Are administrative records an alternative to sample surveys? The case of the Italian Survey on Household income and Wealth

Andrea Neri

Economic and Financial Statistics Department, Banca d'Italia, Rome, Italy

e-mail: andrea.neri@bancaditalia.it

Abstract

Since 1965, Banca d'Italia has conducted a survey of household income and wealth (SHIW) to study the economic behaviors of Italian households. In the recent years, new sources of information have become available such as tax records, data on the ownership and the value of dwellings, data relating to financial assets and liabilities held by households. In the paper, I discuss the experience of Banca d'Italia in the use of such administrative records for the analysis of household economic situations. I investigate the degree of "fitness for use" of any source of information using the following criteria: relevance, timelines, accuracy, accessibility, interpretability and coherence. I argue that administrative records cannot be considered as a viable alternative to survey data, at least in our case. Instead, a more useful research question is how to combine all the information available to fully exploit the benefits of each source.

Keywords   data quality, statistical matching, SHIW, administrative records

1. Introduction

The SHIW is a survey conducted by Banca d'Italia since 1965. The survey consists of a probabilistic sample of around 8,000 households that are representative of the Italian population. Its main focus is the collection of detailed information about household income, wealth and, to a lesser extend, consumption.

In recent years the survey has been integrated in international research projects such as the *Luxembourg Income Study* and the *Luxembourg Wealth Study*, whose aim is to produce a comparable cross-national Data Archive on household income and wealth. 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 finances and consumption in the Euro area (Survey on Household Finance and Consumption).

The survey is one of the tools Banca d'Italia uses to get a deep understanding of the economic situation of Italian households. This is necessary in order to perform its institutional tasks, the main one of which being to contribute to monetary and financial stability. In this context, the SHIW data have been used to understand how does personal consumption respond to wealth shocks (see for instance Paiella, 2004), or how do credit constraints affect consumption and investment (see for instance Magri, 2009).

In recent years, because of the financial crisis, it has become important to understand the implications of changes in household indebtedness for financial stability and the degree of vulnerability of the indebted households (see for instance D'Alessio and Iezzi, 2009).

The SHIW data are also widely used to provide advice to policy makers. They are used into micro-simulation models to understand the impact of policy choices. The bank has a benefit-tax model to simulate the effects of fiscal policies and a dynamic model whose main objective is to make predictions about the current generations' future pensions. Results based on the survey have been recently presented to the Italian parliament in order to explain which groups of the population are mostly affected by the current crisis.

The SHIW has also many external users that range from the general public to academia, from journalists to decision-makers. Micro data and a report are freely available through the web and usually become an important reference for the economic debate.

The comparison between the SHIW data and administrative data has always been one the ways Banca d'Italia has used to assess the quality of survey data. Tables 1 and 2 compare the survey-based estimates of the main components of household income and wealth with the corresponding aggregate administrative figures (for details see Banca d'Italia 2012).

The main discrepancies between the two sources are related to income for self-employment and the value of household financial wealth. The two sources differ by more than the sampling variability would predict.

Since part of the difference is likely to be due to measurement errors and nonresponse problems affecting the survey, during recent years the bank has been exploring the availability of different sources of information that can help in the study of the Italian economic situation by integrating the information coming from the survey.

2. Administrative data

At present, there are four sources of administrative data that are available to the Banca d'Italia.

A first source comes from the financial accounts (FA thereafter) that are produced by the Bank. The financial accounts record the yearly stocks and flows of a country's financial assets and liabilities, divided between the institutional sectors of the economy. They can be used as aggregated information on the household financial wealth.

A second source is the Central Credit Register (CCR). This data set contains information sent by banks and financial companies in Italy that are required to report to Banca d'Italia their bad debts and all their positions in excess of a given amount. The dataset contains information on the size of loans and on the kind and the value of real collateralizations provided and some basic demographic information.

A third source is a database created by the Italian Department of the Finance (MEF, hereafter) that combines administrative archives of the buildings owned by individuals with their average price of sale (based on the recent transactions) in a given area. Banca d'Italia has not a full access to this database but it may retrieve some aggregate information on household real wealth. On the contrary, the information on the minimum and the maximum square meter house price and the typology of the dwelling (villas, cottages, houses,.) in each `OMI' area is fully available (OMI data, hereafter).

The fourth source comes from the tax records. They contain all the statements of income in all their different forms: UNICO, 730, CUD. Also this microdata are not available to the bank, but it can access to some aggregate information about individuals' income.

It is worth mentioning that, at present, Banca d'Italia does not collect any personal identifier of the SHIW respondents. This is done to increase household's trust in the confidentiality of the information it provides. So, any use of administrative records cannot be based on an exact matching with survey data.

3. SHIW data versus administrative data

Statistics Canada (2009) provides a framework that could be used to contrast survey and administrative data. The quality of the information provided by a given source can be defined in terms of "fitness for use" and it can be evaluated using six criteria: *accuracy*, *relevance*, *timeliness*, *accessibility*, *interpretability* and *coherence*.

Accuracy refers to the degree to which data are close to the phenomenon they are meant to measure. It is widely accepted that survey data suffers from accuracy problems. This is particularly the case for surveys like the SHIW that aims at

collecting very sensitive information on household wealth and income. Previous research using the SHIW data has shown that unit nonresponse and measurement error are two major issues. Neri and Ranalli (2011) show that the consequences are particularly severe for the measurement of household financial assets. Very rich people have a lower propensity to participate in the survey and they tend to concentrate most of the total financial wealth. Moreover, the reporting of the financial assets owned and their value seems to be affected by a significant underreporting. On the income side, Neri and Zizza (2010) estimate that the reported average income is around 12 percent lower that the true one, while the corresponding relative standard error is about 1.4 percent.

As far as the administrative records mentioned above, there is hardly any study on data accuracy. A quite common view in the literature is to assume that they are error-free or that the amount of measurement error is negligible. In fact, the absence of evidence does not imply that this assumption holds true. Whenever there is a measurement process, there is often a measurement error. Let's consider the financial accounts for instance. Bonci et al. (2005) use the average amount of revisions each single item is subject to, as a proxy of accuracy. They find that some items like *deposits* are reliable, while other like the *shares* held by households are subject to higher revisions. As a matter of fact, this item is estimated as a residual, after the figures for other sectors are estimated. Therefore, the estimate of the share held by households is subject to all the errors of the other estimations. Also fiscal data are subjected to the measurement error (mainly in the form of underreporting due to tax evasion). Some indirect evidence is provided by Marino and Zizza (2011) who use the SHIW data (adjusted for measurement error) to estimate the amount of underreporting in fiscal data.

The second dimension of quality is *relevance* of statistical information which reflects the degree to which it meets the real needs of users.

In our situation it can be interpreted in terms of *coverage* and *variety* of analysis.

Coverage refers to the extent to which the source produces information for the whole Italian population. Even if we could combine all the administrative records together, we would not have a perfect coverage of the Italian population. For instance, financial accounts don't include the share of the population that does not hold a bank account (around 10% according to the survey data). The same holds for tax records which exclude those who don't fill out the tax form because their income is below a given threshold. The problem is that the missing part of the population cannot be considered as a random subsample of the Italian population. For instance, those who are not obligated to file taxes are likely to be poorer than the others.

As to survey data, they suffer from nonresponse which can have similar consequences to those of noncoverage. The main difference with respect to the previous situation is that some paradata information is collected during the fieldwork that can be used to mitigate the possible bias at the estimation stage.

The second aspect of *relevance* relates the variety of possible analysis. The main advantage of micro-level data is that they usually have a lot of additional information besides the variable of interest. This allows researchers to better understand the data and to investigate a large variety of topics, such as the study of the distributional aspects. For instance, the implications of changes in a tax reform cannot be adequately judged from aggregate data alone, given that it is important to identify in more detail the characteristics of the subgroups of the population that are going to benefit or pay because of the reform. Moreover, the SHIW survey is often used to collect information on specific topics of interest in a given moment. This is done by including special modules such as capital gains, inheritance, risk aversion, unpaid work, economic mobility, social capital, tax evasion, financial literacy.

A third dimension of quality is *timeliness* which refers to the time lag between the reference period of the statistic and its availability for the user. As to the survey, there is generally a gap of 10 months between the reference period of the main variables and

the moment in which edited micro-data becomes available. The internal administrative information (FA and CCR) has a shorter time lag since it is updated on regular basis. The same doesn't hold for the external source of information. For instance, in recent study, Neri and Monteduro (2013) compare SHIW data and the most updated OMI database and have to use data that refers to 2008 (the most recent year for which OMI data are available).

Another relevant aspect of data quality is *coherence* that reflects the degree to which it can be successfully brought together with other statistical information. Also from this point of view, the SHIW offers more advantages with respect to the other sources. Administrative record may not even guarantee comparability across time. For instance, the threshold for indebted households to be recorded in the Central Credit Register was lowered from 75,000 euros to 30,000 euros in December 2008. The same might hold for the tax threshold under which it is not necessary to file taxes. As a consequence, CCR and tax records may be formed of different segments of populations in different years.

From the other side, the core questionnaire of the SHIW has not changed since 1987. Moreover, the survey is part of international projects and it enables researchers to make cross-country comparisons.

The last two elements are *accessibility* and *interpretability.* The first refers to the ease with which information can be obtained from the producer. The second relates the auxiliary information that is provided to users in order to help them to interpret and use data appropriately. The main problem is that administrative data are not created for doing research, they are meant to serve other purposes. This implies that are not usually easily accessible to external users and they don't have a well-developed documentation like most surveys have.

4. Is there a dualism between administrative records and survey data?

From the previous discussion it should be clear that administrative records cannot be considered as a viable alternative to survey data, at least in our case. Nevertheless, in recent years Banca d'Italia has started several research projects studying how to exploit the information coming from administrative records to overcome the *accuracy* concerns relating the SHIW.

In fact, the real challenge is how to combine them together, under the constraint that respondents' personal identifiers are not available. In our case, two more promising statistical uses of administrative records include: (1) use to supplement/update an existing frame at the design stage and (2) indirect use at the estimation stage.

As to the first point, average house prices in a given area (OMI data) could be used as an additional stratification variable. House prices are correlated with household housing wealth which, in turn, is the main component of household wealth. For this reason, one might expect some gains in terms of efficiency/reduction of nonresponse bias. Moreover, the information could be used to oversample rich households.

As to the second point, administrative data that are judged to be reliable can be used to adjust micro-data, at the estimation stage. For instance, Neri and Monteduro (2013) include in the actual SHIW weighting scheme the information on the number of dwelling owned by individuals contained in the MEF data. As another example, Neri and Zizza (2010) use imputation methods to align the number of secondary jobs in the survey to the one coming from National Accounts.

Yet, further research needs to be done. A first necessity is to have a deeper understanding of the level of accuracy of administrative records in order to understand what information that can be considered as a benchmark for the survey.

A second field of further investigation relates the methodology. Combine administrative and survey data without having the possibility of using exact matching techniques requires modeling and assumptions. Usually one issue to deal with is how to reconcile the differences in the concepts and definitions between the two sources. For instance, the SHIW collects information on wealth at the household level, while in

the administrative records it is available at individual level (this is the case for the CCR and MEF data). This might be a major drawback. Economic decisions, such as buying a house or how much to save, are assumed at the household level. Moreover, to fully understand the level of well-being of an individual, one need to consider also the economic situation of the household he/she lives in. So, any adjustment requires a first passage from household level to unit level in the survey data, which, in turn, requires modeling assumptions.

Tables

Table 1: A comparison between   micro and macro data: income account

(per cent)

Items	Ratio between micro and macro	
	2008	2010
Payroll income	89.0	87.8
Imputed rents [1]	211.9	224.7
Self-empl. income in businesses with < 6 employees and rents [2] ..	53.3	53.1
Self-empl. income in businesses with > 5 employees [3]...............	11.2	8.4
Entrepreneurial income and income from financial assets [4]	12.3	6.2
Income from pension and transfers [5]...	75.1	72.4
Disposable income	**76.2**	**79.4**

Source: Banca d'Italia 2012. After-tax incomes are derived from National Accounts by assigning each type of income a proportional share of direct taxes and social contributions.
The corresponding items from National Accounts are: (1) gross operating surplus; (2) Withdrawals from the income of quasi-corporations; (3)   Withdrawals from the income of quasi-corporations ; (4) dividends and interest; (5)   Social benefits other than social transfers in kind.

Table 2: *A comparison between micro and macro data: capital account (per cent)*

Items	Ratio between micro and macro estimates	
	2008	2010
Real estate	97.8	101.2
Land..	53.1	61.3
Business equity [1]	63.0	43.4
Total real assets	92.3	93.6
Deposits..	29.0	28.5
Government bonds	27.7	44.4
Other securities..	18.6	34.7
Total financial assets	22.9	33.1
Financial liabilities [2]	41.7	35.7
Total net wealth..	**71.0**	**76.6**

Source: Banca d'Italia (2012).
(1)   It includes non residential dwellings.   (2)   It includes all the debts towards banks and other financial institutions.

References

Bonci R., Marchese G., Neri A. (2005) "Financial wealth in the financial accounts and in the Survey of Household Income and Wealth," *Temi di discussione (Economic working papers)* 565, Banca d'Italia.

D'Alessio G., Iezzi S. (2009) "Fragilità finanziaria delle famiglie indebitate in Italia", working paper, Banca d'Italia.

Magri, S. (2009) "Household Wealth and Entrepreneurship: Is There a Link?", *Temi di discussione (Economic working papers)* 719, Banca d'Italia.

Marino, M.R., Zizza R. (2011) "Evasion of the personal income tax in Italy: an estimate by taxpayer's type", in Michael Pickhardt and Aloys Prinz (eds.), *Tax Evasion and the Shadow Economy*, Cheltenham: Edward Elgar.

Paiella M. (2004) "Does wealth affect consumption? Evidence for Italy", *Temi di discussione (Economic working papers)* 510, Banca d'Italia.

Neri A., Monteduro M.T. (2013) "The housing wealth of Italian households: a comparison of administrative and survey data," *Questioni di Economia e Finanza (Occasional Papers)* 146, Banca d'Italia.

Neri A., Ranalli M.G. (2012) "To misreport or not to report? The measurement of household financial wealth," *Temi di discussione (Economic working papers) 870,* Banca d'Italia.

Neri A, Zizza R, (2010) "Income reporting behaviour in sample surveys," *Temi di discussione (Economic working papers)* 777, Banca d'Italia.

Statistics Canada (2009), *Quality Guidelines*, Fifth Edition, available at http://www.statcan.gc.ca/pub/12-539-x/12-539-x2009001-eng.pdf.