starting from the 2020 wave, the one we use in this paper, the latter are stratified based on household income and indebtedness to improve the quality of the estimators produced for economic analysis. The sample in the most recent surveys consists of about 7,000 households (16,000 individuals) distributed over 300 Italian municipalities. The survey collects granular information on many topics, ranging from the household's and its members' socio-demographic characteristics to the different sources of income, to the household's assets and liabilities, to consumption and saving behaviors. Concerning their debts, households are asked to distinguish between loans for property purchases, consumer credit, and other debts.

[3] For a general description of the survey, see https://www.bancaditalia.it/statistiche/tematiche/indagini-famiglie-imprese/bilanci-famiglie/index.html?com.dotmarketing.htmlpage.language=1

The Italian Credit Register (CR) is an archive managed by the Bank of Italy that contains information on outstanding loans granted to borrowers in Italy by all financial intermediaries operating in the Italian territory. This source lists all loans from borrowers with a total debt with a reporting intermediary of at least 30,000 Euros. Intermediaries are required by law to report this information. Loans are distinguished into three classes: revolving credit lines, loans backed by account receivables and term loans. The archive also contains information on the type of collateral of the loan (real, personal, or none).

The CTC archive is a dataset on consumer credit which contains information on all consumer credit loans reported by intermediaries who are members of the Consorzio per la Tutela del Credito. The dataset accounts for about 61 percent of all consumer credit granted in Italy. The dataset is somehow complementary to the CR, as consumer credit is typically below the CR reporting threshold. Although the dataset has loan-level granularity, people's ids are encrypted, so it is impossible to find a borrower's identity and link this information with other databases.

The Financial Accounts are the primary source of aggregate information on the financial activity of the institutional sectors; they are compiled by the national central banks (and, for the Euro Area as a whole, by the ECB[4]). They provide the total debt held by households, but they do not allow to categorize it by the purpose of lending.

Finally, the MFI balance sheet items are aggregate data from supervisory reporting that monetary financial institutions make to central banks[5]. Regarding loans granted to households, the database distinguishes between credit for consumption (loans granted for mainly personal consumption of goods and services), loans for property purchases (loans given to invest in housing for own use or rental, including building and refurbishments, or for the purchase of land) and other loans (mainly for business).

3 Debt corrections

We use two different procedures to correct loans for property purchases and consumer credit, depending on the characteristics of the related credit registers. The information available in CR is unencrypted and has a 30,000 euro threshold. For this reason, we adopt exact matching at the individual level with CR and we consider the SHIW data as correct if the loan amount is under 30,000 euros and does not appear in CR. On the other hand, the information available in CTC is encrypted and accounts for about 61 percent of all consumer credit granted in Italy. Therefore, we will mainly use it to obtain a lower bound of the number of households with consumer credit, and impute consumer credit in a model-based way using SHIW information.

[4] https://www.ecb.europa.eu/stats/macroeconomic_and_sectoral/sector_accounts/html/index.en.html
[5] https://www.ecb.europa.eu/stats/money_credit_banking/mfi_balance_sheets/html/index.en.html

In sections 3.1 and 3.2 we describe in more detail the two procedures, while in section 3.3 we describe an optional calibration procedure that allows to equate the aggregate debt amounts with the macro values obtained from external sources.

3.1 Loans for property purchases correction

To correct the loan survey data for property purchases, we use the information available in CR, which contains all loans above the 30,000 euros threshold made by Italian intermediaries to households, irrespective of the underlying reason (see Section 2 for more details). It is possible to exactly match CR and SHIW at an individual level, therefore we follow this approach and then aggregate the information at the household level using the household composition available in SHIW.

Although there is no loans for property purchases category in CR, it can be approximated by term loans to consumer households with real collateral. The focus on consumer households excludes non-residential properties bought for business purposes. The total amount of this aggregate at the end of December 2020 is 379 billion euros, a value close to that of 451 billion euros derived by macro sources[6]; the missing amount is likely related to the CR threshold (as loans for property purchases below it do not appear in CR) and, to a lesser extent, to differences in the perimeters of the two aggregates.

The main issue of the imputation is the presence of the CR threshold. For this reason, we use the following rule: if the household has no loans for property purchases in CR but reports them in SHIW for an amount below 30,000 euros, we consider the SHIW value correct.

The matching allows for improving the quality of estimators based only on the survey data, correcting both for nonreporting errors, if a debt is found in CR and the household does not report it in SHIW, and for misreporting errors, when a different amount is found in CR for a debt reported in SHIW by the household (this can happen, for instance, when a household declares a loan but does not recall its exact residual amount as of December 2020).

3.2 Consumer credit correction

Contrary to CR, it is not possible to match at the individual level the information on consumer credit from CTC with SHIW records[7]. For this reason, we adopt a three-step estimation process consisting of:

1. estimation of the share of households participating in consumer credit;
2. development of a model that associates to each household, based on its characteristics, the probability of resorting to consumer credit. The model will be used to assign consumer credit to households that do not declare it but have a high estimated probability of having it until the overall participation share is equal to the one estimated in the previous point;

[6] The macro estimate of loans for property purchases is obtained by multiplying the total debt held by households (Financial Accounts) by the share of loans for property purchases on total household loans (MFI balance sheet statistics).

[7] Almost all consumer credit is below the CR threshold, and therefore does not appear in it.

3. assignment to each "new" household participating in consumer credit of the amount borrowed, based on its characteristics.

The final result of the process will be an adjusted value of the consumer credit borrowed by each SHIW household that will partially correct for non-reporting errors.

3.2.1 Estimating the share of households with consumer credit

The first step of the consumer credit correction is to estimate the share of households having consumer credit. To this end we use the CTC database, which contains loan-level information for a sample of financial intermediaries representing 61 percent of all consumer credit granted in Italy.

First, we want an estimate of the number N_p of people with consumer credit. We have two main options. The first option is to consider only the borrowers in the CTC dataset, obtaining a conservative estimate (lower bound) of the total number of consumer credit borrowers, which underestimates the true number but that relies on very few hypotheses. The second option is to try to estimate the number of consumer credit borrowers not present in the CTC dataset on the basis of the information available, such as the share of consumer credit covered by the CTC dataset and the share of CTC borrowers that have consumer credit with multiple banks.

For the following we will use the first option, which, although it likely underestimates the number of total consumer credit borrowers, is more conservative and does not make assumptions about the distribution of consumer credit related to intermediaries for which no information is available. This gives us the count $N_p = 8{,}996{,}176$. As a robustness check, we describe the second option and the relative estimates in Appendix 2.

Second, we need to go from the number of people with consumer credit to the number of households with consumer credit. From the SHIW data, we can estimate the share of households by the number of income earners of the household and by the number of consumer credit loans that the household has taken. The results are reported in Table 1.

Table 1. Share of income earners and loans for consumptions for households with consumer credit

	1 loan for consumption	2 loans for consumption	3 loans for consumption	4 loans for consumption or more	**Subtotal**
1 income earner	28.4%	4.4%	0.6%	0.0%	**33.4%**
2 income earners	40.0%	6.3%	1.9%	0.9%	**49.1%**
3 income earners	12.0%	1.9%	0.1%	0.1%	**14.1%**
4 income earners or more	2.4%	0.8%	0.1%	0.0%	**3.3%**
Subtotal	**82.8%**	**13.4%**	**2.7%**	**1.0%**	**100.0%**

Continuing with the conservative approach, we assign each consumer credit to a different income earner, i.e. assuming that an household with i credits for consumption and j income earners has $\min(i,j)$ people with consumer credit (as an example, for households that have three income earners and two consumer credits we assume the number of members with consumer credit to be two). It is important to notice that most households have either a single income earner or a single credit for consumption: for these households, the number of people with consumer credit is one.

Let $\mathrm{s(i,j)}$ be the share of households with i credits for consumption and j income earners, and R be the ratio between the total number of people with consumer credit and the total number of households with consumer credit. Then, we can estimate R as

$$R = \sum_{1 \le i,j \le 4} s(i,j) \cdot \min(i,j) = 1.12$$

This allows us to obtain an estimate of the number N_h of households with consumer credit as

$$N_h = \frac{N_p}{R} = 8{,}032{,}300.$$

Finally, we derive the share of households with consumer credit by dividing the previous estimate by the total number of households:

$$S = \frac{N_h}{N} = 31.7\%.$$

3.2.2 Modeling consumer credit participation

In the previous section we estimated the household participation share in consumer credit and saw that it is significantly higher than the SHIW estimate, 31.7 percent vs 14.3 percent respectively. However, we don't know which households to attribute the difference to.

For this reason in the current section we develop a model which, using SHIW information, estimates the probability that an household has consumer credit based on its characteristics. The idea is to look for households without consumer credit with characteristics similar to those who declare to have it, and assign the ownership of consumer credit to them.

Our model estimates the relationship

$$Y = f(X, \varepsilon)$$

where $Y \in \{0,1\}$ is a boolean variable indicating whether the household has consumer credit, X is a vector of characteristics of the household, and ε is the irreducible error. We select the variables using previous studies that predict participation in consumer credit or in the debt market in general, such as Neri and Ranalli (2012) and Attinà et al. (2019)[8]. The final list includes 15 variables that refer to households income, wealth and consumption, including recent purchases of some categories of durable goods, as well as the socio-demographic characteristics of the main income earner (see Table A1 in the Appendix).

As a predictive model we use the random forest (Breiman, 2001), a machine learning model based on the aggregation of decision trees that has been shown in several works to obtain more accurate predictions than traditional methods such as the logistic regression and the linear discriminant analysis (Muchlinski et al, 2016; Couronné et al, 2018; Moscatelli et al, 2020; Alonso and Carbo, 2020). In addition to the advantage in terms of accuracy, the use of the random forest allows us to capture automatically and in a data-driven way nonlinearities and interactions between variables, and to treat as efficiently as possible ordered categorical variables present in the survey such as the level of education.

The final model has an out-of-sample AUC of 73.4 percent (calculated using an 80:20 train/test split of the dataset), showing a good ability to distinguish households that take consumer credit from those that do not. Figure A1 in the appendix shows the importance of each variable for the model, defined as the decrease in out-of-sample AUC that the model would experience if the variable were randomly permuted across the test dataset (Molnar, 2020; Cascarino et al, 2022). The most important variables for the model are whether the household owns a mortgage, the occupation of the main income earner, the household annual income, the household financial assets and the household annual consumption.

[8] In the first paper, SHIW household characteristics are used to estimate the probability of owning financial liabilities with actual possession inferred from an external reliable source, while, in the second paper, the probability of owning consumer credit in the current survey is estimated, for households belonging to the panel component of the SHIW, using the SHIW characteristics of households in the previous survey.

Using the probabilities generated by the model we assign participation to consumer credit to households that do not declare it but have an high probability of having it, until we reach the overall participation share estimated in the previous section.

3.2.3 Attributing the amount of consumer credit borrowed

As a last step, we attribute, to each household for which participation in the consumer credit market has been assigned, the amount of consumer credit it holds.

After comparing several attribution methods, namely a random forest model, a logit model, and means and medians of the amount held by groups with the same characteristics, we decide to attribute them the average amount held by households in the same quintile of income. The choice made is mainly due to the fact that the attribution via averages is less volatile than the attributions by model, avoiding to attribute particularly high consumer credit to households that do not report it. Following a conservative approach, we believe that mistakenly attributing a high debt to a household that does not actually have is worse than the opposite. Moreover, we believe it is more likely that a household forgets a credit for consumption, or considers less important to report it, if it is small in amount and weighs little on its budget.

3.3 Calibration

It is often useful for aggregate survey amounts to be consistent with macro data obtained from external sources. In this section, we propose a joint calibration that can be used to equate SHIW aggregate debt amounts with macro amounts derived from Financial Accounts and MFI balance sheet statistics. This can be beneficial because Financial Accounts data is reliable and highly comparable with SHIW debt information (Ahnert et al, 2020).

We follow the generalized raking approach of Deville et al (1993). Let w the set of SHIW weights and x the (multivariate) set of variables of interest. In our case, x will include loans for property purchases held by each quintile of the debtors[9], consumer credit, and other debts. We want to determine a set of calibration coefficients $\{c_1, \dots, c_n\}$ for each household such that:

1. $\sum_{k=1}^{n} w_k x_k c_k = \sum_{j=1}^{N} x_j$

2. $\{c_1, \dots, c_n\}$ are "as close al possibile" to one.

The first condition says that, after the calibration, the SHIW aggregate amounts must be equal to the macro amounts for all the variables of interest. The second condition says that, among all the possible sets of values $\{c_1, \dots, c_n\}$ that satisfy the first condition, we choose the set that minimizes a given distance

[9] While the total amount of loans for property purchases is obtained from Financial Accounts and MFI balance sheet statistics, the share held by each quintile is computed using the Italian Credit Register.

function centered in one to avoid changing too much the actual sample values. As distance function we choose the logit (0.1, 10) method, which has the advantage, unlike other methodologies, of not generating negative calibration coefficients and of not giving excessively high coefficients to one or to a small number of households.

The obtained calibration coefficients gives us a new amount of debt held by each household as $x'_k := x_k c_k$, which not only makes the total SHIW aggregates equal to the amounts derived from macro data, but that also generates a concentration of debt more consistent with the one actually derived from granular data, due to inclusion of the share held by each quintile computed using CR. This is another benefit of integrating sample data with credit register information.

4 Results

4.1 Household participation to the debt market and amount of debt held

In this section, we present the new estimates on the share of households participating in the debt market and the amount of debt they hold. We will present the estimates without the calibration procedure, since the calibration would make the survey aggregate amounts equal by definition to the macro amounts.

Table 2 shows household participation in the debt market according to the corrections adopted. The share of households with debt - which in our definition includes, in addition to loans for property purchases and consumer credit, also other debts - based on unadjusted SHIW data is 11 percent lower than the one obtained according to the new corrections (26.9 percent and 37.8 percent, respectively). The new estimate is also quite close to the 42.2 percent participation share provided by a report of the Central Risk Information Function (CRIF), a company that provides credit information and financial risk management and that owns the credit register with the arguably widest coverage in Italy[10]. The increase is mainly due to an underestimation of the share of households that own consumer credit, which increases from 14.3 to 31.7 percent, and to a lesser extent to an underestimation of the participation in the loans for property purchases market, which increases from 15.2 to 20.3 percent.

[10] https://www.crif.it/area-stampa/indebitamento-famiglia-italiane-2020-con-prudenza/. It is however important to highlight that, although conceptually similar, the two indicators refer to different populations since CRIF estimates the share of people with debt out of the total number of Italian adults.

Table 2. Share of household participating to the debt market

	Raw SHIW data	CR correction	CTC correction	Both corrections
Loans for property purchases	15.2%	20.3%	15.2%	20.3%
Consumer credit	14.3%	14.3%	31.7%	31.7%
Both loans for property purchases and consumer credit	*4.5%*	*5.6%*	*10.5%*	*15.3%*
Any type of debt	26.9%	30.7%	37.4%	37.8%

Table 3 shows the total amount of debt held by households. According to the new estimates, consumer credit increases from 27.6 to 65.1 billion, going from 21.9 to 51.6 percent of the Financial Accounts macro data[11]. Part of the residual difference is related to the inability to fully adjust misreporting and to the conservative choices made in the estimation of household participation to the consumer credit market. On the other hand, the attribution method does not seem to have a significant impact: attributing the amount of consumer credit using predictive models (instead of the average consumer credit held by households of the quintile of income to which the household belongs) generates very similar aggregate amounts.

As to the loans for property purchases, the corrections bring the total debt from 428 to 661 billion, exceeding the macro estimate of 451 billion. This is mainly due to the outstanding debt of the one-shot oversampling of indebted households in the 2020 edition of the SHIW[12], which made the total debt for loans for house purchases in the raw SHIW data very close to the macro estimate, instead of quite lower as it should be due to nonreporting (conversely, the effect of the oversampling on the new estimates of debt ownership presented before is minimal[13]). Due to the closeness of the two aggregates, and to the fact that the weighting system used in the survey does not take into account the underreporting of debt amounts as it does not include administrative data as a benchmark in the final calibration stage (for a description of the weighting process see Gambacorta and Porreca, 2022), when credit register data are used to correct for nonreporting the total outstanding debt increases greatly[14]. As a takeaway, while the

[11] Since the Financial Accounts do not distinguish household debts in consumer credit, loans for property purchases and other loans, the amount is obtained by re-proportioning the total debt obtained from the Financial Accounts with the relative shares obtained from the MFI balance sheet statistics.

[12] If we exclude the oversample of indebted households - equal to around 15 percent of SHIW households - and recalibrate the weights, the total debt for loans for property purchases goes from 661 to 371 billion.

[13] Removing the oversampled households decreases the share of indebted households by less than two percentage points. This is mainly linked to the fact that the bulk of the share of indebted households comes from consumer credit, which is estimated externally.

[14] That is, weights cannot correct for the additional measurement error introduced by the one-shot oversampling of indebted households.

raw SHIW data produce consistent and reliable results on the share and the distribution of debt across households based on reported data, integrating the survey with credit register data requires a further adjustment. If integrated data were to be used for analyses, we propose to deal with such an issue by a calibration approach that includes benchmark information of outstanding debt, as described in section 3.3.

Table 3. Amount of debt held by the households

	Raw SHIW data	CR correction	CTC correction	Both corrections	FA/BSI macro data
Loans for property purchases	428.5	661.1	428.5	661.1	451.1
Consumer credit	27.6	27.6	65.1	65.1	126.7
Total debt	551.0	783.7	588.6	821.3	738.0

4.2 Characteristics of non-reporting households

Finally, an interesting question is whether the non-reporting of debt, i.e. when a household has debts but does not declare them in SHIW, depends somehow on the characteristics of the household. To answer the question we focus on loans for property purchases, for which it is possible to make a matching at household level of CR and SHIW data and determine in which cases there are undeclared loans for property purchases.

Table 4 shows, for the subset of households that have a loan for property purchase in CR, the results of the logistic regression having as dependent variable the presence of a loan for property purchase in SHIW and, as covariates, several characteristics of the household. The model presents a good fit, with a McFadden's pseudo-R^2 of 0.144[15]. The probability of non-reporting a loan for a property purchase is significantly higher for households belonging to the wealthiest quintiles of the population, residing in the South and Islands, and for which the reference person has low financial education[16].

It is important to emphasize that the factors that increase the probability of non-reporting loans for property purchases (which we can estimate using CR data) are not necessarily the same that characterize the non-reporting of other types of loans such as low-amount consumer credit, for which some factors, such as the fact that the debtor is not the person responding to the survey, are likely to be way more important.

[15] Although there is no straightforward interpretation of the pseudo-R^2, as a common rule of thumb a model with a pseudo-R^2 between 0.2 and 0.4 is considered to have an "excellent fit".

[16] Other factors that influence non-response are age and employment status.

Table 4. Logistic model of the probability of non-reporting a loan for property purchase

Variable	Coeff.	Std. Err.	p-value	
Constant	-0.42	0.74	0.572	
35 <= Age <= 44	-0.57	0.32	0.073	*
45 <= Age <= 54	-0.33	0.31	0.287	
55 <= Age <= 64	0.18	0.31	0.560	
Age >= 65	0.74	0.36	0.040	**
Middle school	-0.24	0.54	0.660	
High school	-0.53	0.54	0.324	
University degree	-0.65	0.54	0.226	
Self-employed	0.57	0.16	0.000	***
Not employed	0.34	0.23	0.146	
2nd wealth quintile	0.57	0.26	0.031	**
3rd wealth quintile	1.08	0.26	0.000	***
4th wealth quintile	1.47	0.25	0.000	***
5th wealth quintile	1.10	0.26	0.000	***
5,000<=Municipality size<=20,000	-0.60	0.47	0.202	
20,000<=Municipality size<=40,000	0.06	0.41	0.877	
40,000<=Municipality size<=500,000	0.11	0.41	0.792	
Municipality size>=500,000	-0.21	0.41	0.608	
Center	0.25	0.16	0.123	
South and Islands	0.88	0.16	0.000	***
Proxy respondent	0.16	0.13	0.238	
Good financial education	-1.08	0.16	0.000	***

Observations: 1220.
Pseudo-R^2 (McFadden) = 0.144.

5 Conclusions

Several papers have shown that the discrepancy between survey data and Financial Accounts totals is primarily linked to the phenomena of non-reporting and misreporting, meaning that the households interviewed, for various reasons ranging from lack of will to poor knowledge or memory, do not report the debts held or report them with an incorrect amount. In this work, we use the information in two granular credit databases available to the Bank of Italy, the Italian Credit Register and the CTC consumer credit database, to partially correct these phenomena and obtain less biased estimates.

According to the new estimates, debt participation increases significantly and the total amount of debt held is greater, showing that the survey measurement error on debt appears to be non-negligible. Moreover, non-reporting does not appear to be random: households belonging to the wealthiest quintiles of the population, residing in the South and Islands, and for which the reference person has low financial education are more likely not to report a loan for property purchase they have.

There are some limitations to the approach. The two most notable are that loans reported in CR have a minimum threshold of 30,000 euros, thus loans for property purchases below that threshold cannot be corrected, and that the CTC database for consumer credit is encrypted; therefore identifying assumptions are needed to map the overall household participation share to the consumer credit market to the individual SHIW households. Nevertheless, in the trade-off between limitations and benefits, we argue that the latter outweighs the former.

The new estimates can be used, among other things, to obtain more accurate policy-relevant indicators (such as financial vulnerability, inequality, and characteristics of debtors) and as new starting data for constructing the Distributional Wealth Accounts.

There are several possible follow-ups of the work. The first one is to study the new estimates' impact on policy-relevant indicators such as those of financial vulnerability. Another one is to use the CR-SHIW matching to study the drivers of households' credit default and compare the performance of different models in predicting it. Yet another is, should granular data on consumer credit become available, to improve the estimates on the holding of consumer credit using individual matching instead of allocative models in a similar way to what has been done for the loans for property purchases.

References

Ahnert, H., Kavonius, I. K., Honkkila, J., and Sola, P. (2020). Understanding household wealth: linking macro and micro data to produce distributional financial accounts (No. 37). European Central Bank.

Alonso, A., and Carbo, J. M. (2020). Machine learning in credit risk: Measuring the dilemma between prediction and supervisory cost.

Attinà, C. A., Franceschi, F., and Michelangeli, V. (2019). Modelling households' financial vulnerability with consumer credit and mortgage renegotiations. Banca d'Italia.

Baffigi, A., Cannari, L., and D'Alessio, G. (2016). Cinquant'anni di indagini sui bilanci delle famiglie italiane: storia, metodi, prospettive. Bank of Italy Occasional Paper, (368).

Biancotti, C., D'Alessio, G., and Neri, A. (2008). Measurement error in the Bank of Italy's Survey of Household Income and Wealth. Review of Income and Wealth, 54(3), 466-493.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Cascarino, G., Moscatelli, M., and Parlapiano, F. (2022). Explainable Artificial Intelligence: interpreting default forecasting models based on Machine Learning. Bank of Italy Occasional Paper, (674).

Couronné, R., Probst, P., and Boulesteix, A. L. (2018). Random forest versus logistic regression: a large-scale benchmark experiment. BMC bioinformatics, 19(1), 1-14.

D'Alessio, G., and Faiella, I. (2002). Non-response behaviour in the Bank of Italy Survey of Household Income and Wealth (No. 462). Bank of Italy, Economic Research and International Relations Area.

Deville, J. C., Särndal, C. E., & Sautory, O. (1993). Generalized raking procedures in survey sampling. Journal of the American statistical Association, 88(423), 1013-1020.

Di Salvatore, Ilardi and Neri (2022), 'L'uso della Centrale dei rischi per migliorare la qualità delle stime del debito basate sull'Indagine sui bilanci delle famiglie italiane', Banca d'Italia, mimeo

ECB (2020). New experimental Distributional Wealth Accounts (DWA) for the household sector. Methodological note.

Engel, J., Riera, P. G., Grilli, J., and Sola, P. (2022). Developing reconciled quarterly distributional national wealth–insight into inequality and wealth structures.

Gambacorta, R., and Porreca, E. (2022). Bridging techniques in the redesign of the Italian Survey on Household Income and Wealth. Bank of Italy Occasional Paper, (719).

Michelangeli, V., and Pietrunti, M. (2014). A microsimulation model to evaluate Italian households' financial vulnerability. Bank of Italy Occasional Paper, (225).

Molnar, C. (2020). Interpretable machine learning. Lulu.com.

Moscatelli, M., Parlapiano, F., Narizzano, S., and Viggiano, G. (2020). Corporate default forecasting with machine learning. Expert Systems with Applications, 161, 113567.

Muchlinski, D., Siroky, D., He, J., and Kocher, M. (2016). Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. Political Analysis, 24(1), 87-103.

Neri, A., and Ranalli, M. G. (2012). To misreport or not to report? The measurement of household financial wealth. The Measurement of Household Financial Wealth (July 26, 2012). Bank of Italy Temi di Discussione (Working Paper) No, 870.

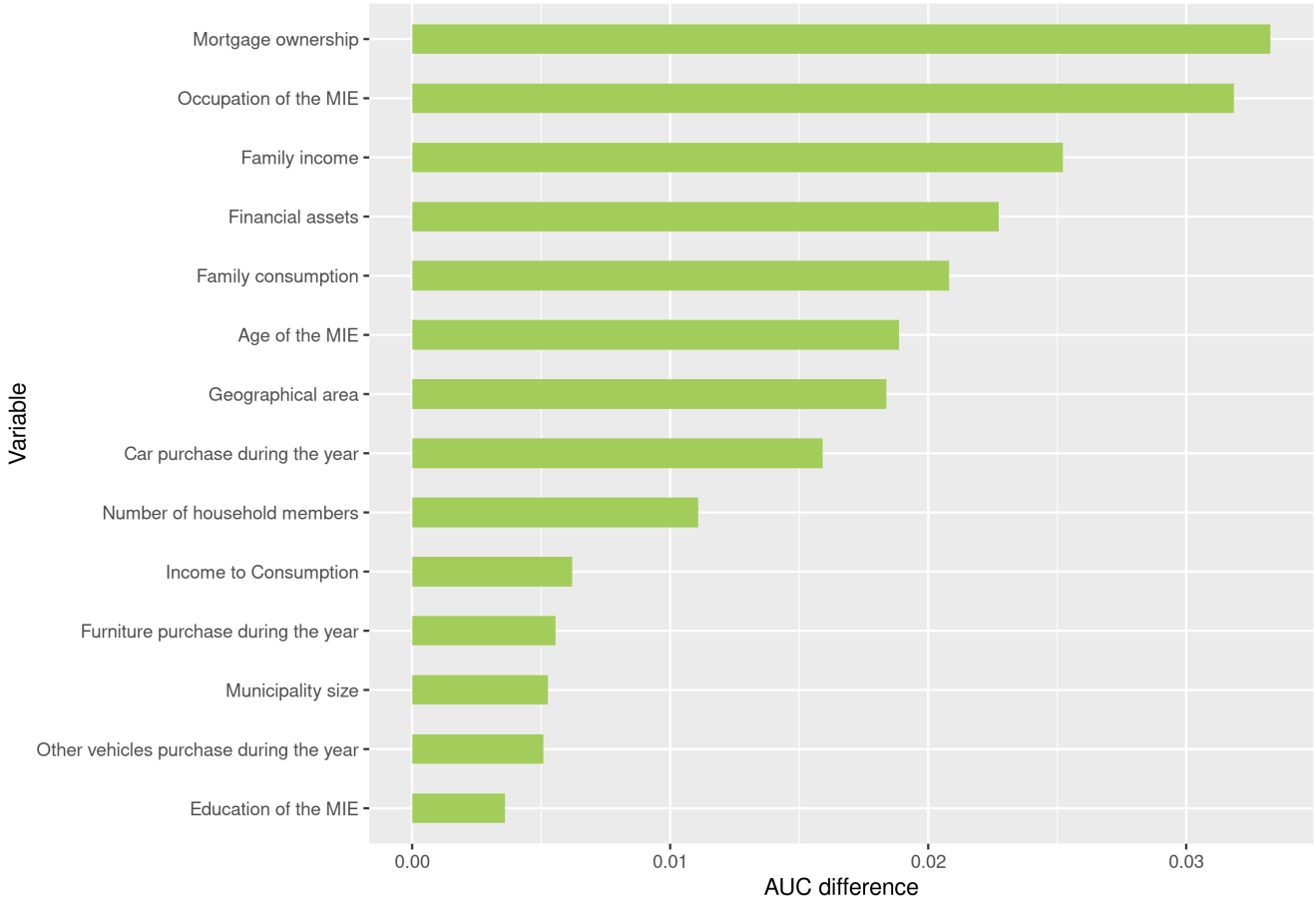
Neri, A., Spuri, M. and Vercelli, F. (forthcoming). The distributional accounts of Italian households.

Appendix 1 – Additional tables and figures

Table A1. Description of the variables of the random forest model.

Variable	Description
Consumer credit ownership	Dummy variable indicating whether the household owns consumer credit (target variable).
Family income	Annual household income.
Family consumption	Household consumption in the year.
Financial assets	Value of the financial assets held by the household.
Mortgage ownership	Dummy variable indicating whether the household owns a mortgage.
Number of household members	Number of household members.
Car purchase during the year	Dummy variable indicating whether the household purchased a car during the year.
Other vehicles purchase during the year	Dummy variable indicating whether the household purchased other vehicles during the year.
Furniture purchase during the year	Purchase of furniture, kitchen, appliances, computers or similar during the year.
Income to Consumption	Ratio between household income and household consumption in the year.
Municipality size	Size of the household municipality.
Geographical Area	Geographical region of the household.
Age of the MIE	Age of the main income earner.
Occupation of the MIE	Occupation of the main income earner.
Education of the MIE	Level of education of the main income earner.

Figure A1. Variable importance of the random forest model.



Appendix 2 – An alternative estimation of the number of people with consumer credit

In this section we describe an alternative estimation of the number N_p of people with consumer credit, which adds to the number of borrowers in the CTC dataset an estimate of the number of borrowers not already in it.

Let $N_{CTC} = 8{,}996{,}176$ be the number of borrowers in the CTC dataset, $s = 0.61$ the coverage ratio of CTC with respect to the total amount of consumer credit outstanding in Italy, and $d = 0.43$ an estimation of the share of borrowers of the intermediaries not participating in CTC that are also borrowers of intermediaries participating in CTC (that must, therefore, be excluded from the count to avoid duplications)[17].

If we assume the same CTC coverage ratio for the number of borrowers as for the amount granted, we can obtain an estimate of the number N_{other} of consumer credit borrowers not already present in the CTC dataset as

$$N_{other} = N_{CTC} \frac{1-s}{s} (1 - d),$$

equal to 5,751,653. In the equation above, $N_{CTC} \frac{1-s}{s}$ represents the number of borrowers of the intermediaries not participating in CTC, and $(1 - d)$ the share of non-duplicated ones.

Using this valure we can obtain a new estimate of the total number of borrowers with consumer credit as $N_p = N_{CTC} + N_{other}$, which is equal to 12,274,618 (about 36 percent larger than the previous estimate); this, in turn, would imply an estimate of the share of households with consumer credit equal to 43 percent.

[17] To estimate the share of borrowers that are already present in the CTC dataset we compute, for each intermediary of the CTC dataset, the share of its borrowers who have credit lines with other intermediaries in the dataset, and then calculate the weighted average of these estimates. The value obtained is equal to 43 percent, with the five largest intermediaries having values very close to it.