BANCA D'ITALIA

EUROSISTEMA

Questioni di Economia e Finanza

(Occasional Papers)

Combining survey and administrative data to estimate the distribution of household deposits

by Andrea Neri, Matteo Spuri and Francesco Vercelli

October 2023

Number 802

BANCA D'ITALIA
EUROSISTEMA

Questioni di Economia e Finanza

(Occasional Papers)

Combining survey and administrative data to estimate the distribution of household deposits

by Andrea Neri, Matteo Spuri and Francesco Vercelli

Number 802 – October 2023

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

COMBINING SURVEY AND ADMINISTRATIVE DATA TO ESTIMATE THE DISTRIBUTION OF HOUSEHOLD DEPOSITS

by Andrea Neri, Matteo Spuri and Francesco Vercelli *

Abstract

Several initiatives have combined survey data with macroeconomic aggregates from national accounts to produce more reliable and timely statistics on the distribution of household income and wealth. This paper builds on the methodology developed by the Expert Group on Distributional Financial Accounts, created by the European Central Bank, and proposes a novel approach for Italy to align survey-based total deposit estimates to the corresponding macroeconomic figures, by exploiting administrative records that are linked to the Survey on Household Income and Wealth and the aggregate information coming from supervisory data. The proposed method leads to a slight decrease in inequality of the deposit distribution compared to survey data and generally guarantees a closer match with aggregates from supervisory data compared to the ESCB approach.

JEL Classification: D14, D31, G51.
Keywords: micro-macro linkage, deposits distribution, wealth distribution.
DOI: 10.32057/0.QEF.2023.0802

Contents

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

1. Introduction[1]

In past years, several initiatives have combined survey data with macroeconomic aggregates from national accounts to produce more reliable and timely statistics on the distribution of household income and wealth. These statistics are commonly referred to as Distributional National Accounts.

In 2015, the European Central Bank established an expert group with the mandate to understand, quantify and explain the main differences between the Household Finance and Consumption Survey (HFCS) and the Financial Accounts (FA) and to develop distributional information on household wealth (EG-LMM, 2020). The work continued in 2019 thanks to the ECB Expert Group on Distributional Financial Accounts, which has fully implemented an estimation method to compile experimental quarterly results both for several European countries and the Euro area as a whole (see, for instance, Engel et al., 2022; Cantarella et al., 2023). The Distributional Wealth Accounts (DWA) will be published by the end of 2023. Depending on the availability of country-specific sources, each country may enrich the general method to improve the quality of the estimates. One of the preliminary and necessary steps to produce the DWA is to reconcile survey data and national accounts so that they produce coherent statistics on total household wealth. Surveys on household income and wealth commonly suffer from two quality issues, namely the difficulty in enrolling wealthy households and the reticence of respondents to report their incomes and assets honestly. Because of these issues, the coverage gap – i.e., the ratio of aggregates obtained from survey-based statistics and the corresponding macroeconomic figures from the national account balance sheet – is generally low. The gap between the two sources requires the development of a methodology to redistribute the missing wealth (i.e., the difference between national accounts and totals from survey data) among the households in the survey. Assumptions are necessary in the absence of reliable external information, such as administrative records. The DWA procedure developed by the European System of Central Banks (ESCB) includes some ad hoc adjustments on survey observations on deposits because this instrument represents a significant share of household gross wealth (more than one-third of financial assets), and its coverage ratio is low (below 50% for the Euro area). In the absence of external information, the ESCB adjustments on deposits are based on identifying outlier observations and their replacement with average values by income class. This paper proposes an alternative method drawing on additional information available for Italy. In

[1] The views expressed in this article are those of the authors alone and do not necessarily represent the position of Banca d'Italia or the Eurosystem. We thank Silvia Fabiani, Luigi Infante, Tullio Jappelli, and Alfonso Rosolia for their useful comments on this paper.

particular, we exploit the aggregate information coming from bank supervisory data and administrative records relating to individual fiscal income, housing wealth, and debt linked to the 2016 Survey on Household Income and Wealth (SHIW) conducted by Banca d'Italia, which is the Italian component of the HFCS.

The paper is structured as follows: Section 2 shows the data used in the analysis (SHIW, individual administrative registers, supervisory reports); Section 3 explains the methods used while Section 4 discusses the main results. Section 5 presents some robustness checks and finally, Section 6 concludes.

2. The data

2.1 The survey on household income and wealth

The SHIW is a survey conducted by the Banca d'Italia since 1965. The survey comprises a probabilistic sample of around 8,000 households selected from population registers. Its primary focus is collecting detailed information about household income, wealth, and, to a lesser extent, consumption expenditure. In particular, the survey collects the following information on the characteristics of the household and of its members (number of income earners, gender, age, education, job status, and dwelling type); income (wage and salaries, income from self-employment, pensions and other financial transfers, income from financial assets and real estates); consumption and saving (food consumption, expenses for housing, health, insurance, spending on durable goods, and household saving); wealth in terms of real estate, financial assets, liabilities. Data collection is entrusted to a specialized company using professional interviewers and CAPI methodology.

Since 2010, the survey has also been part of a project conducted by the European Central Bank to produce a harmonized survey on household finance and consumption in the Euro area (Household Finance and Consumption Survey, HFCS). Several studies have shown that these types of surveys suffer from errors, such as the under-representation of wealthy households in the sample and the reticence of respondents to provide correct information on issues generally perceived as highly sensitive. The analysis of measurement errors in the SHIW dates back to the seventies. Ulizzi (1970), describing the findings of the 1968 survey, observed that "Among the mentioned errors [non-sampling errors], special reference is due to those attributable to the reticence of respondents about the financial assets held. The experience gained in numerous analyses, some specific to the subject, has revealed considerable reluctance on the part of families to provide information on the ownership of financial assets (...). For savings and income, the collaboration of respondents is generally better,

being less the aversion to providing data on flows than on stocks".

Following this first analysis, many other studies have focused on measuring financial assets within the survey. D'Alessio et al. (1990) performed a statistical matching of the financial assets declared by SHIW respondents with data provided by a sample of commercial bank clients from a survey carried out by the bank. The authors used statistical matching to model non-reporting and under-reporting behavior and to adjust SHIW data. Although the adjusted estimates were much higher than the standard SHIW estimates, the difference between micro and macro estimates remained significant. Cannari and D'Alessio (1993) refined the previous experiment with a more complex model-based methodology and showed that the adjustment did not significantly affect the Gini concentration index. D'Alessio and Faiella (2002) studied a sample of about 2,000 households whose information had been matched anonymously with some banking information; in this case, they showed that non-response is not random but is more frequent among the wealthiest families. The bias detected for financial assets was significant (with adjusted estimates 15 to 30 per cent higher than unadjusted ones). D'Aurizio et al. (2008) replicated the statistical matching between commercial bank data and SHIW data. The adjusted estimates of financial assets averaged more than twice the original figures, reaching 85 per cent of the aggregate. The adjustment was more significant for households whose head is old or poorly educated. The paper also adjusted financial liabilities, whose corrected values were, on average, about 40 per cent higher. Neri and Ranalli (2011), using the results of a telephone survey conducted on SHIW non-respondents, reported higher difficulty obtaining interviews from the wealthiest households and proposed a corresponding adjustment of sampling weights. The result was confirmed by D'Alessio and Iezzi (2015). D'Alessio and Neri (2015) conducted several adjustment experiments on SHIW data, making wide use of calibration techniques, which produce estimates consistent with the macro-economic information to be used in the adjustments; however, when the sample estimates are very distant from the aggregate figures, calibrations produce unstable estimators. The results suggest that the unadjusted SHIW data underestimate wealth inequality.

This paper primarily uses the 2016 wave, the only wave for which data linkage with administrative records is currently available. In 2016, the survey-based estimate of the total deposit was about 34 percent of the aggregate total from national accounts. The macro figure, about 1,500 billion euros, can be considered a high-quality benchmark since it is not affected by nonresponse and measurement errors typical of survey data. Moreover, the deposit definition is highly comparable between the two sources. So, our adjustment method redistributes about 1,000 billion euros of deposits among the households in the SHIW sample.

2.2 Administrative data

Administrative data on household balance sheets exist in almost all European countries. Yet, only a few HFCS countries make substantial use of them for improving survey data. The main challenge is limited access because of legal, institutional, and practical constraints. In this paper, we use two sets of administrative data: the first relates to registers linked to the survey by individual identifiers. These registers include data on fiscal income (from tax registers), housing wealth (from cadastral records), and debts. They are used to identify a sub-group of respondents in the SHIW that may be considered highly reliable in the regression model for predicting the outstanding amounts of deposits. The second type of register data consists of aggregate banking supervisory reports; we use them to adjust the final distribution of deposits in the DWA through calibration techniques.

2.2.1 Administrative records on fiscal income, housing wealth, and loans

Tax records contain information on the revenues generated by individuals and the tax paid. They comprise income from employment, profits from sole trading, partnership income, company dividends, rental income, and private pension, while investment income is excluded. The definitions of income are sometimes different between tax records and the SHIW. However, thanks to the detailed information gathered in the survey, we can reconstruct income variables using the same definitions adopted in the administrative source. Since we use the difference between income data from the tax register and from the SHIW as a proxy of the reliability of the SHIW respondents, we construct within the SHIW income definitions that are as closest as possible to the ones used in the tax records. For example, in the tax records, social pensions are not included in the pension income; therefore, we exclude them from the definition of pension in the SHIW.

The cadastral register is an inventory of the real property present throughout the national territory, and consists of two distinct sub-systems: the first is the Land Cadastre which – comprises the list of all rural properties and unbuilt land plots, the second is the Urban Building Cadastre, which includes buildings for civil, industrial and commercial use. Each property has a market value computed based on of the average prices of rents and housing transactions in specific micro-zones, and the cadastral value of each property. This valuation criterion differs significantly from the one used in the SHIW, based on the respondent's self-assessment. Therefore, the number of properties is the only information we use from the cadastral records for assessing the reliability of the SHIW respondents. We only use the market value of the real property in the regression model for predicting the outstanding amounts of deposits.

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 who have 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). We take information on debt positions based on CR for all the SHIW respondents who appear in the register.

2.2.2 Banking Supervisory Reports

Italian Banking Supervisory Reports (BSR) include some distributional information on deposits held by households. Since 2008,[2] twice a year, at the end of June and December, banks provide the number of clients and the outstanding amounts of deposits by asset range of clients' deposits. The ranges are:

1) up to €12,500
2) €12,500-50,000
3) €50,000-250,000
4) €250,000-500,000
5) over €500,000

The aggregate value of deposits in the BSR data by asset range represents nearly 90% of the outstanding amounts of deposits in the Financial Account statistics (FA). The lower amount in the BSR data is mainly due to the absence of postal bonds issued by the Central Government (*Buoni postali fruttiferi*) and deposits held abroad.[3]

A client is defined through the personal fiscal code. If a client holds multiple checking accounts in the same bank, she is assigned to the range corresponding to the overall deposit amount. However, joint accounts are not split between holders but are considered different clients. For example, if two clients have one bank account each and one joint account, the bank registers them as three separate

[2] Data are available since 2008, but data quality improves significantly since 2013.
[3] The BSR also contain data on the overall outstanding amounts at market value for securities, listed shares, and investment fund shares (SSF, for brevity) held in custody at the reporting bank. The ranges for SSF are the same as for deposits, except for the first two intervals which are condensed into a unique class (below €50,000). The SSF outstanding amounts in BSR cover around 80% of debt securities, listed shares, and mutual fund shares in the FA and the difference depends on the estimates of financial assets held abroad. The present paper focuses on the estimation of deposits within the DWA procedure. However, the same method applied to deposits can be extended to SSF.

clients. Therefore, although the supervisory statistics are formally based on the definition of a client, the underlying concept is closer to the number of accounts.

The unit of observation in banking statistics differs from the one in the SHIW and FA. As mentioned above, a first departure is related to the statistical management of joint accounts. Second, different components of the same household unit are treated as separate clients. Third, the same household may hold checking accounts at more than one bank: this is the most relevant reason for the divergence between banking statistics and SHIW/FA.[4] According to SHIW, households hold, on average, two bank accounts.

Table 1 reports aggregate data from BSR on the outstanding amounts of deposits by asset range.[5] For example, the class of clients with more than €500,000 held around €90 billion in 2016, corresponding to about 9% of the overall deposits.[6]

Table 1: Amounts of clients' deposits by asset range from the BSR data

(*annual data; millions of euros and percent*)

Year	<12.5k	12.5-50k	50-250k	250-500k	>500k	Total	<12.5k (%)	12.5-50k (%)	50-250k (%)	250-500k (%)	>500k (%)	Total (%)
2013	132,736	272,734	359,55	77,075	74,114	916,209	14.5	29.8	39.2	8.4	8.1	100.0
2014	133,876	274,093	373,384	84,263	81,400	947,016	14.1	28.9	39.4	8.9	8.6	100.0
2015	132,967	274,802	388,707	88,438	86,820	971,734	13.7	28.3	40.0	9.1	8.9	100.0
2016	131,213	280,656	426,778	91,742	92,444	1,022,833	12.8	27.4	41.7	9.0	9.0	100.0
2017	132,199	283,860	440,306	94,304	95,544	1,046,213	12.6	27.1	42.1	9.0	9.1	100.0
2018	132,166	287,876	454,984	98,629	99,779	1,073,434	12.3	26.8	42.4	9.2	9.3	100.0
2019	130,400	292,579	485,909	109,835	112,125	1,130,848	11.5	25.9	43.0	9.7	9.9	100.0
2020	137,054	319,605	528,516	116,808	114,740	1,216,723	11.3	26.3	43.4	9.6	9.4	100.0

Note: this table reports for each deposit range (0-12.5k, 12.5-50k, 50-250k, 250-500k, >500k) the amount of deposits held by clients whose deposits fall in the deposit range.

[4] Suppose, for example, that an individual owns €600,000 of deposits. She is registered correctly in the wealthiest class if she holds the entire amount within a single account. However, if she splits her deposits into two accounts in two different banks, €300,000 each, she would be registered as two clients in the second most affluent class. If she splits the holdings into €10,000 and €590,000, she would show up as two different clients, one in the wealthiest class and the other in the poorest class.

[5] Data on the distribution of postal savings accounts between 2014 and 2016 are interpolated using data on 2013 and 2017, due to errors in the original reports.

[6] BSR data also contain information on the number of clients. Unfortunately, such information cannot be used as a benchmark. For example, the number of clients in the wealthiest class does not represent an upper or a lower bound for the number of people with at least €500,000 of financial wealth. Suppose, for example, an individual holds €1 million in deposits. If she splits them into three equally-sized amounts and deposits them at three different banks, she would be registered as three other people in the second most affluent class. So we would underestimate the number of the wealthiest households. Instead, if she splits into two equally-sized amounts, she would be registered as two different people in the most affluent class: we would overestimate the number of the rich.

Table 2 compares aggregate estimates for 2016 based on the SHIW with the distributional information from BSR. The total obtained in the SHIW is less than 40% of the BSR aggregates.[7] The shares of deposits held by the two wealthiest classes are similar, whereas the differences are remarkable for the first three classes. Specifically, thanks to the BSR values it is possible to identify a lower bound of the overall amount of deposits that are over a certain threshold. For example, households with deposits larger than €500,000 held at least €90 billion in 2016. Some households that should belong to this class may own part of their deposits at different banks, ending up with deposits lower than €500,000. Hence, those deposits would be classified in a lower asset range. The same reasoning holds for the other thresholds. For example, households with deposits greater than €50,000 hold at least €610 billion (i.e., the sum of the holdings of the three wealthiest classes). If we applied the proportional allocation method to fill the coverage gap, i.e., we used the SHIW distribution to the BSR total (see the last two columns of Table 2), we would end up with an overall amount of deposits held by the three wealthiest classes equal to €461 billion. Therefore, the proportional allocation would imply an undervaluation of at least €150 billion for the amount held by households with deposits higher than €50,000. This outlines the importance of including BSR distribution information within the DWA estimation procedure.

Table 2: Total deposits by asset range: BSR and SHIW

(*year=2016; millions of euros and per cent*)

	Values (€ billions)		Percentage		Difference	Coverage ratio	Proportional	Difference
	SHIW	BSR	SHIW	BSR	SHIW-BSR	SHIW/BSR	allocation	Prop.All.-BSR
<12,5k €	84,910	131,213	22.1	12.8	-46,303	64.7	226,046	94,833
12,5-50k €	125,851	280,656	32.8	27.4	-154,805	44.8	335,489	54,833
50-250k €	99,830	426,778	26.0	41.7	-326,948	23.4	265,937	-160,841
250-500k €	34,543	91,742	9.0	9.0	-57,199	37.7	92,055	313
>500k €	38,787	92,444	10.1	9.0	-53,657	42.0	103,306	10,862
Total	383,921	1,022,833	100.0	99.9	-638,912	37.5	1,022,833	0

3. Methodology

The method we use to adjust deposits in the DWA procedure consists of three steps.

In the first one, we select a subset of highly reliable households exploiting the data linkage with administrative records at the individual level. Then, we estimate a relationship between deposits and

[7] In the 2020 release of the SHIW, thanks to methodological changes to improve the statistical coverage of high-income households, the coverage ratio of deposits is around 50%.

some socio-demographic characteristics for the group of highly reliable households, and we then use the estimated coefficients to predict the value of deposits for the less reliable ones. Finally, we replace the values declared in the survey with the predicted ones when the latter ones are higher than the former.

In the second step, we calibrate the imputed results to the aggregate statistics from banking supervisory reports. We use traditional calibration methods widely used by national statistical offices but in a different way. While the standard practice is to apply them to the survey sampling weights, we apply them directly to the value of deposits.

Lastly, in the third step, we extend the methodology to the years when data linkage is not available.

The method that we propose is an alternative approach to the one used by the ESCB to produce DWA (see Appendix A.2 for a brief overview of such a procedure).

3.1 Imputation based on data linkage

In this section, we show the first step of our methodology, which selects a subset of highly reliable households to estimate deposits for less reliable households. The procedure exploits data linkage with administrative records at the individual level to adjust the amount of deposits declared by the SHIW respondents. However, the linkage applied only to 2016 data. It will be possible to apply the same estimation technique for successive releases of the data when they will become available.

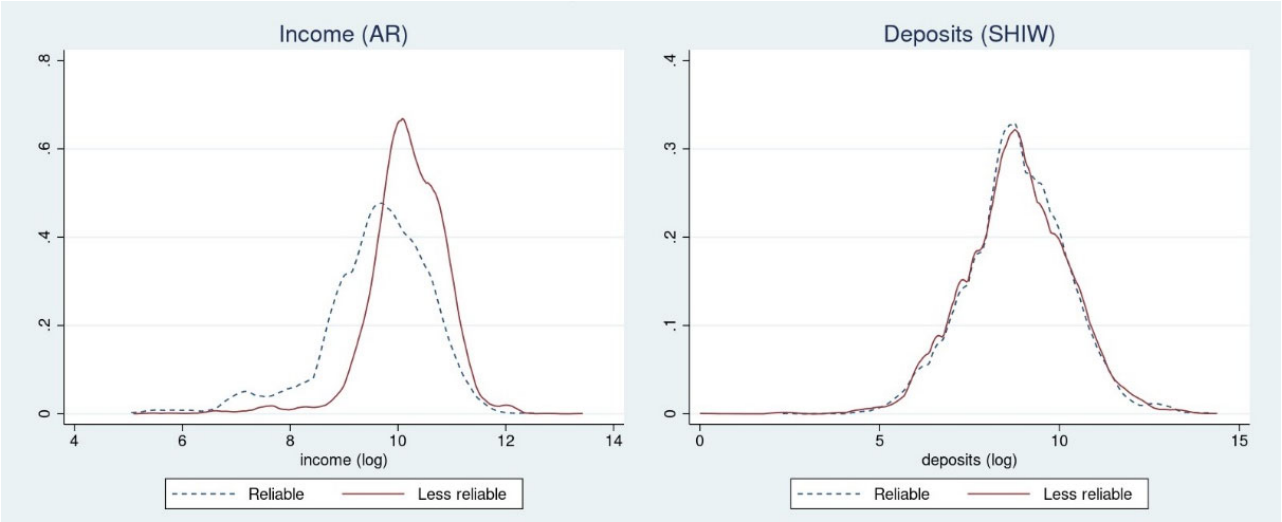
3.1.1 Identification of highly reliable households

We define highly reliable the households who reported in the survey a total income (from labor, pensions, transfers, and rents) that differs by less than 5 percent from the corresponding one in tax records. However, since evasion is a significant issue in fiscal data, we consider highly reliable also the households whose income is higher in the survey than in tax records. According to this definition, more than 40 percent of households are highly reliable.

We assume that households who underreport on their incomes are likely to underreport also on their financial assets, and in particular on deposits. Since administrative data on financial assets at the individual level are unavailable, we can only provide indirect evidence to support our working hypothesis. The left panel of Figure 1 compares the distribution of AR incomes between highly

reliable and less reliable households, while the right panel shows the distributions of deposits for the same two groups of households.[8] Less reliable households display higher incomes than highly reliable ones, whereas the two groups show similar distributions of deposits. Given the plausible assumption that households with higher levels of income also hold higher levels of deposits, the figure suggests that deposits of the less reliable households reported in the SHIW suffer from under-reporting.

Figure 1: Distributions of incomes and deposits.



Note: Highly reliable households are defined as those whose SHIW income is at least 95% of AR income. The distribution of deposits does not include observations with zero deposits.

As a robustness check, we also use six alternative definitions of highly reliable households (see Table 3). The first alternative definition considers those households with survey incomes at least 10 percent lower than the administrative ones as less reliable. The second defines as less reliable those households who own, according to registers, at least one property over the number declared in the survey. In our third definition, we consider less reliable those households with survey incomes less than 90 percent of administrative ones and with deposits less than 10 percent of monthly income; this second condition corresponds to the "income criterion" adopted in the standard ESCB procedure to identify outlier observations.[9] In our fourth definition, we add the so-called "asset criterion" in the standard ESCB procedure, i.e., households whose share of gross wealth held as deposits is lower than 0.8 percent are considered less reliable.[10] Our fifth definition mimics the ESCB method by identifying outliers according to the income criterion and/or the asset criterion. Finally, our sixth definition

[8] We exclude households with zero deposits. However, the share of households with zero deposits is broadly similar between the two groups.
[9] However, we do not apply this condition for households with a low annual income (less than €10,000) and with credit card debt
[10] This second condition does not apply to households with overdraft credit, mortgage debt, and null gross wealth

considers highly reliable those households that turn out reliable according to at least two criteria among the first, the second, and the fifth ones.

Table 3: Number of highly reliable households in the SHIW, according to different definitions.

No.	Definition	No. of obs.
	Baseline: SHIW income > 0.95*Admin. records (AR) income	3,069
1	SHIW income > 0.90*AR income	3,756
2	SHIW no. of properties > (AR no. of properties) - 100 (perc. points)	3,226
3	SHIW income > 0.90*AR income and meeting the ESCB income criterion	3,524
4	SHIW income > 0.90*AR income and meeting the ESCB income and asset criterion	3,121
5	meeting both the ESCB income and asset criteria	5,663
6	meeting at least 2 criteria: income (1); no. of properties (2); not an ESCB's outlier (5)	4,403
	Number of observations	7,130

3.1.2 Imputation technique

Then, we estimate a model for predicting deposits of highly reliable households. First, we run the following linear regression:

$$y_i = \alpha + \sum_{j=1}^{K} \beta_j \cdot x_{i,j} + \epsilon_i$$

where y_i represents deposits in the SHIW for household i, $x_{i,j}$ the j-th variable among the set of K covariates for household i, ϵ_i the idiosyncratic error. The covariates used in the estimates are:

- wages, pensions, self-employed incomes, profits, and rents (source: administrative records AR);
- real estate (source: AR);
- loans (source: AR);
- financial assets, other than deposits (source: SHIW);
- expenditures using banknotes (source: SHIW);
- durables and non-durable consumption (source: SHIW);
- overdraft credit and credit card debt (source: SHIW);
- savings (source: SHIW);
- age of the head of the household (source: SHIW);

- geographical macro area of residence (source: SHIW);
- household composition (source: SHIW);
- sector of occupation of the respondent (source: SHIW)

The dependent variable as well as the regressors, excluded demographic variables, are expressed in log terms.

Table A.1 reports the estimates obtained using 6 different sets of covariates. The sample is restricted to the group of highly reliable households based on our preferred definition (the income declared in the SHIW is at least 95% of the one in the AR). We exclude households with zero deposits from the analysis, either because they don't have a bank account or report a zero balance. As expected, in all the models, the deposit amount is positively associated with household income, financial wealth, real estate, and expenditure. Since we are primarily interested in the predicting power of the model, we perform 10-fold cross-validation, and we compute the average Root Mean Squared Error (RMSE) across folds.[11] We select model 6 which shows the lowest average RMSE.

Finally, we use the estimated coefficients in our selected model to predict deposits for the subsample of less reliable households. Since deposits are expressed in log terms, we obtain predictions applying Duan's smearing transformation, which does not need any particular assumption on the distribution of the residuals (Duan, 1983). We assign the predicted values to less reliable households when predictions are higher than the observed deposits. Then, the micro-level database with adjusted deposits can be included within the ESCB DWA procedure to obtain inequality indicators on household net wealth.

3.2 Calibration techniques using BSR data

In the second step, we calibrate the imputed results to the aggregate statistics from supervisory data. As described in Section 2.2.2, the BSR data include information on the outstanding deposits by asset range of clients' accounts. These statistics cover nearly 90 percent of overall deposits from financial accounts. Since the procedure should match national accounts aggregates, BSR data are rescaled to official figures.

[11] The procedure starts by splitting the sample into 10 equally-sized groups. The regression is performed using only 9 groups out of 10; then, the estimated coefficients are applied to the tenth group, and the RMSE is stored. This step is replicated leaving out one group at each step. In the end, we take the average of the RMSEs obtained at each step. The lowest the average RMSE, the better the model.

Comparing banking statistics and SHIW data by asset range is not straightforward because of the different observation units. The SHIW reports the number of deposit accounts for each household, which is insufficient to split deposits by account. In the 2020 release of the SHIW, a new question was introduced about the share of deposits held in their main deposit account. On average, households with more than one deposit account have around 66% of their deposits in the main account. This percentage slightly declines with the increase in deposits but remains over 63% for households with five bank accounts. We use this information to estimate how the deposits of each household are distributed across bank accounts.[12] Then, we transform our dataset at the household level into a bank account-level database. Therefore, we can apply standard calibration techniques to match deposits at the bank-account level with aggregate figures from BSR data.

Let 1,..*i*,..*n* be the observations on bank accounts. Let the vector a_i denote the adjustment factors at the bank account level that allow reducing the gap with aggregate BSR data. Let w_i be the vector of survey weights and x_i the vector of deposit amounts. Moreover, let $I_{i,c}$ be an indicator function that identifies to which asset range the *i-th* bank account belongs among C = 5 asset ranges. Finally, let X_c the total deposits of asset range $c \in C$ from BSR data. The calibration solves the following problem:

$$\min_{a} \sum_{i=1}^{n} \frac{(w_i a_i - w_i)^2}{w_i}$$

$$s.t. \quad \sum_{i=1}^{n} w_i \cdot (a_i \cdot x_i) \cdot I_{i,c} = X_c$$

$$a_i \in [min_a, max_a] \; \forall i \in \{1, \dots, n\}$$

The adjustment factors a_i are constrained to range from 0.5 to 10.[13] After convergence, we multiply the estimated factors a_i by the observed deposit amount, obtaining imputed values of deposits $\hat{x}_i = x_i \cdot \hat{a}_i$. Since the constraints in the calibration depend on the initial classification of deposits across asset ranges (i.e., in the formulas above $I_{i,c}$ does not depend on a_i), convergence does not guarantee that constraints are met. Indeed, for some observations, the asset range of the imputed deposits may differ from the one associated with the initial values pre-calibration $I_{i,c}$. Therefore, we perform

[12]We assume that households with more than two accounts hold 20% of the overall amount in their second-largest account, and we split the residual into equal amounts across the remaining accounts. For the releases before 2020, we attribute 66% of deposits to the first account. When the number of bank account information is missing, we impute it using average values by estimating in the SHIW. In particular, we attribute one bank account if deposits are lower than €25,000, 2 if deposits are between €25,000 and €75,000, 3 if they are higher.

[13] We perform robustness checks using other parameters. However, depending on the wave, the range cannot be restricted too much otherwise convergence is not achieved.

several iterations of the calibration procedure and select the iteration step with the lowest mean squared error. We also penalized observations moving from one class to another to preserve the original distribution broadly.

It is worth stressing that in the standard approach, the adjustment factors are multiplied by the survey weights to get a new set of weights that aligns the survey-based totals with the external benchmarks. Instead, we apply these factors directly to the deposit amounts. This solution produces the same result as the traditional approach. Still, it has the merit of not modifying the sampling weights; therefore, it does not affect all the other statistics.

3.3 The method applied for waves where administrative data are missing

The starting period for the Italian DWA statistics is 2010, in correspondence with the first wave of HFCS. The step described in Section 3.1 requires administrative data from tax files and the cadastral register which are available in 2016.

First of all, for those waves, we need an alternative procedure to identify highly reliable households. For panel observations, i.e. those households included in the 2010 and 2014 releases who were interviewed also in 2016 (the third wave), we assume that the reliability of their answers is constant across waves. Of course, while nearly one-third of the 2014 respondents were also interviewed in 2016, the percentage was just around 15 percent in 2010, so the number of highly reliable respondents based on panel observations would be too low. Therefore, we use additional information available in the survey. In particular, the interviewer provides a score on the quality of the respondents' answers, ranging from one to 10, and reports if the respondent consulted some documentation to answer the questionnaire. We consider reliable those households with a high-quality score (at least 9) and who consulted some documentation during the interview. A similar approach applies to the 2020 wave.

Second, we run a regression using the same set of covariates of the model chosen for the 2016 wave, restricting the estimates to the highly reliable households. In contrast with wave 3, in this case we do not have any regressor based on administrative sources but only on the SHIW.

Finally, as in the standard method, we apply the estimated coefficient to less reliable households and we obtain predicted values of log deposits. As usual, we transform log deposits through Duan's smearing transformation and we assign the predicted values to less reliable households when they are

higher than the observed deposits.

Concerning the step described in Section 3.2, the BSR data needed for the calibration constraints are available for the entire time span of the DWA statistics.

4. Results

According to SHIW data, the distribution of deposits shows a high degree of concentration Table 4). The top 5 percent holds about half the total amount, while the bottom 50 percent is around 5 percent. The adjustment method leads to a slight decrease in the inequality of deposit distribution. The Gini index drops from 74 percent to 72. The top 5 percent's share increases by seven p.p. to some 55 percent. At the same time, households in the bottom 50 percent increase their share by about four p.p., to 9% of the total.

Table 4: Effect of the adjustment on the deposit distribution (2016).

	Gini	Median Wealth (TEUR)	Deposit Share (%) held by:				
			Top 1%	Top 5%	Top 10%	50-90%	Bottom 50%
Raw SHIW data	0.74	5.00	25.34	47.62	61.17	33.75	5.08
Step 1: Imputation based on data linkage	0.68	20.71	32.94	47.89	57.32	33.09	9.59
Step 2: Calibration correction	0.76	11.33	23.53	52.00	67.93	25.94	6.13
Final estimates (Imputation and Calibration)	0.72	17.12	26.70	54.72	64.48	26.81	8.71
Final estimates with the ESCB method	0.73	17.39	36.36	53.37	63.40	28.67	7.93

Most of the results align with those obtained with the ESCB baseline method, which is our primary benchmark. The median value of deposits is about 17,000 euro, and the Gini coefficient is around 72 percent, while it is slightly larger in the ESCB method (73%). However, the share of deposits held by the top 1 percent (in terms of deposits) is much higher according to the ESCB approach (36%) than in the proposed method (27%). Indeed, as shown in Figure 2, thanks to the calibration step, our approach generally guarantees a closer match with aggregates from banking statistics compared to the ESCB approach, which tends to overestimate the wealthiest part of the deposit distribution. Conversely, the calibration performs poorly for a few combinations of wave and asset range (for example, for the class €250-500,000 in waves 1 and 3).

Figure 2: Calibration techniques: comparison with administrative data

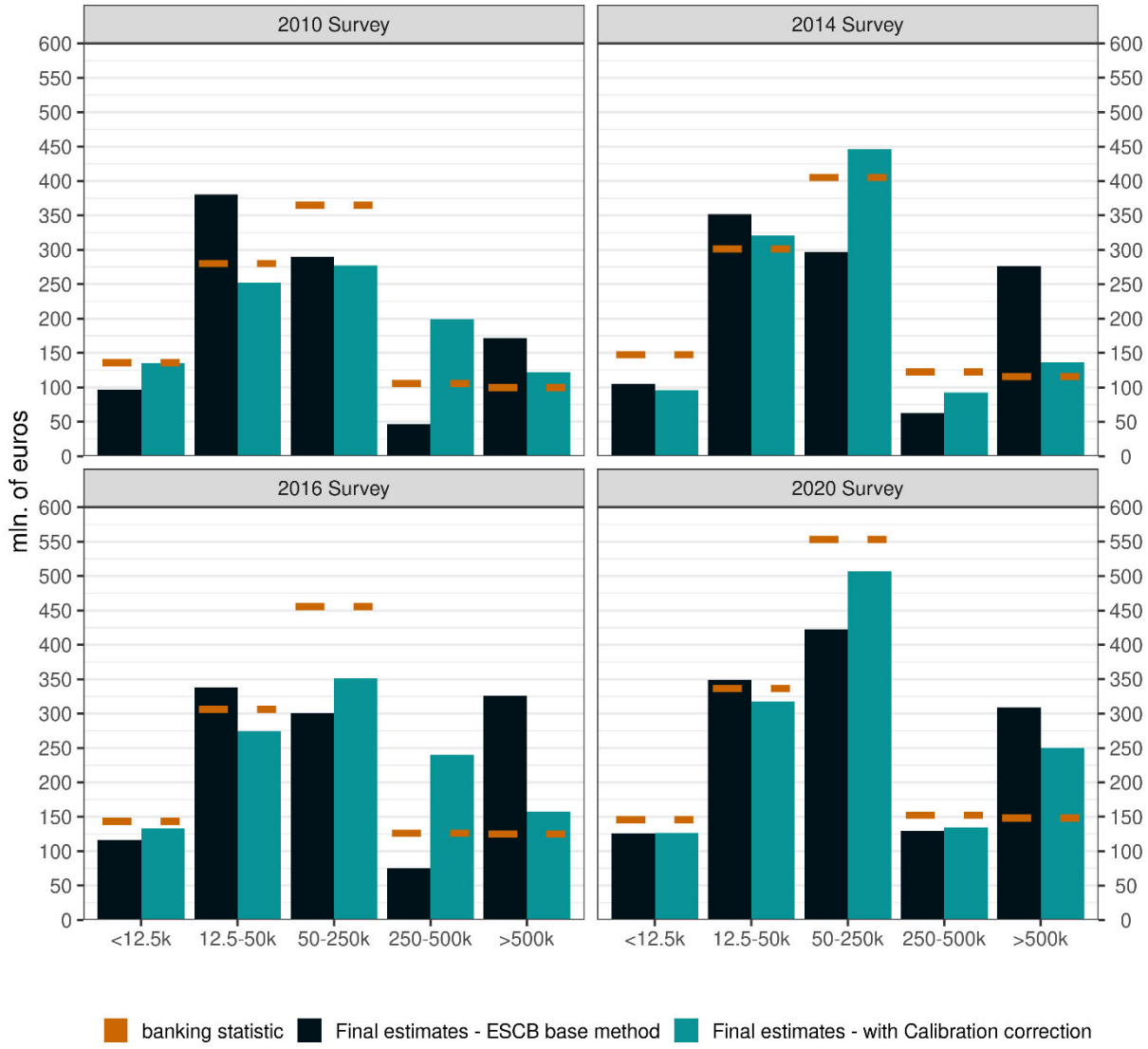
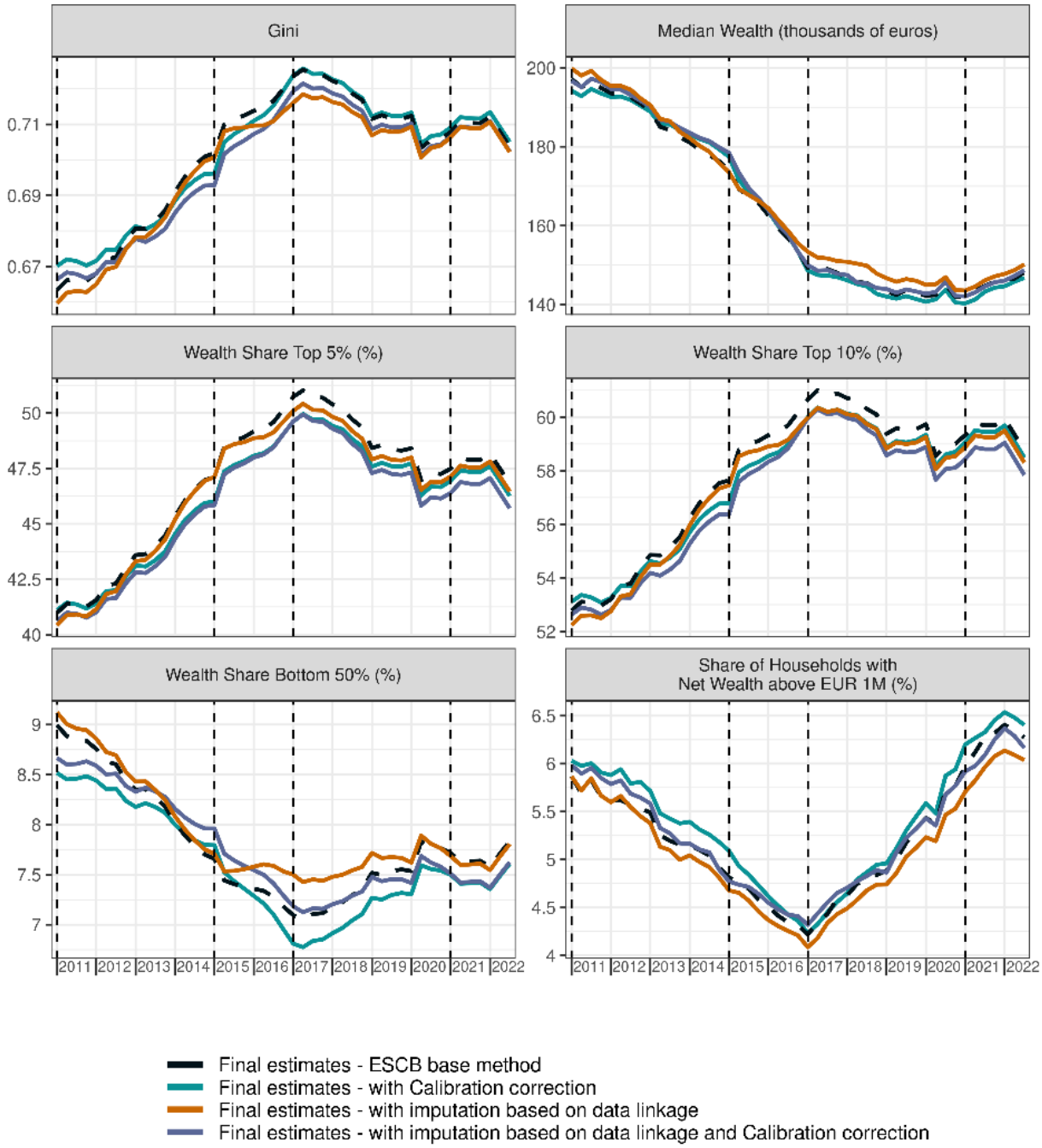


Figure 3 shows the effect of our methodology on the inequality of net wealth distribution compared with the ESCB approach. Overall, the two methods provide a very consistent picture. From 2011 to 2016, the wealth share held by the top 5 and 10 percent increased while the median value decreased, leading to a rise in the Gini index. Since 2017, the wealth shares of the top 5 and 10 percent have decreased, while the median wealth remained almost flat, producing a slight decrease in the Gini index. Minor differences between the two methods concern the top part of the distribution. While the ESCB method attributes more wealth to the households at the top of the distribution, our approach produces a slightly less unequal net wealth distribution.

Figure 3: The impact of the methodology on net wealth inequality



5. Robustness checks

We perform four exercises to assess the robustness of our results.

First, in Table A.2 we run the selected model 6 using different definitions of reliable household introduced in Table 3. The coefficients are generally stable across models using different definitions. In particular, the coefficient on income is always statistically significant and ranges between 0.13

using definition 2 and 0.20 using definition 5 (it is 0.16 according to the base definition).

Second, we use hurdle models to assess the effects of excluding households with zero deposits from the group of reliable households, on which we run the deposit imputation model. These models represent an alternative to linear regressions and allow treating corner solutions as observed instead of censored (Cragg, 1971). They consist of a selection equation, which in our case regards the probability of owing deposits, and an outcome equation, which determines the relation of deposits to other explanatory variables, given that deposits are positive. Table A.3 reports the estimates obtained using different sets of variables, both in the selection and in the outcome equations. The sample is restricted to the group of highly reliable households according to our first definition (the income declared in the SHIW is at least 95% of the one in the tax registers). We perform a 10-fold cross-validation and we compute the average RMSE across folds to select the model with the highest ability of prediction.[14] In the selection equation, we always exclude the overall amounts of financial assets, loans, debts, and expenditures, since we consider these variables more useful for explaining the level of deposits than the probability of holding them.

As a third exercise, we also assess estimates from the Hurdle model (sixth specification of Table A.3) using all the different definitions of reliable respondents (Table A.4).

Finally, Table A.5 presents some inequality indicators on the net wealth distribution obtained from the DWA procedure after using different combinations of imputation models and definitions of reliability. Overall, the results show a low variability across different models and definitions of reliability. The median net wealth is about 153,000 euro with a range of variation of around one p.p.; similarly, the share held by the top 10% ranges between 59.6 and 60.6 percent and the Gini index between 71 and 72.3 percent. The percentage of households with over one million net wealth remains around 4 percent. Compared to the ESCB method, our methodology results in a slightly lower inequality in the net wealth distribution. According to the ESCB method, the share of wealth held at the top of the distribution is always larger or equal to the one observed in our approach, and the percentage at the bottom of the distribution is lower. The median wealth is €153,000 on average across the different combinations of models and definitions compared to the €149,000 from the ESCB method. Overall, the Gini coefficient reduces from 72.3% to 71.6%.

[14] We set predicted deposits to zero when the predicted probability of holding deposits from the selection equation is lower than 0.5.

Although hurdle models are attractive for considering corner solutions, they display some shortcomings, which may be relevant in compiling DWA statistics. First, there is no guarantee that the estimation through MLE converges. Second, they require extra assumptions to define a selection equation. Third, the Pseudo-R^2 is relatively low (in model 6 is less than 0.1), suggesting that the model's predictive ability is not very satisfactory. Fourth, the final impact on coverage and inequality measures does not differ markedly from the regression models. Moreover, the share of households declaring zero deposits is shrinking over time. Therefore, we prefer to follow a simple linear regression model.

6. Conclusions

The DWA statistics developed by the ECB Expert Group on Distributional Financial Accounts provide a comprehensive view of the distribution of household wealth by adjusting survey data to obtain aggregate figures coherent with national accounts. A vital adjustment on survey data concerns deposits since this instrument represents a significant share of gross household wealth, and its coverage of national figures is low. Because of the lack of external information, the adjustment is based on identifying outlier observations and their replacement with average values. This paper proposes an alternative method for Italian data drawing on additional information from administrative records and banking supervisory reports.

First, we use register data to identify subsets of respondents that may be considered highly reliable. Then, we estimate a relationship between deposits and some socio-demographic characteristics for the group of highly reliable households, and we use the estimated coefficients to predict the value of deposits for the less reliable ones.

We then use aggregate statistics from banking supervisory reports, which regard the outstanding deposits by asset range of clients' holdings. Finally, we adjust survey observations using calibration techniques to match aggregate information by asset range from supervisory reports.

The proposed adjustment method leads to a slight decrease in the inequality of deposit distribution in 2016. The Gini index in terms of deposit holdings drops from 74 percent in the survey data to 72. The top 5 percent's share increases by seven p.p. to some 55 percent. At the same time, households in the bottom 50 percent increase their share by about 4 p.p., to 9 percent of the total. The proposed method generally guarantees a closer match with aggregates from banking statistics

compared to the ESCB approach, which tends to overestimate the deposits in the wealthiest part of the deposit distribution. However, regarding net wealth, the inequality indicators obtained are pretty in line with those from the ESCB baseline method, with a slightly lower share of net wealth held at the top of the distribution.

Further extensions of the calibration techniques presented in this paper will be implemented in future research projects. Banking statistics by asset range are available semi-annually so that they can be used for improving the interpolation and extrapolation of the DWA quarterly time series when survey data are unavailable. Moreover, a similar methodology based on supervisory reports can be applied to debt securities, listed shares, and investment fund shares. Other improvements of the Italian DWA estimation procedure include the usage of administrative data on debts and real estate properties.

References

Cannari, L. and D'Alessio, G. (1993). Non-reporting and under-reporting behavior in the bank of Italy's survey of household income and wealth. Bulletin of the International Statistical Institute–Proceedings of the 49th ISI Session, 55(3):395–412.

Cannari, L., D'Alessio, G., Raimondi, G., and Rinaldi, A. (1990). Le attività finanziarie delle famiglie italiane. Temi di discussione 136, Banca d'Italia.

Cantarella, M., Neri, A., and Ranalli, G. (2023). Estimating the distribution of household wealth: methods for adjusting survey data estimates using national accounts and rich list data, Review of Income and Wealth, forthcoming.

Cragg, J. G. (1971). Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods. Econometrica, 39(5):829–844.

D'Alessio, G. and Faiella, I. (2002). Non-response behaviour in the Bank of Italy's Survey of Household Income and Wealth. Temi di discussione (Economic working papers) 462, Bank of Italy, Economic Research and International Relations Area.

D'Alessio, G. and Iezzi, S. (2015). How the Time of Interviews Affects Estimates of Income and Wealth. Bank of Italy Occasional Paper 273, Bank of Italy, Economic Research and International Relations Area.

D'Alessio, G. and Neri, A. (2015). Income and Wealth Sample Estimates Consistent with Macro Aggregates: Some Experiments. Bank of Italy Occasional Paper 272, Bank of Italy, Economic Research and International Relations Area.

D'Aurizio, L., Faiella, I., Iezzi, S., and Neri, A. (2008). The under-reporting of households' financial assets in Italy. In for International Settlements, B., editor, The IFC's contribution to the 56th ISI Session, Lisbon, August 2007, volume 28 of IFC Bulletins chapters, pages 415–420. Bank for International Settlements.

Duan, N. (1983). Smearing estimate: A nonparametric retransformation method. Journal of the American Statistical Association, 78(383):605–610.

Engel, J., Riera, P. G., Grilli, J., and Sola, P. (2022). Developing Reconciled Quarterly Distributional National Wealth – Insight into Inequality and Wealth Structures. Working Paper Series, no. 2687, European Central Bank.

Expert Group on Linking macro and micro data for the household sector (EG-LMM), (2020). Understanding household wealth: linking macro and micro data to produce distributional financial accounts. Statistics Paper Series 37, European Central Bank.

Neri, A. and Ranalli, M. G. (2011). To misreport or not to report? The measurement of household financial wealth. Statistics in Transition, 12(2): 281--300.

Ulizzi, A. (1970). Risparmio e struttura della ricchezza delle famiglie italiane nel 1968. In Banca d'Italia, editor, Bollettino, volume 1, pages 103–167. Banca d'Italia.

A Appendix

A.1 List of Tables

Table A.1: Linear regression models with different sets of covariates.

	(1) deposits (log) b/se	(2) deposits (log) b/se	(3) deposits (log) b/se	(4) deposits (log) b/se	(5) deposits (log) b/se	(6) deposits (log) b/se
income (AR) (log)	0.273*** (0.0482)		0.147*** (0.0557)		0.229*** (0.0512)	0.160*** (0.0522)
real estate (AR) (log)	0.036*** (0.0079)	0.041*** (0.0081)	0.030*** (0.0073)	0.031*** (0.0076)	0.033*** (0.0079)	0.026*** (0.0071)
financial assets (excl. deposits) (log)	0.037*** (0.0102)	0.046*** (0.0099)	0.024** (0.0104)	0.024** (0.0104)	0.036*** (0.0100)	0.019* (0.0105)
loans (AR) (log)	0.004 (0.0091)	0.004 (0.0095)	0.002 (0.0088)	0.002 (0.0090)	0.003 (0.0091)	0.007 (0.0084)
wages (AR) (log)		0.016 (0.0102)		-0.004 (0.0100)		
pensions (AR) (log)		0.045*** (0.0102)		0.034*** (0.0103)		
self-employed income, profits, rents (AR) (log)		0.047*** (0.0123)		0.014 (0.0129)		
expenditures using banknotes (log)			0.218*** (0.0616)	0.188*** (0.0614)		0.212*** (0.0626)
durables (log)			0.033*** (0.0110)	0.040*** (0.0115)		0.035*** (0.0106)
non-durable consumption (log)			0.432*** (0.1029)	0.588*** (0.0975)		0.562*** (0.0964)
overdraft credit (log)			-0.066** (0.0295)	-0.060** (0.0285)	-0.053* (0.0315)	-0.073** (0.0296)
credit card debt (log)			0.005 (0.0269)	0.008 (0.0264)	0.009 (0.0309)	0.007 (0.0299)
savings (log)					0.042*** (0.0120)	
age of the head of the household						0.013 (0.0138)
age of the head of the household (squared)						-0.000 (0.0001)
constant	5.832*** (0.4553)	7.862*** (0.1234)	1.381* (0.8230)	1.204 (0.8739)	6.000*** (0.4597)	-0.680 (0.9176)
Adjusted R^2	0.106	0.091	0.151	0.156	0.119	0.209
10-fold CV RMSE (ave)	1.322	1.326	1.304	1.296	1.313	1.285
Observations	2332	2332	2332	2332	2332	2332

Note: Regressions are estimated using the subsample of highly reliable households (SHIW income >0.95 * AR income). Model (1) includes as regressors: the overall income from tax registers, the value of real estate properties from the cadastral register, total financial assets from the SHIW, and loans from the credit register. Model (2) includes wages, pensions, self-employed income, profits, and rents separately. Model (3) includes covariates related to expenditures, durables, and debts other than loans (overdraft credit and credit card debt). In model (4), we split incomes again into their components. Model (5) excludes variables on expenditures and durables, and includes overall savings. Finally, model (6) includes some demographic information, such as the age of the household head (also entering with a quadratic term), the macro area of residence (North-West, North-East, Center, South, Isles), the household composition (e.g., single, married without children) and the sector of occupation of the respondent.

Table A.2 – Base regressions: different definitions of highly reliable households

	Base def. 0.95*AR	Def. 1 0.90*AR	Def. 2 property	Def. 3 Def. 2 + ESCB(inc)	Def. 4 Def. 2 + ESCB (inc-ass)	Def. 5 ESCB (inc-ass)	Def. 6 Def 2+3+6
	b/se	b/se	b/se	b/se	b/se	b/se	b/se
income (AR) (log)	0.160***	0.180***	0.130**	0.187***	0.168***	0.203***	0.183***
	(0.0522)	(0.0505)	(0.0570)	(0.0503)	(0.0506)	(0.0414)	(0.0447)
real estate (AR) (log)	0.026***	0.028***	0.016**	0.028***	0.033***	0.030***	0.032***
	(0.0071)	(0.0067)	(0.0070)	(0.0067)	(0.0067)	(0.0051)	(0.0055)
financial assets (excl. deposits) (log)	0.019*	0.020**	0.028***	0.019**	0.026***	0.025***	0.025***
	(0.0105)	(0.0091)	(0.0106)	(0.0090)	(0.0091)	(0.0063)	(0.0081)
loans (AR) (log)	0.007	-0.005	-0.008	-0.004	-0.005	-0.005	-0.010
	(0.0084)	(0.0077)	(0.0088)	(0.0076)	(0.0074)	(0.0053)	(0.0069)
expenditures using banknotes (log)	0.213***	0.193***	0.215***	0.176***	0.127**	0.084*	0.141***
	(0.0626)	(0.0563)	(0.0632)	(0.0555)	(0.0544)	(0.0430)	(0.0494)
durables (log)	0.035***	0.043***	0.036***	0.043***	0.037***	0.032***	0.039***
	(0.0106)	(0.0095)	(0.0123)	(0.0094)	(0.0093)	(0.0073)	(0.0086)
non-durable consumption (log)	0.562***	0.546***	0.563***	0.544***	0.599***	0.544***	0.562***
	(0.0964)	(0.0924)	(0.1042)	(0.0915)	(0.0952)	(0.0673)	(0.0807)
overdraft credit (log)	-0.073**	-0.071***	-0.105***	-0.060**	-0.077***	-0.084***	-0.078***
	(0.0296)	(0.0265)	(0.0275)	(0.0258)	(0.0258)	(0.0242)	(0.0247)
credit card debt (log)	0.007	-0.029	-0.126***	-0.032	-0.045	-0.051**	-0.056*
	(0.0299)	(0.0316)	(0.0360)	(0.0315)	(0.0311)	(0.0231)	(0.0292)
age of the head of the household	0.013	0.006	0.011	0.006	0.009	0.021**	0.016
	(0.0138)	(0.0134)	(0.0145)	(0.0134)	(0.0128)	(0.0100)	(0.0115)
age of the head of the household (squared)	-0.000	0.000	0.000	0.000	0.000	-0.000	-0.000
constant	-0.677	-0.453	-0.641	-0.408	-0.604	-0.764	-0.843
	(0.9179)	(0.8595)	(1.0132)	(0.8541)	(0.8663)	(0.6451)	(0.7589)
Adjusted R^2	0.209	0.199	0.204	0.204	0.246	0.235	0.238
Observations	2332	2978	2306	2970	2756	5228	3715

Note: Each column displays the estimates of Model 6 (Table A.1) restricting the sample to different definitions of highly reliable households reported in Table 3. All the estimates include also dummies on geographical residence, household composition, and sector of occupation of the respondent.

Table A.3 – Hurdle models with different sets of covariates: estimates.

	(1) deposits (log) b/se	(2) deposits (log) b/se	(3) deposits (log) b/se	(4) deposits (log) b/se	(5) deposits (log) b/se	(6) deposits (log) b/se
Outcome model						
income (AR) (log)	0.070***			0.070***		0.031**
	(0.0151)			(0.0151)		(0.0147)
real estate (AR) (log)	0.042***	0.039***	0.039***	0.042***	0.040***	0.025***
	(0.0073)	(0.0077)	(0.0077)	(0.0073)	(0.0077)	(0.0069)
financial assets (excl. deposits) (log)	0.047***	0.048***	0.048***	0.047***	0.048***	0.021**
	(0.0096)	(0.0097)	(0.0097)	(0.0096)	(0.0098)	(0.0101)
loans (AR) (log)	0.005	0.003	0.003	0.005	0.003	0.005
	(0.0089)	(0.0091)	(0.0091)	(0.0089)	(0.0091)	(0.0081)
wages (AR) (log)		0.015	0.015		0.016*	
		(0.0096)	(0.0096)		(0.0096)	
pensions (AR) (log)		0.043***	0.043***		0.043***	
		(0.0088)	(0.0088)		(0.0089)	
self-employed income, profits, rents (AR) (log)		0.042***	0.042***		0.043***	
		(0.0114)	(0.0114)		(0.0114)	
overdraft credit (log)					-0.041	-0.076***
					(0.0300)	(0.0283)
credit card debt (log)					0.016	0.023
					(0.0268)	(0.0294)
expenditures using banknotes (log)						0.184***
						(0.0605)
durables (log)						0.038***
						(0.0101)
non-durable consumption (log)						0.687***
						(0.0919)
age of the head of the household						0.021*
						(0.0128)
age of the head of the household (squared)						-0.000
						(0.0001)
constant	7.744***	7.922***	7.922***	7.744***	7.918***	-0.632
	(0.1401)	(0.0977)	(0.0977)	(0.1401)	(0.0982)	(0.8891)
Selection model						
income (AR) (log)	0.075***	0.075***		0.059***		
	(0.0123)	(0.0123)		(0.0137)		
real estate (AR) (log)	0.024***	0.024***	0.020***	0.037***	0.032***	0.032***
	(0.0072)	(0.0072)	(0.0074)	(0.0083)	(0.0085)	(0.0085)
wages (AR) (log)			0.041***		0.028**	0.028**
			(0.0105)		(0.0121)	(0.0121)
pensions (AR) (log)			0.042***		0.053***	0.053***
			(0.0097)		(0.0131)	(0.0131)
self-employed income, profits, rents (AR) (log)			0.043***		0.038***	0.038***
			(0.0117)		(0.0124)	(0.0124)
age of the head of the household				-0.072***	-0.072***	-0.072***
				(0.0181)	(0.0181)	(0.0181)
age of the head of the household (squared)				0.001***	0.001***	0.001***
				(0.0002)	(0.0002)	(0.0002)
constant	0.221**	0.221**	0.353***	2.449***	2.686***	2.686***
	(0.1108)	(0.1108)	(0.0916)	(0.5435)	(0.5399)	(0.5399)
lnsigma						
constant	0.274***	0.271***	0.271***	0.274***	0.270***	0.198***
	(0.0213)	(0.0218)	(0.0218)	(0.0213)	(0.0219)	(0.0205)
Pseudo R2	0.034	0.035	0.037	0.055	0.058	0.090
10-fold CV RMSE (ave)	3.652	3.654	3.654	3.649	3.629	3.542
Observations	2993	2993	2993	2993	2993	2993

Note: Each column displays the estimates of Model 6 restricting the sample to highly reliable households (SHIW income >0.95 * AR income). The selection equation in models 4-6, as well as the outcome model in model 6, include also dummies on geographical residence, household composition, and sector of occupation of the respondent.

Table A.4 – Hurdle models: different definitions of highly reliable households.

	Base def. 0.95*AR	Def. 1 0.90*AR	Def. 2 property	Def. 3 Def. 2 + ESCB(inc)	Def. 4 Def. 2 + ESCB (inc-ass)	Def. 5 ESCB (inc-ass)	Def. 6 Def 2+3+6
	b/se	b/se	b/se	b/se	b/se	b/se	b/se
Oucome model							
income (AR) (log)	0.031**	0.030**	0.017	0.031**	0.025*	0.036***	0.027**
	(0.0147)	(0.0146)	(0.0153)	(0.0146)	(0.0143)	(0.0127)	(0.0133)
real estate (AR) (log)	0.025***	0.028***	0.016**	0.029***	0.035***	0.033***	0.033***
	(0.0069)	(0.0065)	(0.0068)	(0.0065)	(0.0064)	(0.0049)	(0.0054)
financial assets (excl. deposits) (log)	0.021**	0.022**	0.029***	0.022**	0.028***	0.027***	0.027***
	(0.0101)	(0.0088)	(0.0104)	(0.0087)	(0.0088)	(0.0061)	(0.0078)
loans (AR) (log)	0.005	-0.006	-0.008	-0.005	-0.005	-0.005	-0.011
	(0.0081)	(0.0074)	(0.0083)	(0.0073)	(0.0070)	(0.0052)	(0.0066)
expenditures using banknotes (log)	0.184***	0.167***	0.181***	0.151***	0.111**	0.080*	0.120**
	(0.0605)	(0.0549)	(0.0622)	(0.0541)	(0.0526)	(0.0423)	(0.0484)
durables (log)	0.038***	0.045***	0.038***	0.046***	0.039***	0.034***	0.040***
	(0.0101)	(0.0092)	(0.0117)	(0.0091)	(0.0090)	(0.0072)	(0.0083)
non-durable consumption (log)	0.687***	0.671***	0.683***	0.674***	0.718***	0.647***	0.681***
	(0.0919)	(0.0872)	(0.0995)	(0.0867)	(0.0889)	(0.0636)	(0.0764)
overdraft credit (log)	-0.076***	-0.073***	-0.104***	-0.063**	-0.081***	-0.087***	-0.081***
	(0.0283)	(0.0254)	(0.0271)	(0.0248)	(0.0249)	(0.0240)	(0.0239)
credit card debt (log)	0.023	-0.014	-0.124***	-0.017	-0.030	-0.038	-0.043
	(0.0294)	(0.0308)	(0.0347)	(0.0307)	(0.0307)	(0.0235)	(0.0288)
age of the head of the household	0.021*	0.015	0.017	0.015	0.019	0.026***	0.022**
	(0.0128)	(0.0125)	(0.0134)	(0.0124)	(0.0118)	(0.0095)	(0.0108)
age of the head of the household (squared)	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
constant	-0.632	-0.267	-0.581	-0.217	-0.549	-0.243	-0.506
	(0.8891)	(0.8277)	(0.9562)	(0.8225)	(0.8346)	(0.6226)	(0.7285)
Selection model							
wages (AR) (log)	0.028**	0.029**	0.054***	0.059***	0.020	0.062***	0.055***
	(0.0121)	(0.0112)	(0.0113)	(0.0154)	(0.0201)	(0.0155)	(0.0120)
pensions (AR) (log)	0.053***	0.045***	0.063***	0.103***	0.076***	0.091***	0.063***
	(0.0131)	(0.0127)	(0.0121)	(0.0142)	(0.0213)	(0.0171)	(0.0130)
self-employed income, profits, rents (AR) (log)	0.038***	0.041***	0.010	0.064***	0.084***	0.072***	0.062***
	(0.0124)	(0.0114)	(0.0111)	(0.0143)	(0.0214)	(0.0179)	(0.0123)
real estate (AR) (log)	0.032***	0.032***	0.034***	0.035***	0.047***	0.061***	0.058***
	(0.0085)	(0.0079)	(0.0079)	(0.0091)	(0.0139)	(0.0120)	(0.0080)
age of the head of the household	-0.072***	-0.062***	-0.030**	-0.051***	-0.039	-0.053**	-0.061***
	(0.0181)	(0.0172)	(0.0148)	(0.0197)	(0.0281)	(0.0244)	(0.0172)
age of the head of the household (squared)	0.001***	0.001***	0.000***	0.000***	0.000	0.001**	0.001***
	(0.0002)	(0.0001)	(0.0001)	(0.0002)	(0.0002)	(0.0002)	(0.0001)
Constant	2.686***	2.319***	1.219***	1.836***	1.668**	2.015***	2.091***
	(0.5399)	(0.5153)	(0.4499)	(0.5892)	(0.8388)	(0.6973)	(0.5165)
lnsigma							
constant	0.198***	0.217***	0.219***	0.203***	0.138***	0.146***	0.178***
	(0.0205)	(0.0187)	(0.0206)	(0.0184)	(0.0200)	(0.0147)	(0.0167)
Pseudo R2	0.090	0.084	0.099	0.102	0.104	0.098	0.103
Observations	2993	3717	3199	3469	3003	5551	4326

Note: Both the selection and the outcome equations include also dummies on geographical residence, household composition, and sector of occupation of the respondent.

Table A5: Different models: impact on coverage and inequality indicators.

Model	Var. Model	Reliab. Def.	% of HH wealth > €1 mil.	Top 5%	Top 10%	Top 20%	Bottom 50%	Gini	Median Wealth
ESCB			4.2%	50.7%	60.7%	73.3%	7.1%	72.3%	149,338
Lin. Reg.	1	1	4.1%	49.8%	59.6%	72.2%	7.9%	71.0%	155,856
Lin. Reg.	2	1	4.0%	49.9%	59.6%	72.2%	7.8%	71.1%	154,976
Lin. Reg.	3	1	4.1%	50.0%	59.9%	72.5%	7.6%	71.4%	153,903
Lin. Reg.	4	1	4.0%	50.1%	60.0%	72.6%	7.5%	71.6%	154,02
Lin. Reg.	5	1	4.1%	49.9%	59.7%	72.2%	7.8%	71.2%	156,024
Lin. Reg.	6	1	4.1%	50.1%	60.0%	72.6%	7.5%	71.6%	153,388
Lin. Reg.	6	2	4.0%	50.2%	60.1%	72.8%	7.4%	71.8%	152,253
Lin. Reg.	6	3	4.2%	50.3%	60.3%	73.0%	7.1%	72.1%	151,556
Lin. Reg.	6	4	4.0%	50.1%	60.0%	72.6%	7.5%	71.6%	153,419
Lin. Reg.	6	5	4.1%	49.9%	59.8%	72.5%	7.5%	71.5%	153,521
Lin. Reg.	6	6	4.1%	50.7%	60.6%	73.2%	7.1%	72.3%	150,152
Lin. Reg.	6	7	4.1%	50.2%	60.2%	72.9%	7.2%	72.0%	152,529
Hurdle	1	1	4.0%	49.9%	59.7%	72.2%	7.9%	71.0%	156,184
Hurdle	2	1	4.0%	49.8%	59.6%	72.2%	7.8%	71.1%	155,041
Hurdle	3	1	4.0%	49.8%	59.6%	72.2%	7.8%	71.1%	155,041
Hurdle	4	1	4.0%	49.9%	59.7%	72.2%	7.9%	71.1%	155,927
Hurdle	5	1	4.0%	49.9%	59.6%	72.2%	7.8%	71.1%	154,956
Hurdle	6	1	4.1%	50.2%	60.1%	72.7%	7.5%	71.7%	153,229
Hurdle	6	2	4.0%	50.3%	60.2%	72.8%	7.3%	71.9%	152,350
Hurdle	6	3	4.2%	50.4%	60.4%	73.1%	7.1%	72.2%	150,655
Hurdle	6	4	4.0%	50.2%	60.1%	72.7%	7.4%	71.7%	153,363
Hurdle	6	5	4.1%	49.9%	59.9%	72.5%	7.5%	71.5%	153,888
Hurdle	6	6	4.2%	50.6%	60.5%	73.1%	7.2%	72.2%	150,887
Hurdle	6	7	4.1%	50.3%	60.3%	72.9%	7.2%	72.1%	152,251

A.2 The DWA methodology

The Distributional Wealth Account (DWA) statistics developed by the ECB Expert Group on Distributional Financial Accounts provide distributional information on household wealth since 2010, including outstanding amounts of financial and non-financial instruments by net wealth decile and several inequality indicators, like the Gini coefficient and the wealth share of the top 10 per cent. The dataset is not yet publicly available since it is still under development.

The ESCB methodology to produce DWA integrates microeconomic and macroeconomic data based on different sources: HFCS; "World's Billionaires List", published by Forbes; macroeconomic aggregates coming from national accounts. This process involves several adjustments and estimations.

Variables collected in the HFCS are matched with the definitions of the national accounts. Due to conceptual issues and poor comparability, some instruments (e.g. currency, pension entitlements, other accounts) were not included. Nonetheless, included instruments cover more than 86% of the total households' assets and liabilities.

Then, the full reconciliation of the totals derived from the surveys (using sampling weights) with the ones coming from national accounts is achieved through four different steps: first, survey observations are adjusted to take into account the bias deriving from *zero-reporting* and *under-reporting*. In particular, the procedure focuses on identifying outlier observations on deposits, i.e. when deposit holdings are very small compared to household income (income criterion) and/or the share of the household portfolio held as deposits is too small (asset criterion), and replace them with the average values by income class. The second step addresses the well-known issue of poor coverage of the wealthiest households in surveys like the HFCS. The correction is based on the key assumption that the right tail of the wealth distribution follows a Pareto distribution. The rich list from Forbes is added to the sample and used to estimate the Pareto tail distribution parameters. Synthetic households sampled from the Pareto tail, with wealth bounded between the HFCS' richest households and the rich list's poorest ones, complement the survey sample. Lastly, a proportional allocation is performed, i.e. for each instrument the remaining gap between the Financial Accounts total to the adjusted HFCS total is allocated proportionally to all households.

Following these adjustments, microdata are then interpolated and extrapolated based on the information derived from the quarterly national accounts. This allows obtaining quarterly time series on the distribution of household wealth.