IARIW 2023

IARIW – BANK OF ITALY 2023

Wednesday, March 29 - Saturday, April 1

# Combining Survey and Administrative Data to Estimate the Distribution of Household's Deposits

Andrea Neri
(Bank of Italy)
andrea.neri@bancaditalia.it

Matteo Spuri
(Bank of Italy)
Matteo.Spuri@bancaditalia.it

Francesco Vercelli
(Bank of Italy)
Francesco.Vercelli@bancaditalia.it

Paper prepared for the Conference on Central Banks, Financial Markets, and Inequality
March 29 – April 1, 2023

Session 8: Advancing Statistical Methods

Time: Friday, March 31, 2023 [17:00-18:00 PM CEST]

# COMBINING SURVEY AND ADMINISTRATIVE DATA TO ESTIMATE THE DISTRIBUTION OF HOUSEHOLDS' DEPOSITS

**Andrea Neri, Matteo Spuri and Francesco Vercelli [1]**

March 3, 2023

## 1. Introduction

In the last years, several initiatives have been launched to combine survey data with macroeconomic aggregates coming 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 (DNA).

In 2015, the European Central Bank created 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 (Ahnert et al., 2020). The work was continued in 2019 by 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 for the Euro area as a whole (Engel et al., 2021; Cantarella et al., 2021). Depending on the availability of country-specific sources, each country may enrich the general method in order to improve the quality of the estimates. One of the preliminary and necessary steps to produce the Distributional Wealth Accounts (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 very rich households and the reticence of respondents to report truly their incomes or their assets. 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. This requires the development of a methodology to redistribute the missing wealth (i.e., the difference between totals from survey data

[1] Bank of Italy, Directorate General for Economics, Statistics and Research.
The opinions expressed and conclusions drawn are those of the authors and do not necessarily reflect the views of the Bank of Italy and of the Eurosystem.

and national accounts) among the households in the survey. In the absence of reliable external information, such as administrative records, assumptions must be adopted. The DWA procedure developed by the ECB 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). Essentially, in absence of external information, the ECB adjustments on deposits are based on the identification of 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 particular, we exploit the aggregate information coming from supervisory data and administrative records relating to fiscal individual income, housing wealth, and debt that are 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 and the impacts on the DWA estimates; finally, Section 4 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 consists of a probabilistic sample of around 8,000 households selected from population registers. Its main focus is the collection of 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.

Starting from 2008, 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 the very rich households in the sample, and the reticence of respondents to provide correct information on issues that are 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 of which are specific on the subject, has revealed considerable reluctance on the part of families to provide information on the ownership of financial assets (...). For savings and income, collaboration of respondents is generally better, being less the aversion to provide data on flows than on stocks”.

Following this first analysis, many other studies have focused on the measurement of 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 Gini concentration index is not significantly affected by the adjustment. 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 percent 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 percent of the aggregate. The adjustment was larger for households whose head is old or poorly educated. The paper also adjusted financial liabilities, whose corrected values were on average about 40 percent higher. Neri and Ranalli (2012), using the results of a telephone survey conducted on SHIW non-respondents, reported greater 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 a 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 the Gini concentration indexes of income and wealth.

## 2.2 Administrative data

Administrative data on households' balance sheets do exist in almost all European countries. Yet, only a few HFCS countries make substantial use of them for the survey.  The main challenge is limited access because of legal, institutional, and practical constraints. In this paper, we use two different sets of administrative data: the first relates to registers that are linked to the survey by individual identifiers. These are 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. The second type of register data consists of aggregate banking supervisory reports that are used 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

Administrative records (AR) from tax files and cadastral register relative to 2016 are available thanks to a memorandum of understanding signed between Banca d'Italia and the Ministry of Economics and Finance (MEF). The agreement foresees the linkage of register data on a sample of individuals selected by Banca d'Italia through fiscal identifiers and limits its usage to specific goals. Thanks to this linkage we are able to reconstruct for each household in the SHIW the net income resulting from tax records, and the number of its real estate properties as well as their cadastral value. We obtain administrative data on household debt from the Bank of Italy's Credit Register, which contains detailed information on household credit relationships with intermediaries operating in Italy. We then compare AR data with the corresponding information collected in the survey to identify subsets of respondents that may be considered highly reliable.

The available information allows defining "highly reliable" households according to different criteria. Given the well-known under-reporting issue, we expect that survey incomes are generally equal or lower than administrative ones: the lower the incomes are, the less reliable the observations are. Therefore, in our first definition, we consider less reliable those households whose survey incomes are at least 5 percent lower than the administrative ones. The second definition is a little less restrictive and considers as not reliable those households with survey incomes at least 10 percent lower than the administrative ones. Both survey and administrative data provide information on the number of real estate properties (which are expressed as percentage points, depending on the held share of property). In our third definition, we consider as non reliable those households who hold, according to administrative data, at least one property in excess with respect to the number declared in the survey. An alternative way to distinguish highly and less reliable households may be ground in the identification of outlier observations. In the ECB's procedure for estimating the DWA, for example, households are considered outliers when deposit holdings are very small compared to household income (income criterion) and/or the share of household portfolio held as deposits is too small (asset criterion). Therefore, in our fourth definition we include an income criterion for outlier detection: we consider as less reliable households with survey incomes less than 90% of administrative ones and with deposits less than 10% of monthly income; however, this second condition does not hold for households with very low annual income (less than €10,000) and with credit card debt. In our fifth definition, we add the ECB asset criterion for outlier detection, i.e. households whose share of gross wealth held as deposits is lower than 0.8% are considered as less reliable; however, this second condition does not hold for households with overdraft credit, mortgage debt and null gross wealth. Our sixth definition does not depend on administrative data and simply mimics the ECB's method to identify outliers, considering as non-reliable those households which do not respect the income or the asset criterion. Finally, our seventh definition consider as reliable those households who turn out highly reliable according to at least 2 of definitions among the second, the third and the sixth ones. Table 1 reports the number of highly reliable households according to the different definitions.

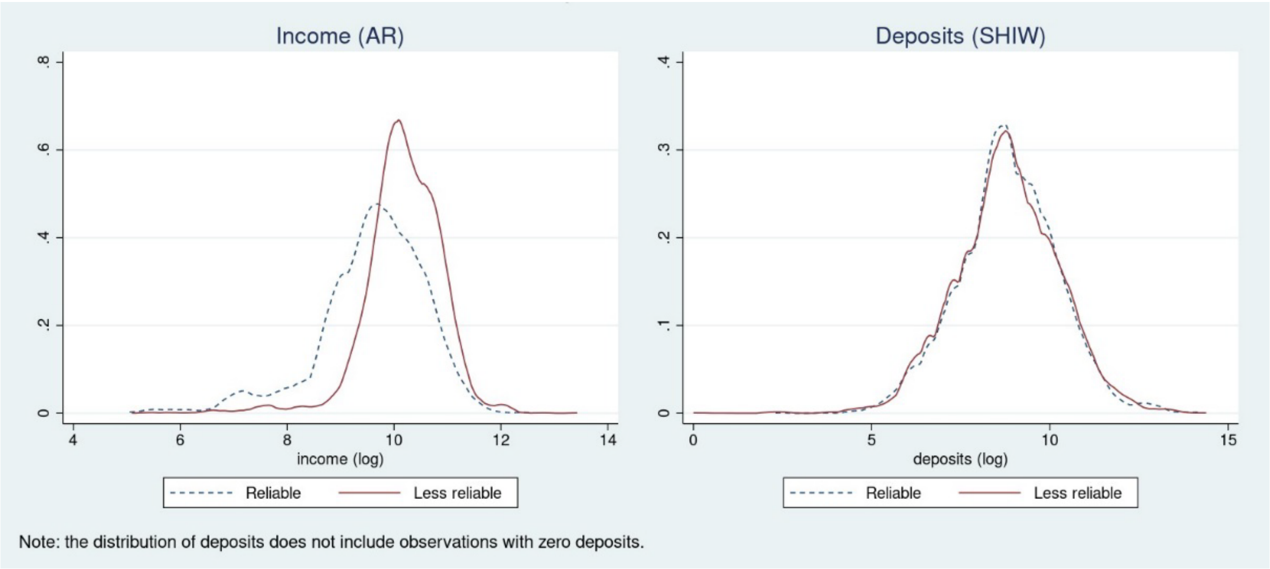
Table 1: Number of highly reliable households in the SHIW, according to different definitions.

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

The left panel of Figure 1 compares the distribution of AR incomes between highly reliable and less reliable households. Less reliable households display higher incomes, showing that under-reporting is stronger in the upper part of the distribution. Instead, as reported in the bottom panel, the two groups show similar distributions of deposits.[2] 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. The fact that less reliable households display similar levels of deposits but earn higher incomes also emerges in Figure A.1. Moreover, both groups include households declaring null deposits.

Figure 1: Distributions of incomes and deposits.

(*Highly reliable households: SHIW income >0.95 * AR income*)



Note: the distribution of deposits does not include observations with zero deposits.

### 2.2.2 Banking Supervisory Reports

Italian Banking Supervisory Reports (BSR) include some distributional information on deposits held by households. 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

[2] The right panel with the distribution of deposits does not include households with zero deposits. However, the two distributions would be very similar also including those households

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 her personal fiscal code. If a client holds more than one checking account at the same bank, she is assigned to the range corresponding to the overall amount of deposits held at the bank. However, joint accounts are not split between holders but considered as a different client. For example, if two clients have one bank account each and one joint account, the bank registers three different clients. Therefore, although the supervisory statistics are formally based on the definition of client, the underlying concept is closer to the number of accounts.

Obviously, the unit of observation in banking statistics differs from the one in the SHIW and in the FA. A first departure is related to the statistical management of joint accounts, as we have mentioned above. 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 of divergence between banking statistics and SHIW/FA.[4] According to SHIW, households hold on average two bank accounts.[5]

The number of clients holding deposits according to the BSR are reported in Table A.1. Unfortunately, these data cannot be used as reliable estimates on the number of households with deposits higher than a certain threshold. For example, the number of clients in the richest class does not represent neither an upper nor a lower bound for the number of people with at least €500,000 of financial wealth.[6] Therefore, the number of clients by asset range is not very informative for the

[3] The BSR also contain data on the overall outstanding amounts at market value of 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 on 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.

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

[5] However, the survey does not distinguish between checking and securities accounts, so we do not know the average number of either checking or securities accounts per household: we just know that it must be lower than two.

[6] Suppose, for example, that an individual holds €1 million of deposits. If she splits them into 3 equally-sized amounts and deposit them at 3 different banks, she would be registered as 3 different people in the second richest class: so we

construction of the DWA.

Table 2: 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 |

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

On the contrary, data on the outstanding amounts of deposits by asset range, reported in Table 2, can be useful.[7] For example, we observe that the class of clients with more than €500,000 holds around €90 billion in 2016, which correspond to about 9% of the overall amount of deposits. Table 3 compares aggregate estimates based on the SHIW with the distributional information from BSR. The total obtained in the SHIW is less than 40% of the BSR aggregates.[8] The shares of deposits held by the two wealthiest classes are quite similar, whereas the differences are remarkable for the first 3 classes. To be precise, the BSR values allow identifying a lower bound of the overall amount of deposits held by households with deposits over a certain threshold. For example, we know that 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 whose deposits are larger 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 applied the SHIW distribution to the BSR total (see the last two columns of Table 3), we would end up with an overall

would underestimate the number of the richest households. Instead, if she splits into 2 equally-sized amounts, she would be registered as 2 different people in the richest class: we would overestimate the number of the rich.

[7] 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.

[8] 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%.

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 3: Total deposits by asset range: BSR and SHIW

(*year=2016; millions of euros and percent*)

| | 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. Methods

The method we propose to adjust deposits in the DWA procedure consists of two steps:

1) in the first one we select a subset of highly reliable households comparing individual level administrative records and survey data; 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 impute the predicted values when they are larger than the values observed in the survey.

2) in the second step we calibrate the imputed results to the aggregate statistics from banking supervisory reports.

### 3.1 Deposit adjustments based on individual administrative records

As explained in Section 2.2.1, by comparing AR and survey data we can select a subset of highly reliable households. We aim at estimating a model for predicting deposits of highly reliable households. First, we run the following linear regression model:

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

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 $i$, $\epsilon_i$ the idiosyncratic error. The covariates used in the estimates are:

- wages, pensions, self-employed incomes, profits and rents (AR);
- real estate (AR);
- loans (AR);
- financial assets, other than deposits (SHIW);
- expenditures using banknotes (SHIW);
- durables and non-durable consumption (SHIW);
- overdraft credit and credit card debt (SHIW);
- savings (SHIW);
- age of the head of the household (SHIW);
- geographical macroarea of residence (SHIW);
- household composition (SHIW);
- sector of occupation of the respondent (SHIW)

All income and financial variables are expressed in log terms.

Table 4 reports the estimates obtained using different sets of covariates. 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 AR), holding strictly positive amounts of deposits. We are mostly interested in the ability of the model to predict the amount of deposits. In order to select the model with the highest prediction performance, we perform a 10-fold cross-validation and we compute the average Root Mean Squared Error (RMSE) across folds.[9] The first model includes a small subset of regressors: the overall amount of incomes 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. As expected, all the coefficients display a positive sign, although the one referred to loans is not statistically significant.[10] In the second model we include separately wages, pensions, and self-employed income, profits and rents: the average RMSE slightly increases, suggesting a preference for the first model. Interestingly, the coefficient on wages is close to zero and not statistically significant, whereas the other income components have a positive impact on

[9] 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. At the end, we take the average of the RMSEs obtained at each step. The lowest the average RMSE, the better the model.

[10] We provide only a quick description of the estimated coefficients. Clearly, we are interested in the prediction performance of the model, not on the economic interpretation of the coefficients.

deposits. In the third model we include covariates related to expenditures, durables and debts other than loans (overdraft credit and credit card debt). The coefficient on overdraft debt is negative as expected; the coefficient on credit card debt is positive but not statistically significant. The average RMSE decreases from model 1 to model 3 from 1.32 to 1.30. In model 4 we split again incomes into their components and, this time, there is a slight decrease in the average RMSE. In the fifth model, we exclude variables on expenditures and durables and we include overall savings: the flow is positively correlated with the level of deposits, but the average RMSE increases. In the sixth model, we include 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. We prefer the sixth model since it displays the lowest average RMSE.

Then, we use the estimated coefficients 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).[11] We assign the predicted values to less reliable households when they are higher than the observed deposits. Table 5 reports the results for different model specifications, showing in column "Coverage" the ratio between aggregate deposits obtained from micro data after the adjustment and financial accounts aggregates. The table also displays several inequality indicators obtained at the end of the DWA procedure, using different deposit adjustment methods.[12] The first line shows the results from the base ECB adjustment method, which allows achieving a coverage ratio of 44.7%. The coverage obtained using the regressions of Table 4 (linear regressions, using the first definition of reliable households) is higher, ranging from 48.9% in model 4 to 52.1% in model 1. Our preferred model, i.e. the sixth one, displays a coverage equal to 49.8%. Inequality is slightly lower using the regression methods than the ECB one. The share of net wealth held by the top 5% of richest households declines from 50.7% in the ECB method to 50.1% in our preferred model and the Gini coefficient reduces from 72.3% to 71.6%. The median wealth rises from €149,000 to €153,000. Figure A.2 graphically shows how individual data on deposits increase due to model predictions (sixth model). As reported in Figure 2, the density function of deposits shifts towards the right side using the ECB method, but the shift is more pronounced when using a regression model based on administrative data. Figure 3 shows how the increase of aggregate deposits is distributed along the pre-adjustment distribution of deposits. The adjustment in the ECB

[11] Alternatively, under the assumption of normal distribution of the error term, we could construct a correction term by exponentiating the mean squared error obtained from the regression.

[12] See Appendix A.3 to a brief overview of the DWA procedure.

method generally concerns households with lower deposits than in the BI method, which identifies relevant under-reporting also for higher levels of declared deposits.

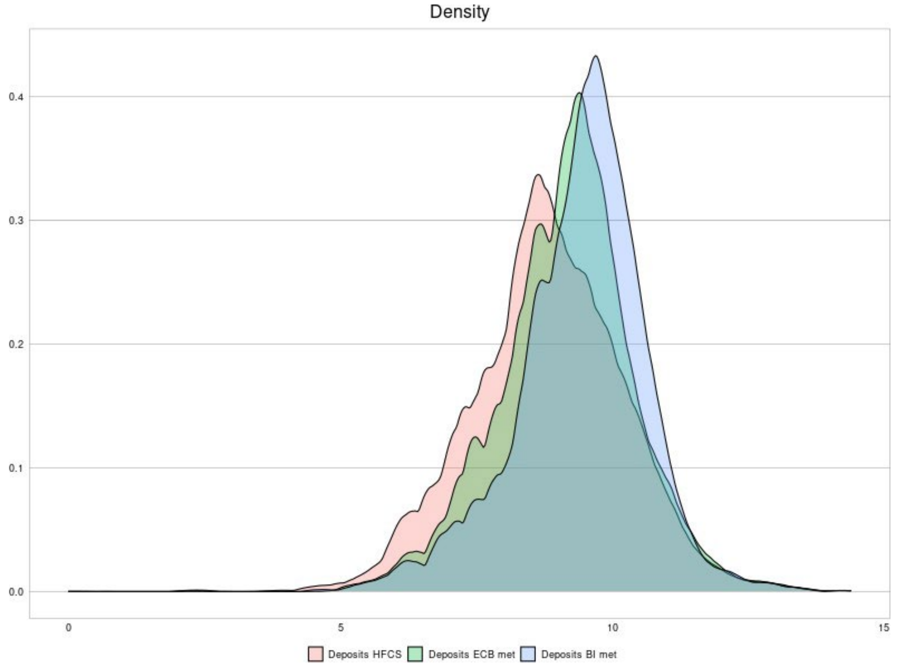
Table 4: Linear regression models with different set of covariates: estimates. Highly reliable households: SHIW income >0.95 * AR income.

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

Table 5: Different models: impact on coverage and inequality indicators.

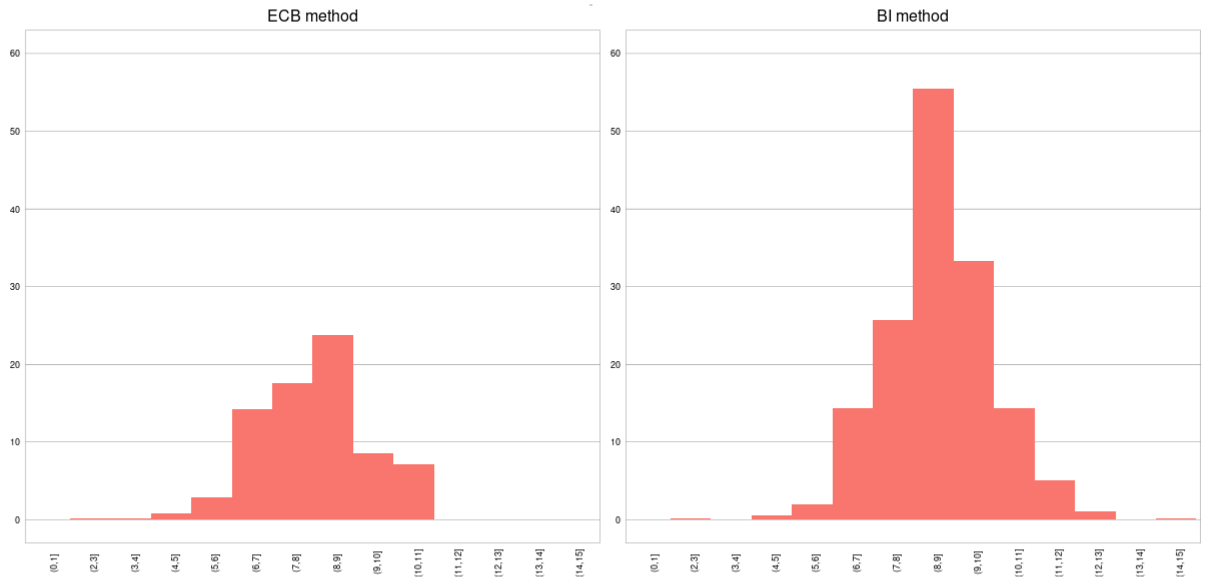
| Model | Var. Model | Reliab. Def. | Coverage | % of HH wealth > €1 mil. | Top 5% | Top 10% | Top 20% | Bottom 50% | Gini | Median Wealth |
|---|---|---|---|---|---|---|---|---|---|---|
| ECB | | | 44.7% | 4.2% | 50.7% | 60.7% | 73.3% | 7.1% | 72.3% | 149,338 |
| Lin. Reg. | 1 | 1 | 52.1% | 4.1% | 49.8% | 59.6% | 72.2% | 7.9% | 71.0% | 155,856 |
| Lin. Reg. | 2 | 1 | 51.4% | 4.0% | 49.9% | 59.6% | 72.2% | 7.8% | 71.1% | 154,976 |
| Lin. Reg. | 3 | 1 | 49.6% | 4.1% | 50.0% | 59.9% | 72.5% | 7.6% | 71.4% | 153,903 |
| Lin. Reg. | 4 | 1 | 48.9% | 4.0% | 50.1% | 60.0% | 72.6% | 7.5% | 71.6% | 154,02 |
| Lin. Reg. | 5 | 1 | 50.9% | 4.1% | 49.9% | 59.7% | 72.2% | 7.8% | 71.2% | 156,024 |
| Lin. Reg. | 6 | 1 | 49.8% | 4.1% | 50.1% | 60.0% | 72.6% | 7.5% | 71.6% | 153,388 |
| Lin. Reg. | 6 | 2 | 46.9% | 4.0% | 50.2% | 60.1% | 72.8% | 7.4% | 71.8% | 152,253 |
| Lin. Reg. | 6 | 3 | 46.6% | 4.2% | 50.3% | 60.3% | 73.0% | 7.1% | 72.1% | 151,556 |
| Lin. Reg. | 6 | 4 | 48.0% | 4.0% | 50.1% | 60.0% | 72.6% | 7.5% | 71.6% | 153,419 |
| Lin. Reg. | 6 | 5 | 51.3% | 4.1% | 49.9% | 59.8% | 72.5% | 7.5% | 71.5% | 153,521 |
| Lin. Reg. | 6 | 6 | 41.7% | 4.1% | 50.7% | 60.6% | 73.2% | 7.1% | 72.3% | 150,152 |
| Lin. Reg. | 6 | 7 | 46.7% | 4.1% | 50.2% | 60.2% | 72.9% | 7.2% | 72.0% | 152,529 |
| Hurdle | 1 | 1 | 50.5% | 4.0% | 49.9% | 59.7% | 72.2% | 7.9% | 71.0% | 156,184 |
| Hurdle | 2 | 1 | 51.5% | 4.0% | 49.8% | 59.6% | 72.2% | 7.8% | 71.1% | 155,041 |
| Hurdle | 3 | 1 | 51.5% | 4.0% | 49.8% | 59.6% | 72.2% | 7.8% | 71.1% | 155,041 |
| Hurdle | 4 | 1 | 50.5% | 4.0% | 49.9% | 59.7% | 72.2% | 7.9% | 71.1% | 155,927 |
| Hurdle | 5 | 1 | 51.7% | 4.0% | 49.9% | 59.6% | 72.2% | 7.8% | 71.1% | 154,956 |
| Hurdle | 6 | 1 | 48.8% | 4.1% | 50.2% | 60.1% | 72.7% | 7.5% | 71.7% | 153,229 |
| Hurdle | 6 | 2 | 46.1% | 4.0% | 50.3% | 60.2% | 72.8% | 7.3% | 71.9% | 152,350 |
| Hurdle | 6 | 3 | 46.2% | 4.2% | 50.4% | 60.4% | 73.1% | 7.1% | 72.2% | 150,655 |
| Hurdle | 6 | 4 | 47.1% | 4.0% | 50.2% | 60.1% | 72.7% | 7.4% | 71.7% | 153,363 |
| Hurdle | 6 | 5 | 50.6% | 4.1% | 49.9% | 59.9% | 72.5% | 7.5% | 71.5% | 153,888 |
| Hurdle | 6 | 6 | 42.0% | 4.2% | 50.6% | 60.5% | 73.1% | 7.2% | 72.2% | 150,887 |
| Hurdle | 6 | 7 | 46.1% | 4.1% | 50.3% | 60.3% | 72.9% | 7.2% | 72.1% | 152,251 |

Figure 2: Distribution of deposits using different estimation methods (ECB and BI)



The estimates in Table 4 are obtained within the sample of households that are considered highly reliable according to the first definition (income in the SHIW is at least 95% of the one in the AR). For robustness check, in Table A.2 we run model 6 using all the definitions introduced in Section 2.2.1. The coefficients are generally quite stable across models using different definitions. In particular, the coefficient on income is always statistically significant, and ranges between 0.13 using definition 3 and 0.20 using definition 6. As shown in Table 5, most definitions are associated with a higher coverage than the one obtained with the ECB method. The fifth definition, which combines both a constraint based on AR-SHIW incomes comparison and the ECB's outlier detection criteria, produces the largest coverage (51.3%). Instead, coverage is lower than using the ECB method only when we apply definition 6 (41.7%), based on the ECB's outlier detection criteria on income and assets. Again, the inequality indicators are generally close to each other. Again, the main divergence relates to the sixth definition, which assign a larger share of deposits to the richest part of the distribution, leading to higher inequality.

Figure 3: Adjustments on deposits using different estimation methods (ECB and BI)



As emerges in Figure A.1, there are households declaring null deposits both within the highly and less reliable groups. It is plausible that some Italian households do not have a bank account. However, this corner solution (zero deposits) may imply biased results in the linear regression model estimated within the sample of highly reliable households.[13] Hurdle models represent an alternative to linear regressions and allow treating corner solutions as observed instead of censored (Cragg, 1971). They combine two models: 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 non-null. 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 in order 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. In the first model, the selection equation includes income and real estates, whereas the outcome equation also includes financial assets and loans. In the second model, the income variable of the outcome equation is split into its main components (wages, pensions and other sources of income); in the third model, income is spit

[13] For this reason, in the base linear regressions of Tables 4 and A.2 we exclude households with null deposits.

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

into its components also in the selection equations. The Pseudo-R$^2$ only slightly increases and the average RMSE is stable. In models 4-6 we add demographics to the selection equation (geographical residence, household composition and sector of occupation of the respondent) and we present different sets of covariates for the outcome equation. The sixth model, which include the larger number explanatory variables, displays the best performance, with the highest Pseudo-R2 and the lowest average RMSE from the 10-fold cross validation.

According to model 6, the coverage of deposits stands at 48.8% after correcting the data of the least reliable observations (Table 5). Therefore, the coverage is just slightly lower than using the linear regression estimates (model 6) and inequality indicators are very close. Figure A.3 graphically shows how individual data on deposits increase due to model prediction. Differently from the results of the linear regression model (Figure A.2), the hurdle model allows for zero values in the prediction of deposits. However, there are few observations with null predicted deposits and this explains the fact that the two method delivers similar aggregate results. The other specifications of the hurdle model predict higher corrections, with coverage ranging between 50.5% and 51.7%. Since these corrections increase deposits in the middle part of the distribution, the Gini coefficient turns nearly one percentage point lower than in the sixth specification.

In Table A.4 we estimate the Hurdle model (sixth specification) using all the different definitions of reliable respondents. As in the linear regression model, the estimates based on the sixth definition (ECB outlier criteria on income and assets) display the lowest coverage (42%), while the coverage for the other estimates ranges between 46.1 and 50.2%. The inequality indicators are quite stable: the Gini coefficient ranges from 71.5 to 72.2%.

Although hurdle models are attractive for the possibility of considering corner solutions, they display some shortcomings, which may be relevant in the compilation of DWA statistics. First of all, there is no guarantee that the estimation through MLE converges. Second, they require extra assumptions to define a selection equation. Third, the Pseudo-R2 is quite low (in model 6 is less than 0.1), suggesting that the predictive ability of the model is not very satisfactory. Fourth, the final impact on coverage and inequality measures does not differ markedly from the regression models. Therefore, we prefer to follow a simple linear regression model.

### 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 amounts of deposits by asset range of clients' accounts.[15] These statistics cover nearly 90% of overall deposits from financial accounts. Since the procedure should match national accounts aggregates, BSR data are rescaled to official figures.

Comparing banking statistics and SHIW data by asset range is not straightforward because of the different unit of observation. For each household, the SHIW reports the number of deposit accounts, but this is not sufficient to split deposits by account. In the 2020 release of the SHIW, households were asked about the share of deposits held in their main deposit account. On average, households with more than one deposit account hold around 66% of their deposits in the main account. This percentage slightly declines with the increase of the number of deposits, but still remains over 63% for the households with 5 bank accounts. We use this information to estimate how the deposits of each household are distributed across bank accounts.[16] Then, we transform our dataset at the household level into a bank-account level database. Therefore, we are ready to apply calibration techniques in order to match deposits at the bank-account level with aggregate figures from BSR data.

Let $n$ be the observations on bank accounts. Let the vector $a \in \mathrm{R}^n$ denote the adjustment factors at the bank account level that allow reducing the gap with aggregate BSR data. Let the vector $w \in \mathrm{R}^n$ the set of survey weights and the vector $x \in \mathrm{R}^n$ the amount of deposits. Let $I_{i,c}$ be an indicator function that identifies to which asset range the bank account $i$ belongs among $C = 5$ asset ranges. Let $X_c$ the amount of aggregate deposits of asset range $c \in C$ from BSR data. Then, we solve the following problem:

[15] In Section 2.2.2 we have explained that the definition of asset ranges by *clients'* holding may be misleading because the unit of observation of these statistics is closer to deposit accounts than to clients.

[16] 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 in equal amounts across the remaining accounts. For the releases prior to 2020, we attribute 66% of deposits to the first account. When the information on the number of bank accounts is missing, we impute it using average values by estimated in the SHIW. In particular, we impute 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.

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

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

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

The estimated adjustment factors multiply the observed amount of deposits, without any change in sampling weights. To avoid excessive changes with respect to the original data, the adjustment factors are allowed to range from 0.5 to 10.[17] After convergence, each observation is associated to an adjustment factor, which may move the bank account to a different asset range, reducing the match with aggregate constraints. Therefore, we perform several iterations of the calibration procedure and we select the iteration step with the lowest mean squared error. We also put a penalty to observations moving from one class to another in order to preserve broadly the original distribution.

The calibration step enters the DWA procedure after the Pareto adjustment step and before the proportional allocation step.[18] At that point the coverage ratio of deposits is still lower than 50% for waves 1 and 2, around 60% for wave 3 and nearly 70% for wave 4. Therefore, the overall amount of deposits that can be assigned is quite relevant and produces a shift of the distribution to the right (Figure A.4). The empirical CDF tends to be more affected when the initial coverage gap is lower, like in waves 1 and 2, while the correction is less remarkable for wave 4. The graphs on the third column of Figure A.4 show that the stronger adjustments in terms of aggregate deposits concern the central part of the distribution.

Figure 4 compares aggregate deposits by asset range (at the account level) in the BSR data with those obtained: before the calibration step; in the final DWA estimates applying the calibration techniques; in the final DWA estimates using the standard methodology. In general, the calibration procedure guarantees a closer match with aggregates from banking statistics with respect to the standard methodology, which tends to overestimate the amount of deposits in the richest part of the distribution. For few combinations of wave and asset range, the calibration performs poorly (for example for the class €250-500,000 in waves 1 and 3).

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

[18] See Appendix A.3 to a brief overview of the DWA procedure.

Figure 5 shows the effect of the calibration step on some major indicators of inequality. The wealth share held by the top 5% decreases by nearly one percentage point in the second and in the third waves. At the same time, the wealth share held by the bottom 50 percent declines, especially since the third wave, so that the impact on the Gini coefficient is very small. The calibration procedure determines a strong increase of the share of deposits held by the ninth decile of the distribution of net wealth: the share stands at 18% whereas it is only 13% in the standard method.

Figure 4: Calibration techniques: comparison with administrative data

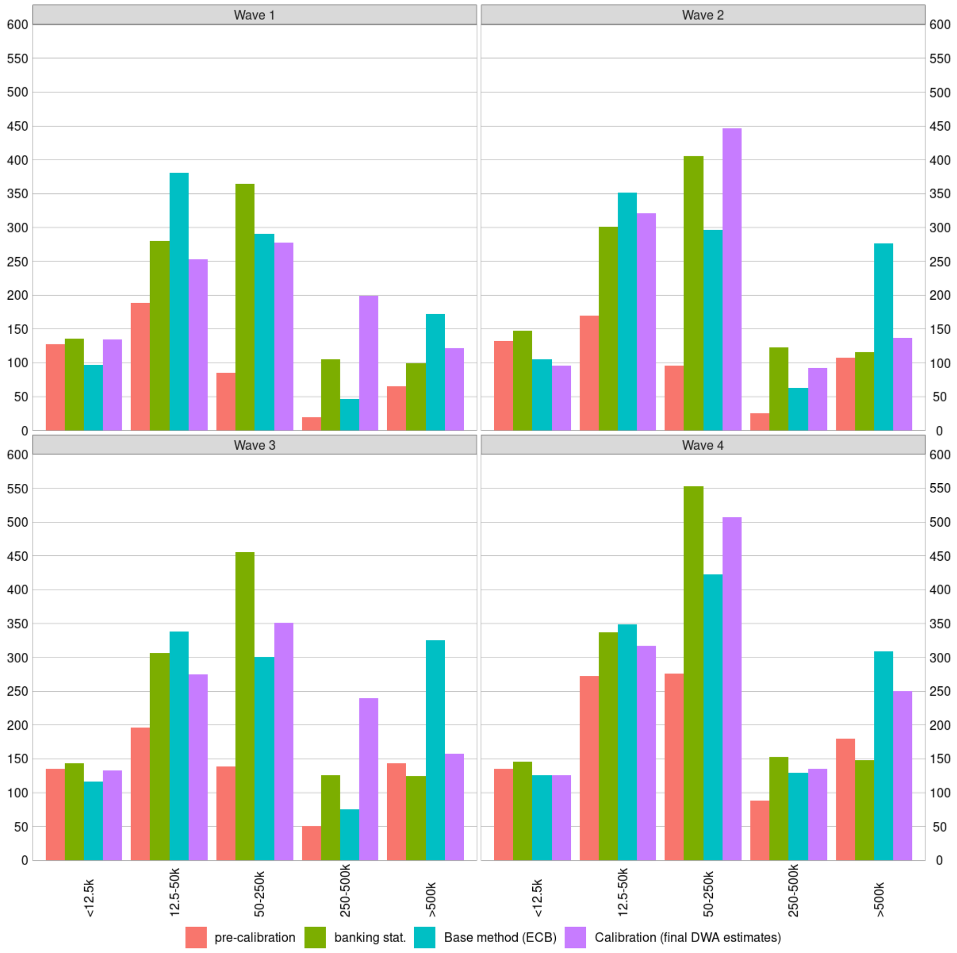
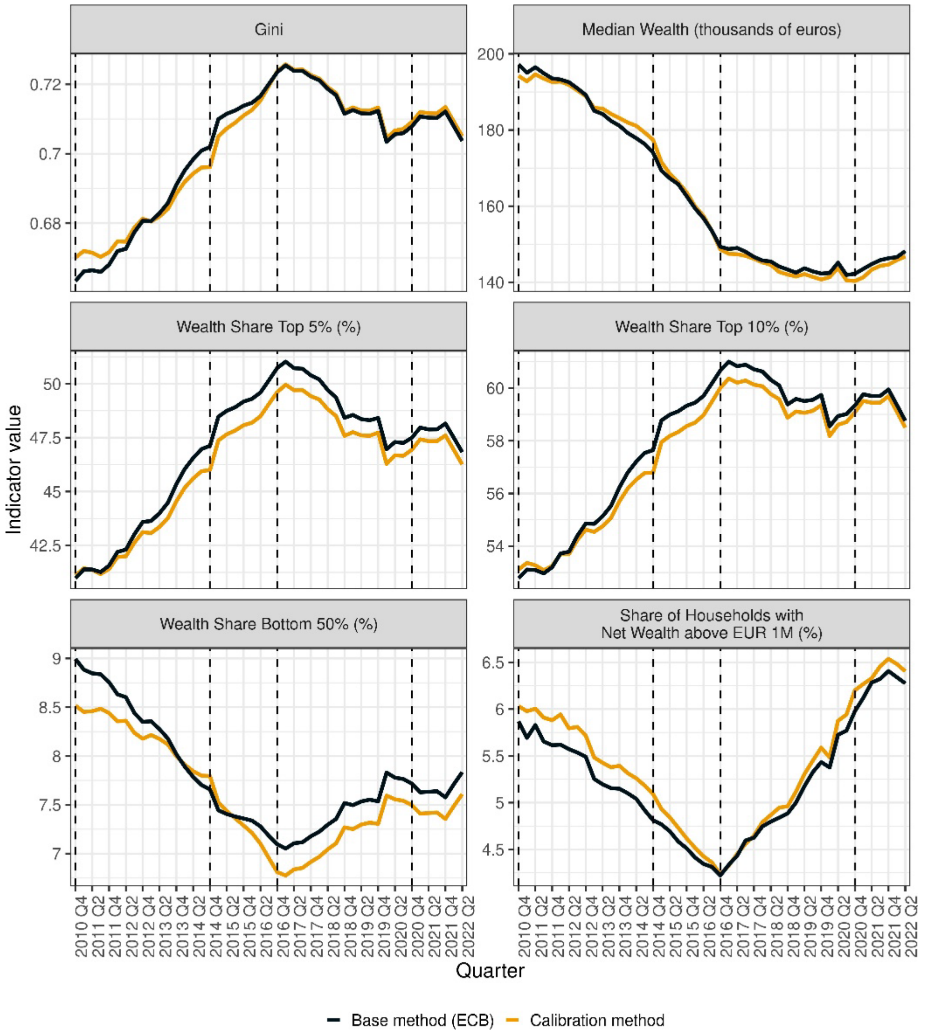


Figure 5: Calibration techniques: the impact on inequality indicators



## 4. Conclusions

The DWA statistics developed by the ECB Expert Group on Distributional Financial Accounts provide a comprehensive view on the distribution of household wealth by adjusting survey data to obtain aggregate figures coherent with national accounts. An important adjustment on survey data concerns deposits, since this instrument represents a significant share of household gross wealth and its coverage of national figures is low. Because of the lack of external information, the adjustment is based on the identification of 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. 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. The imputation method increases the coverage of aggregate deposits from 45 percent in the base ECB method to around 50 percent. The adjustment in the ECB method generally concerns households with lower deposits than in the BI method, which identifies relevant under-reporting also for higher levels of declared deposits. The Gini coefficient slightly decreases from 72.2 percent in the base model to 71.6 percent in the method that we propose.

We then make use of aggregate statistics from banking supervisory reports, which regard the outstanding amounts of deposits by asset range of clients' holdings. Using calibration techniques, we adjust survey observations to match aggregate information by asset range from supervisory reports. The calibration method determines a decline in the wealth share of the richest households as well as of those in the bottom 50 percent, while the share of the ninth decile increases. Overall the Gini coefficient remains quite stable.

Further extensions of the calibration techniques presented in this paper will be implemented in future research projects. Banking statistics by asset range are available on a semi-annual basis so that they can be used for improving the interpolation and extrapolation of the DWA quarterly time series when survey data are not available. 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.

## Bibliography

Ahnert, H., Kavonius, I. K., Honkkila, J., and Sola, P. (2020). Understanding household wealth: linking macro and micro data to produce distributional financial accounts. Statistics Paper Series 37, European Central Bank.

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

Cantarella, M., Neri, A., and Ranalli, G. (2021). Mind the wealth gap: a new allocation method to match micro and macro statistics on household wealth. Questioni di Economia e Finanza (Occasional Papers) 646, Bank of Italy, Economic Research and International Relations Area.

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., Cannari, L., Raimondi, G., and Rinaldi, A. (1990). Le attivit`a finanziarie delle famiglie italiane. Temi di discussione della Banca d'Italia.

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 house- holds' 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. (2021). Developing Reconciled Quarterly Distributional National Wealth – Insight into Inequality and Wealth Structures. mimeo, ECB.

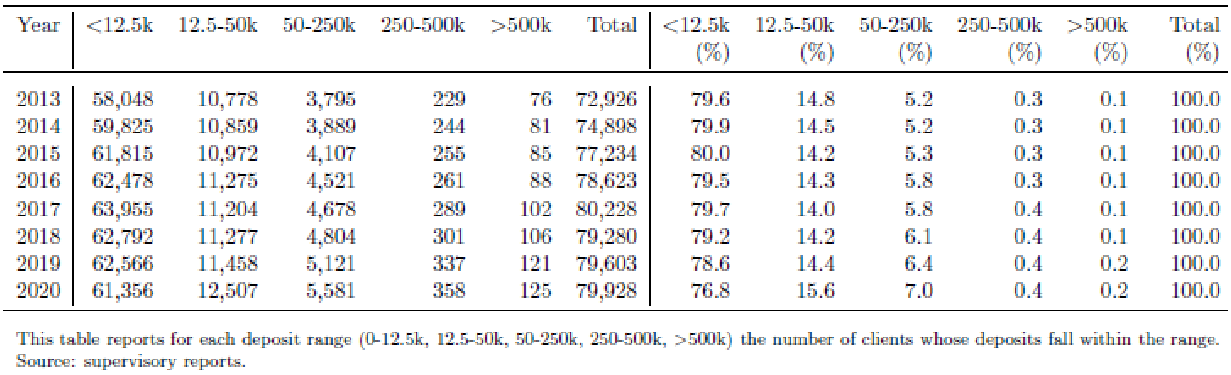
Neri, A. and Ranalli, M. G. (2012). To misreport or not to report? The measurement of household financial wealth. Temi di discussione (Economic working papers) 870, Bank of Italy, Economic Research and International Relations Area.

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

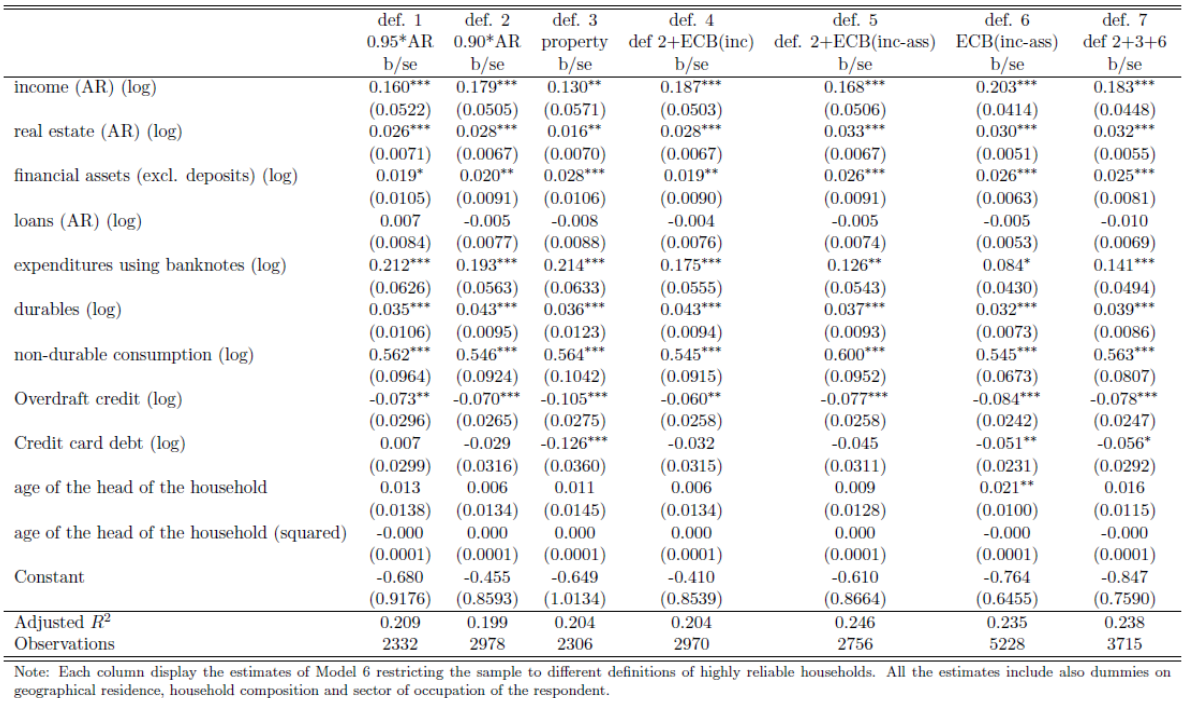
Table A.1 – Number of clients with checking accounts by asset range
*(annual data; thousands of clients and per cent)*



| Year | <12.5k | 12.5-50k | 50-250k | 250-500k | >500k | Total | <12.5k (%) | 12.5-50k (%) | 50-250k (%) | 250-500k (%) | >500k (%) | Total (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2013 | 58,048 | 10,778 | 3,795 | 229 | 76 | 72,926 | 79.6 | 14.8 | 5.2 | 0.3 | 0.1 | 100.0 |
| 2014 | 59,825 | 10,859 | 3,889 | 244 | 81 | 74,898 | 79.9 | 14.5 | 5.2 | 0.3 | 0.1 | 100.0 |
| 2015 | 61,815 | 10,972 | 4,107 | 255 | 85 | 77,234 | 80.0 | 14.2 | 5.3 | 0.3 | 0.1 | 100.0 |
| 2016 | 62,478 | 11,275 | 4,521 | 261 | 88 | 78,623 | 79.5 | 14.3 | 5.8 | 0.3 | 0.1 | 100.0 |
| 2017 | 63,955 | 11,204 | 4,678 | 289 | 102 | 80,228 | 79.7 | 14.0 | 5.8 | 0.4 | 0.1 | 100.0 |
| 2018 | 62,792 | 11,277 | 4,804 | 301 | 106 | 79,280 | 79.2 | 14.2 | 6.1 | 0.4 | 0.1 | 100.0 |
| 2019 | 62,566 | 11,458 | 5,121 | 337 | 121 | 79,603 | 78.6 | 14.4 | 6.4 | 0.4 | 0.2 | 100.0 |
| 2020 | 61,356 | 12,507 | 5,581 | 358 | 125 | 79,928 | 76.8 | 15.6 | 7.0 | 0.4 | 0.2 | 100.0 |

This table reports for each deposit range (0-12.5k, 12.5-50k, 50-250k, 250-500k, >500k) the number of clients whose deposits fall within the range. Source: supervisory reports.

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



| | def. 1<br>0.95*AR<br>b/se | def. 2<br>0.90*AR<br>b/se | def. 3<br>property<br>b/se | def. 4<br>def 2+ECB(inc)<br>b/se | def. 5<br>def. 2+ECB(inc-ass)<br>b/se | def. 6<br>ECB(inc-ass)<br>b/se | def. 7<br>def 2+3+6<br>b/se |
|---|---|---|---|---|---|---|---|
| income (AR) (log) | 0.160$^{***}$ | 0.179$^{***}$ | 0.130$^{**}$ | 0.187$^{***}$ | 0.168$^{***}$ | 0.203$^{***}$ | 0.183$^{***}$ |
| | (0.0522) | (0.0505) | (0.0571) | (0.0503) | (0.0506) | (0.0414) | (0.0448) |
| 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$^{***}$ | 0.026$^{***}$ | 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) | 0.212$^{***}$ | 0.193$^{***}$ | 0.214$^{***}$ | 0.175$^{***}$ | 0.126$^{**}$ | 0.084$^{*}$ | 0.141$^{***}$ |
| | (0.0626) | (0.0563) | (0.0633) | (0.0555) | (0.0543) | (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$^{***}$ | 0.564$^{***}$ | 0.545$^{***}$ | 0.600$^{***}$ | 0.545$^{***}$ | 0.563$^{***}$ |
| | (0.0964) | (0.0924) | (0.1042) | (0.0915) | (0.0952) | (0.0673) | (0.0807) |
| Overdraft credit (log) | -0.073$^{**}$ | -0.070$^{***}$ | -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 |
| | (0.0001) | (0.0001) | (0.0001) | (0.0001) | (0.0001) | (0.0001) | (0.0001) |
| Constant | -0.680 | -0.455 | -0.649 | -0.410 | -0.610 | -0.764 | -0.847 |
| | (0.9176) | (0.8593) | (1.0134) | (0.8539) | (0.8664) | (0.6455) | (0.7590) |
| 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 display the estimates of Model 6 restricting the sample to different definitions of highly reliable households. All the estimates include also dummies on geographical residence, household composition and sector of occupation of the respondent.

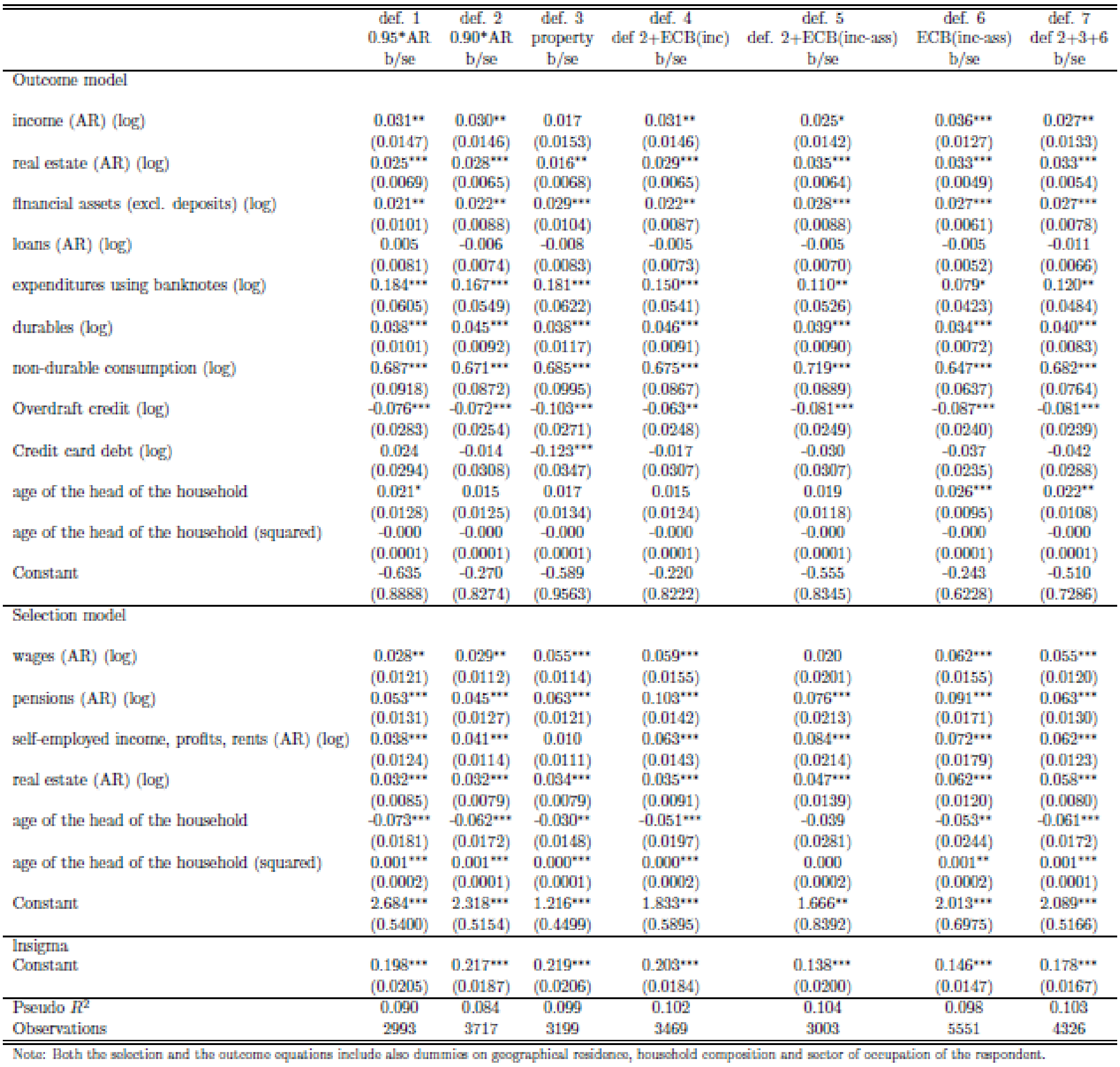
Table A.3 – Hurdle models with different set of covariates: estimates.

*Highly reliable households: SHIW income > 0.95 * AR income*

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | deposits (log) | deposits (log) | deposits (log) | deposits (log) | deposits (log) | deposits (log) |
| | b/se | b/se | b/se | b/se | b/se | 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 | 0.024 |
| | | | | | (0.0268) | (0.0294) |
| expenditures using banknotes (log) | | | | | | 0.184*** |
| | | | | | | (0.0605) |
| durables (log) | | | | | | 0.038*** |
| | | | | | | (0.0101) |
| non-durable consumption (log) | | | | | | 0.687*** |
| | | | | | | (0.0918) |
| 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*** | -0.635 |
| | (0.1401) | (0.0977) | (0.0977) | (0.1401) | (0.0982) | (0.8888) |
| Selection model | | | | | | |
| wages, pensions and rents (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*** | -0.073*** | -0.073*** |
| | | | | (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) |
| Codice ripartizione=2 | | | | 0.062 | 0.044 | 0.044 |
| | | | | (0.1442) | (0.1451) | (0.1451) |
| Codice ripartizione=3 | | | | -0.377*** | -0.396*** | -0.396*** |
| | | | | (0.1339) | (0.1354) | (0.1354) |
| Codice ripartizione=4 | | | | -0.696*** | -0.672*** | -0.672*** |
| | | | | (0.1219) | (0.1224) | (0.1224) |
| Codice ripartizione=5 | | | | -0.764*** | -0.766*** | -0.766*** |
| | | | | (0.1361) | (0.1364) | (0.1364) |
| Constant | 0.221** | 0.221** | 0.353*** | 2.448*** | 2.684*** | 2.684*** |
| | (0.1108) | (0.1108) | (0.0916) | (0.5435) | (0.5400) | (0.5400) |
| 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 $R^2$ | 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 display the estimates of Model 6 restricting the sample to different definitions of highly reliable households. 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.



| | def. 1 | def. 2 | def. 3 | def. 4 | def. 5 | def. 6 | def. 7 |
|---|---|---|---|---|---|---|---|
| | 0.95*AR | 0.90*AR | property | def 2+ECB(inc) | def. 2+ECB(inc-ass) | ECB(inc-ass) | def 2+3+6 |
| | b/se | b/se | b/se | b/se | b/se | b/se | b/se |
| Outcome 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) | (0.0142) | (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*** | 0.150*** | 0.110** | 0.079* | 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*** | 0.685*** | 0.675*** | 0.719*** | 0.647*** | 0.682*** |
| | (0.0918) | (0.0872) | (0.0995) | (0.0867) | (0.0889) | (0.0637) | (0.0764) |
| Overdraft credit (log) | -0.076*** | -0.072*** | -0.103*** | -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) | 0.024 | -0.014 | -0.123*** | -0.017 | -0.030 | -0.037 | -0.042 |
| | (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 | -0.635 | -0.270 | -0.589 | -0.220 | -0.555 | -0.243 | -0.510 |
| | (0.8888) | (0.8274) | (0.9563) | (0.8222) | (0.8345) | (0.6228) | (0.7286) |
| Selection model | | | | | | | |
| wages (AR) (log) | 0.028** | 0.029** | 0.055*** | 0.059*** | 0.020 | 0.062*** | 0.055*** |
| | (0.0121) | (0.0112) | (0.0114) | (0.0155) | (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 | 0.063*** | 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*** | 0.062*** | 0.058*** |
| | (0.0085) | (0.0079) | (0.0079) | (0.0091) | (0.0139) | (0.0120) | (0.0080) |
| age of the head of the household | -0.073*** | -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 | 2.684*** | 2.318*** | 1.216*** | 1.833*** | 1.666** | 2.013*** | 2.089*** |
| | (0.5400) | (0.5154) | (0.4499) | (0.5895) | (0.8392) | (0.6975) | (0.5166) |
| 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 $R^2$ | 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.

## A.2 List of Figures

Figure A.1

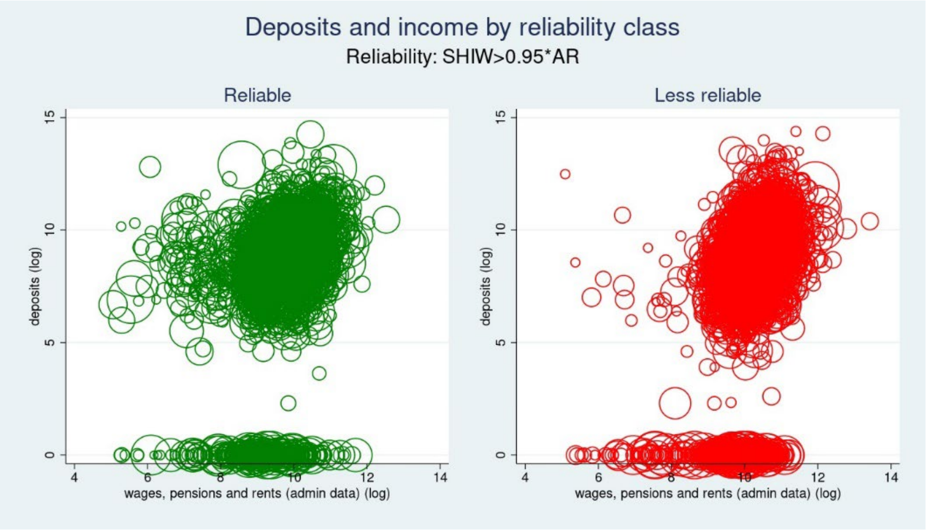


Figure A.2

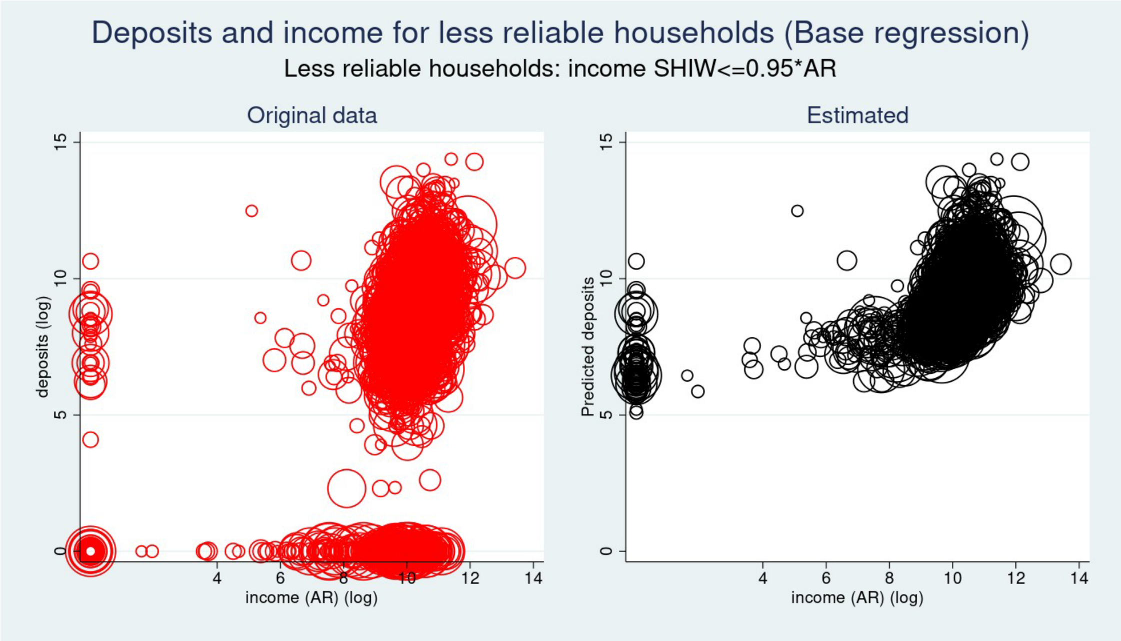


Figure A.3

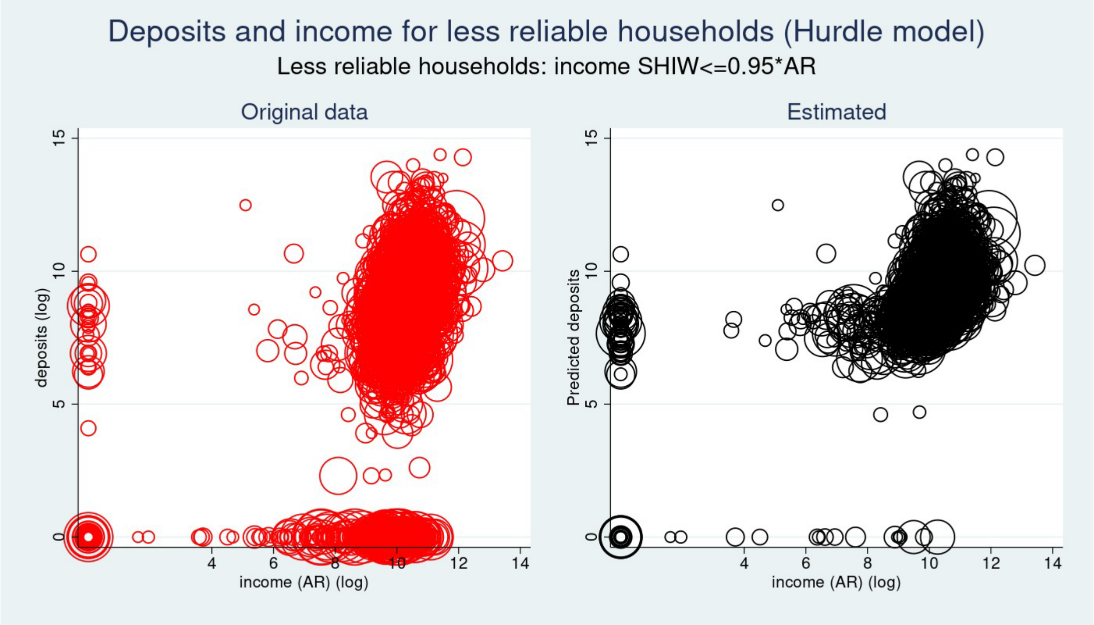
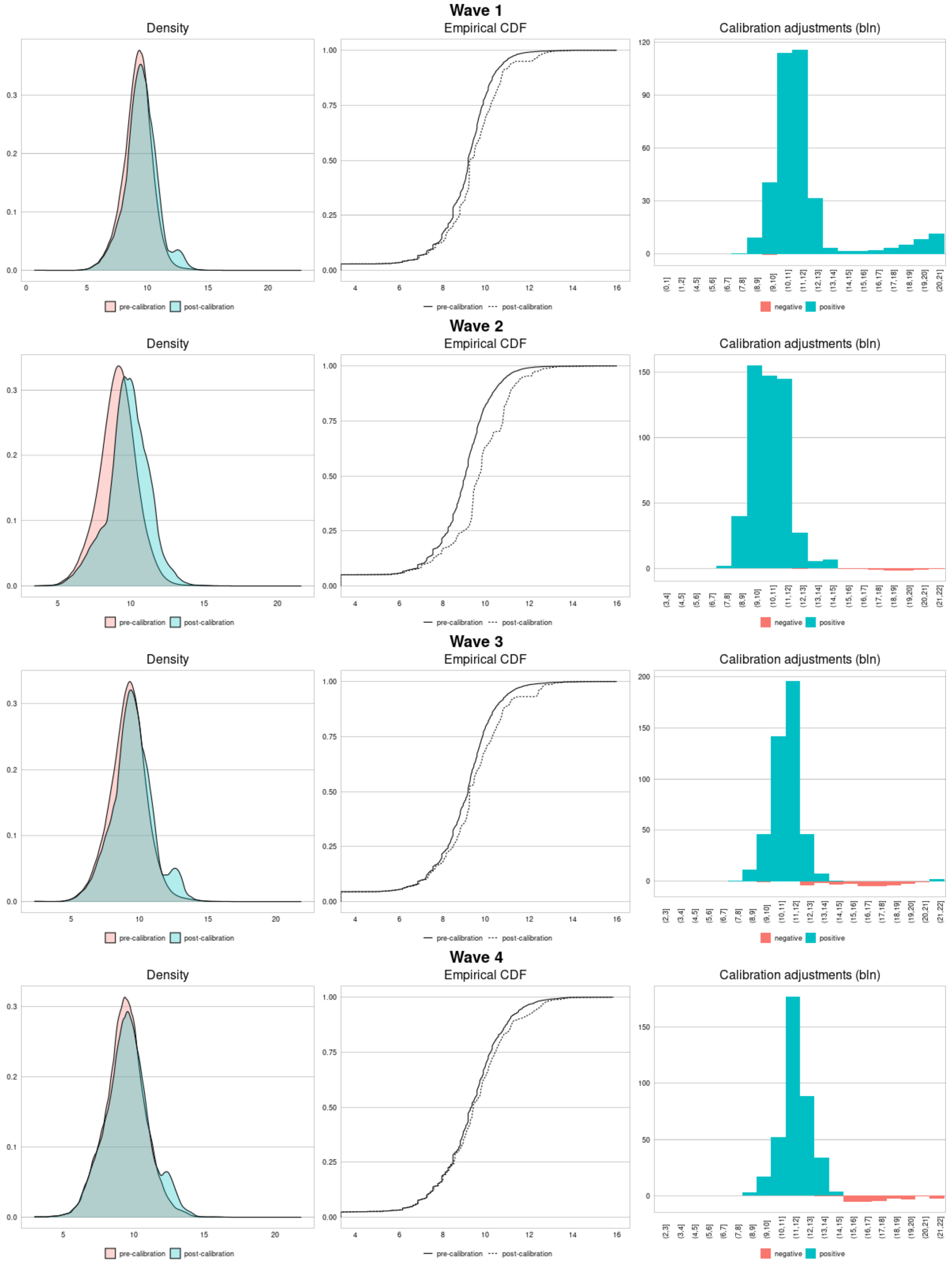

Deposits and income for less reliable households (Hurdle model)
Less reliable households: income SHIW<=0.95*AR
Original data
Estimated
deposits (log)
Predicted deposits
0
5
10
15
4
6
8
10
12
14
income (AR) (log)
income (AR) (log)

Figure A.4: Calibration techniques: the impact on the distribution of deposits



### A.3 DWA: the methodology developed by the ECB Expert Group on Distributional Financial Accounts

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 ECB 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 the 86% of the total of households' assets and liabilities.

Then, the full reconciliation of the totals derived from the surveys (by means of 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 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 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 deriving from the quarterly national accounts. This allows obtaining quarterly time series on the distribution of household wealth.