Chapter 4

Data sources and methods for producing household wealth statistics

This chapter describes the different types of sources and methods that are used to collect and compile household wealth statistics, and discusses the ways in which they are used. The main strengths and weaknesses of the different approaches are presented, with the aim of assisting countries to determine the best approach in particular circumstances.

4.1. Different types of sources and methods

Micro-level data on household wealth are typically obtained from one or more of these sources:

● multipurpose household surveys in which a wealth module is embedded;

● integrated household surveys on wealth and income, or wealth, income and consumption;

● surveys of financial institutions;

● administrative records; and

● specialised databases covering items such as works of art, historical dwellings and luxury vehicles.

The optimal source for compiling statistics on household wealth may vary, depending on the country circumstances and on the budgets available to statistical offices and other data collecting agencies, co-operation agreements between different institutions, privacy laws, content of public records, cultural factors influencing the propensity to participate in sample surveys, and characteristics of the market for private databases.

In general terms, no single source is completely informative or error-free; methodologically sound combinations of different sources may sometimes offer the best results. There may be important conceptual differences in measures of wealth collected in various sources, and careful thought should be given to the congruence of a given measure with a given analytical objective.

4.2. Strengths and weaknesses of different approaches

All approaches involve data quality challenges, but the nature of the challenge varies significantly with the source or method.

Household surveys, discussed in detail in Chapter 6, are currently becoming the prevalent source of micro-level data on household wealth. In the past, some wealth-related items were generally included in multipurpose surveys on economic activity and living conditions of households, mainly focusing on income, employment and/or consumption. After the recent financial crisis, in an effort to gain more insight about financial vulnerability, several countries started running new surveys focusing on wealth, or renovated pre-existing surveys to include detailed information on assets and liabilities.

Household surveys often collect a core set of demographic and socio-economic information along with the topic or topics of particular interest. This core information can be used to classify households into groups and then show the distribution of total wealth or other aspects of wealth across these groups. Such classificatory information is generally not available from other sources. In case of multipurpose surveys, the simultaneous collection of data on various facets of economic activity allows, in principle, for the production of relatively complex statistics (e.g. debt/income ratios, sources of wealth accumulation by income class, etc.), without the added variability or error in results that

comes from most data-merging techniques. When surveys include a panel component, wealth dynamics can also be studied at the micro level using appropriate models.

The main challenge to data quality in cross-sectional surveys comes from the response process. Data can be biased by non-response and misreporting, particularly for sensitive or conceptually complicated topics such as wealth. Minimising these reporting problems and their effects is usually a key part of a survey strategy.

For any type of statistical survey, there is a trade-off between the perceived response burden and the amount of data collected. This can affect the extent to which a collection can obtain data on the many variables of interest for household wealth statistics.

Trade-offs also exist in terms of the precision of different sets of estimates. For example, a survey geared towards measuring the aggregate holdings of complex financial assets needs to be focused on the (typically small) group of households investing in such instruments. In most countries, these households are also the wealthiest ones, and the amount of wealth tends to be inversely proportional to a household's propensity to participate in surveys. Under a fixed budget, resources may be diverted from interviewing more ordinary households to interviewing wealthy ones. Precision in measuring rarely held assets may therefore translate into a higher variability of estimates for other ones that are widely held but not particularly common in the wealthiest segment of the population.

Administrative records provide detailed data on any asset or liability that must be registered, either for fiscal purposes or for other reasons. In most countries, dwellings and vehicles are subject to registration procedures; sometimes, other wealth items – such as savings held in foreign currency or in specific financial instruments – are too. Administrative records might also exist for debts exceeding a certain threshold, especially in countries where credit risk is evaluated by government agencies.

Administrative records provide two significant advantages over competing sources. They normally cover the whole population of an administrative unit (e.g. country, state/province/region, etc.); and, especially in the case of tax data, a great deal of effort and money is spent in ensuring their accuracy.

The use of administrative data is, however, often severely restricted in order to protect the privacy of households; depending on the country, laws might mandate that this information is used only for a specific purpose, such as calculating taxes, or that it is released only in an aggregate form to any institution, public or private, outside the producing agency. Even in the minority of countries where some administrative records are publicly searchable, auxiliary information is often scarce, limiting the possibilities of analysis if using such sources alone.

Quality issues might affect administrative records too. Compared to the case of surveys, where data quality generally varies across the sample, administrative data are more likely to be affected by systematic error; for example, the value of dwellings might be registered based on a census carried out in the distant past, and then not updated to take market dynamics into account. Conceptual differences may also be important. For example, in some instances, the value of real estate assessed for tax purposes follows a formula that has only a rough connection with one of the variety of market prices that might apply. Where administrative data are maintained for fiscal purposes, there may also be an incentive for households to act in ways that cause the value recorded to be minimised.

Private data sets cover a variety of wealth items. Some are created as standard outcomes of business relationships, for private administrative purposes: for example, banks keep

records of financial assets owned by their clients. Some are assembled to support business operations: for example, a category association of real estate brokers might sponsor a database of recent contracts so as to provide a pricing reference to associates.

The main advantages provided by these data sets are timeliness and accuracy. They tend to be updated more often compared to public-sector administrative records and survey databases. Money and time are generally spent to prevent errors in data that may end up damaging business operations.

Other private data sets exist that are more specific: auction houses document their own transactions involving valuables, etc. Also, specialised companies may construct data sets with the sole intention of selling them for marketing, research or similar purposes. Accuracy might be lower in these cases, especially when record-keeping is not mandated by the law, because data producers are separate from data users and therefore may not be as strongly motivated to deliver high-quality output. Most often, information needed to evaluate the quality or the construction of such data sets is not available.

The most accurate private databases are often not available in their full form. Customer information cannot be released by most companies on account of legal provisions, except in aggregate or highly anonymised form. Analogously, most for-research or for-marketing data sets do not contain identifying information; linking them with other sources might be particularly difficult.

Finally, private data sets generally cover segments of the population, e.g. customers of a specific credit institution, owners of a certain brand of car, residents of a certain area. While several data sets, e.g. from different banks, might be combined to get a fuller picture, this results in additional costs that are often incompatible with the budget constraints faced by institutional data producers.

4.3. Combination of different sources: Purposes and methods

Combinations of different sources can occur for several purposes. Data sets originally conceived for separate reasons might be brought together so as to have archives that are more comprehensive, covering a wider span of topics and allowing users to study them simultaneously. Archives might also be merged in order to assess or improve data quality; this happens, for example, when a data set generally known to be reliable is combined with another one that is either new and of unknown quality or known to be affected by various issues. Data from the latter can then be evaluated based on discrepancies with analogous data from the former, and/or substituted where necessary. Finally, data from one source can be used as auxiliary information when designing or using another source, without any immediate comparison taking place. Depending on the goal, different merging techniques are available.

Conceptually, two families of combinations between different sources can be defined: direct, i.e. carried out on individual data records (e.g. linking of survey-based information with tax information at the household level); indirect, i.e. based on the incorporation of data from one source as background information for another (e.g. use of housing prices from a private database in the process of survey design, use of correlations observed on a database in order to construct imputation models for a different one).

Where *direct* combination at the micro level is concerned, two kinds of techniques are available: *exact* matching and *statistical* matching. Exact matching consists in linking data referring to the same household from different archives. It requires individual households

to be identifiable. Statistical matching consists in linking data on similar households from different archives. It requires definition of criteria for assessing similarities between households, based on variables present in all the archives involved.

Exact matching tends to be difficult to carry out, because privacy laws often prevent different data producers from exchanging identifying data about individual households. Sometimes, the producers themselves are prevented from acquiring such data, or from storing them beyond immediate needs. This is the case with many sample surveys, especially when fieldwork activities are outsourced; for example, interviewers may receive a list of addresses to visit from the sponsoring agency, and they will have to acquire the names of households residing at each address for operational reasons, but the law may forbid them from relating this information to the sponsor.

The most favourable conditions for exact matching arise whenever the archives to be matched belong to the same institution, e.g. a National Statistical Office. If this institution has the right to store and use identifying information, then it can merge different archives, e.g. a wealth survey and a register of dwellings. Experiments with exact matching have also been carried out in the context of ad hoc projects where two or more data producers co-operated to assemble a specific data set, with the prior explicit consent of the households involved.

Statistical matching does not hinge on the direct use of personal data; as a consequence, it is generally not subjected to strong legal restrictions, provided that the information contained in the matched data set does not facilitate identification of the respondents. Compared to exact matching, its main disadvantage lies in the uncertain nature of the process.

Statistical matching reflects only micro-level relationships used in the matching algorithm, which can contain errors and/or omissions due to incorrect models or to scarcity of information in at least one of the involved databases. If survey data are used in a matching exercise, survey error enters the process, increasing this margin.

Sophisticated models for statistical matching have been developed in order to overcome these limitations, at least in part, and to estimate the level of uncertainty of the estimates based on matched data sets. Nonetheless, for purposes of complex modelling, the dimensions of the statistical matching define the limits of the possible relationships that can be captured between the survey data and the matched data.

Where wealth data are concerned, most matching experiments aimed at expanding the breadth of available information have been carried out by linking wealth data sets with other ones focused on consumption, income, employment and financial literacy. Socio-demographic characteristics or transformations thereof are generally used as pivot variables. Some experiments aimed at improving data quality have been conducted by matching survey data sets affected by non-response on some wealth items and/or for some categories (e.g. the very rich) with private data sets or administrative records containing information requested of, but not provided by, respondents.
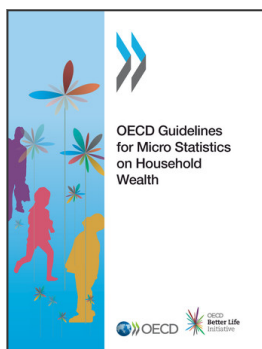
The indirect combination of sources can also take many forms. In most wealth-related cases, a household survey is the central element of the measurement process, but information from other sources may be used in the process of designing the survey, producing estimates, and evaluating the results. For example, sample design for a wealth survey may benefit from estimates of average real estate prices or average credit scores on a territorial basis; this information, extracted from administrative records or private databases, can be incorporated in the structuring of strata and/or in post-stratification (see Chapter 6). Data on assets generally owned by wealthy households only, such as dwellings

in certain premium locations, works of art or leisure boats, can help when there is a need to oversample such households. Macro-level estimates of real and financial wealth, where considered reliable, can be used to benchmark survey results, helping data producers to understand any weaknesses in their data collection process.

4.4. Summary

The key highlights from this chapter can be summarised as follows:

- Micro-level data on household wealth can be obtained from a number of sources, including multipurpose household surveys, household surveys focusing on one or more dimensions of economic well-being, administrative records and records from businesses such as financial institutions.

- The choice of source or sources for micro-level data on wealth depends on a number of factors, such as the budgets available for collecting data and privacy laws.

- Household surveys are the most common source of comprehensive wealth data. They can collect a core set of demographic and socio-economic information that is useful for classifying households into groups of particular interest. However, there is a trade-off between the perceived response burden and the amount of data collected.

- Administrative records do not place an additional response burden on households and often cover the whole population. While they are not subject to sampling error, they may be subject to other errors, and are unlikely to contain all the data required for a detailed analysis of household wealth. Access to administrative records may be restricted by privacy laws.

- Business records can be more up-to-date than administrative records, but they are less likely to cover the whole population and will include only some items of interest. Access to these records is also likely to be restricted by privacy laws.

- Data from multiple sources can be combined using data-matching techniques. Exact matching consists in linking data referring to the same household from different archives. It requires individual households to be identifiable. Statistical matching consists in linking data on similar households from different archives. It requires the definition of criteria for assessing similarities between households, based on variables present in all the archives involved.