8 Linking or matching data across data sources

As multiple micro data sources may be used in the compilation process, linking data across these datasets in a proper way to arrive at coherent and consistent sets of accounts for underlying households is of crucial importance. This chapter describes four methods to achieve this objective, with their main pros and cons.

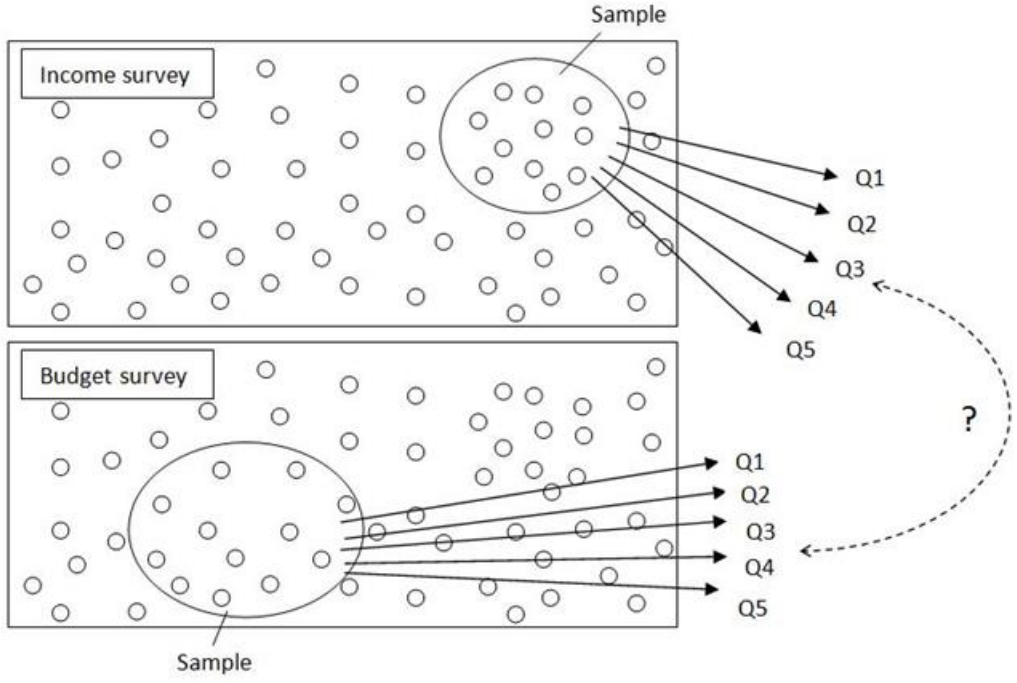
8.1. Introduction

A very important step in the compilation process is the linking or matching of data across various data sets to construct coherent data on income, consumption and saving for the various household groups. In many cases, data from different micro data sources are used, obtained via micro data surveys or administrative data, and the way in which they are combined may seriously impact the overall results.

Sometimes the various data sources may describe exactly the same households, in which case it will be easy to link the data, but in many cases, it will concern different samples of households. The question then arises how these data should be matched to create complete sets of accounts for similar types of households, and to arrive at coherent distributional results for income, consumption and saving for various household groups.

Figure 8.1 provides a simplified example of the issue, showing a country that uses two different sample surveys for its income and its consumption items. As it concerns sample surveys, different households may be selected in the samples. Furthermore, the samples may differ in size.

Figure 8.1. The issue of linking data across surveys



Source: The Author.

In order to arrive at reliable and consistent distributional results for income, consumption and saving across household groups (e.g. income quintiles as shown in Figure 8.1), income and consumption data from different micro data sets need to be matched in a coherent way. Generally, there are four methods to achieve this objective, the first two aiming to link or match data at the micro level[1] and the latter two processing results separately for each data source and only matching results at the aggregated level:

1. Link records on the basis of common household IDs or identifiers present in the data sources (record linking). Results can then be clustered on the basis of these matched micro data.
2. Merge data from different data sets into a single micro data set via statistical matching and modelling. This approach uses matching variables available in all data sets to impute missing

variable(s) available in one specific data set (the donor set) into the data set where this/these variable(s) is/are missing (the recipient set). Results can then be clustered on the basis of information from this "new" synthetic data set.

3. Construct household groupings for each data source separately in case the variable needed for clustering is available in all of them and link these results at the aggregated level.
4. Construct household groupings for each data source separately on the basis of an imputed variable in case this variable is not available in all data sources and link these results at the aggregated level. This is a variant of option c but relying on imputations instead of direct observations to match the data at the aggregated level.

These methods are explained in more detail in the following sections, explaining the basic technique as well as the main pros and cons of each of the methods.

8.2. Linking records on the basis of identifiers

In the first approach, records from different data sources are linked on the basis of unique identifiers that enable the direct linking of records across different data sources. This can for example be done on the basis of social security numbers, fiscal numbers or addresses. This option will often be available in case countries use administrative data as one of their main data sources. Data from these administrative data sources may then be linked to data available from surveys.[2]

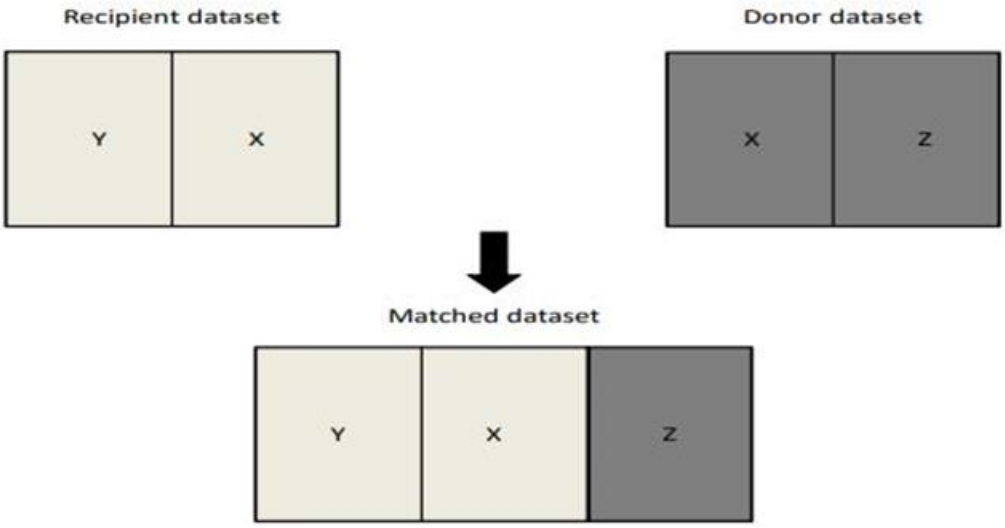
Linking on the basis of identifiers is the preferred method to link data across multiple data sources, as it ensures that the data on income, consumption and saving are fully consistent at the micro level, without the need to rely on any assumptions to link the data. This means that there is no margin of error feeding into the results as a result of the matching exercise (except in case of any errors in the identifiers themselves).

8.3. Integrating data sets through statistical matching

In the second approach, information from different data sources is fused on the basis of statistical matching. In this technique, a specific variable that is missing from one data set (the recipient data set) is imputed from another data set (the donor data set) by looking at common variables available in both. These may concern information on income (group), age, gender, marital status, region, household size, main source of income, occupation, type of labour contract, country of birth, education level, etc.

For example, a compiler may have data from two surveys, i.e. an income survey and a budget survey, in which disposable income (variable Y) is missing from the budget survey and consumption expenditure (variable Z) from the income survey. In order to obtain a data set that includes data on both income and consumption, both data sets can be fused with the help of common variables X available in both data sets. For that purpose, the relation between these common variables with the target variable need to be assessed on the basis of data from the donor data set, assessing the specific matching variables that will be used to conduct the matching. This relation can then be used to impute the target variable in the recipient data set. This will lead to a synthetic (or "matched" or "fused") file, [3] containing records that include both X, Y and Z (see Figure 8.2).

Figure 8.2. Integrating data sets through statistical matching



Source: Balestra and Oehler (2023[1]).

After the data fusion, results are processed according to the step-by-step approach and allocated to the relevant household groups on the basis of the underlying information in the synthetic data file.

A prerequisite for statistical matching is that the populations match across the various data sets and that the matching variables are identical in terms of concepts and reporting.[4] To the extent that this is not the case, adjustments will be needed to ensure a good alignment between the data sources.

The quality of the matching will largely depend on the selection of the matching variables. Balestra and Oehler (2023[1]) explain that these should meet two essential criteria. First of all, they should show homogeneous distributions across the relevant data sources, ensuring that the data sets cover similar types of households with coherent information on the distribution of the matching variables. Secondly, they should have a significant correlation to the target variable(s), i.e. they should behave as good predictor of the target variable(s) to be imputed in the recipient data set.

Ideally, the target variables (i.e. Y and Z) are independent of each other and the full relationship between the two is explained by the common variables (i.e. X). This is known as the conditional independence assumption. However, this assumption rarely holds and is difficult to test in practice (see Eurostat (2013[2])). Balestra and Oehler (2023[1]) explain that auxiliary information may help in increasing the likelihood of meeting this assumption. This may for example be in the form of having a proxy variable for Z in the recipient data set or a proxy for Y in the donor set (e.g. a reported income variable in the budget survey). Normally, the more detail that can be used in the matching, the more accurate the results. A more detailed description of statistical matching is available in Eurostat (2013[2]).

The main advantage of this approach is that households are fused at the micro level on the basis of common characteristics. This is expected to lead to relatively good matches (to the extent they meet the criteria as explained above) and provides the opportunity to assess the plausibility of the results at the micro level at the start of the process. If some records show implausible results for the combinations of income and consumption, edits may be performed before further processing the data. This could be done by correcting either income or consumption results, or by changing some of the characteristics that are at the basis of the matching. Editing the underlying micro data may be particularly relevant in case of large micro-macro gaps for specific items. Instead of applying a proportional allocation to close the gaps for the various items, one could edit those items at the micro level for which the gaps between the micro data and

the relevant national accounts data are most significant and for which the matching may show implausible results for some specific households.

The downside of this approach is that it requires assumptions for the matching which may lead to some degree of uncertainty surrounding the matched results. This will largely depend on the coherence of distributions of the common variables across the various data sets and of the explanatory power of the common variables to explain the target variable. As the statistical matching may not perfectly capture the full relationship between the common variables and the target variable, it is sub-optimal to direct matching on the basis of identifiers (Balestra and Oehler (2023[1])), but preferable over matching at the aggregated level only (as described in the next two sections).

8.4. Construct household groupings for each data source separately on the basis of a common variable

Data from different sources can also be processed and clustered into household groups independently, with matching only taking place at the aggregated level. In that case, the various steps will be processed separately for the various data sets and distributional results will only be linked in a final step. Figure 8.3 presents an example of how this works, showing a country that uses different sample surveys for its income and its consumption items, with different households included in both. The data are processed separately for the two surveys and then combined at an aggregated level on the basis of the targeted household groups.

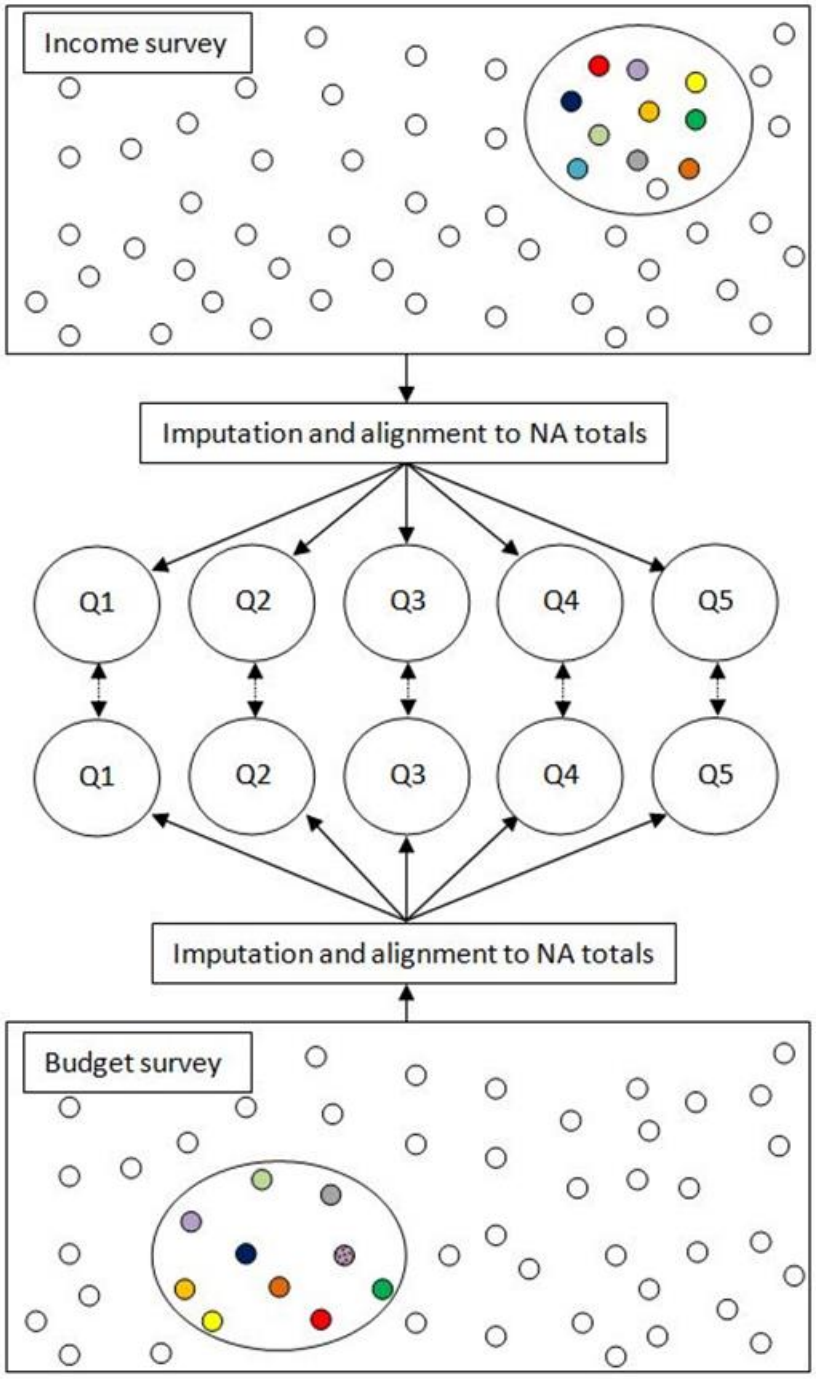
This approach can be applied when the variable necessary for clustering into household groups is available in all relevant data sets and is comparable in terms of concepts and reporting. If that is not the case, this would require specific adjustments and edits to ensure good alignment between the variables. Alternatively, one could opt to impute the relevant variable(s) in the missing data sets (see Section 8.5).

When applying this approach, it is important that the data sources describe the same population and that they show similar distributions for the common items, i.e. the same prerequisites for applying statistical matching. This will increase the likelihood of starting from similar data sets and it will help avoiding incoherent matching results. The latter may occur, for example, with regard to the level of income reported for each of the income groups, the number of households included in each of the household groups, and with regard to the socio-demographic characteristics of the various household groups.

With regard to the first issue, the income levels when clustering households according to their equivalised disposable income may be different when clustering results for each data set separately. It will be important to apply similar kinds of adjustments to income as reported in the various data sources, to ensure similar income concepts across data sources, bearing in mind that the specific adjustments needed may differ across data sources dependent on their concept and coverage of income items.[5] Ideally, compilers would then arrive at similar income levels, upper and lower bounds, and distributions for the various household groups across the different data sources. If this is not the case, compilers should investigate the main underlying reasons for any differences and try to make informed adjustments to bring the results closer in line.

When looking at clustering according to other characteristics than income, it is important to ensure consistency in the number of households for each household group across the data sources used.[6] For example, when clustering according to main source of income, a different number of households may end up in the group “income from self-employment” according to income survey results than according to budget survey data. In case of large differences in numbers of households for specific household groups, compilers should investigate the main reasons for these differences. These may for example relate to differences in concepts, differences in weights and/or differences in reported values. Dependent on the most likely issues, compilers need to make adjustments to ensure closer alignment between the results.

Figure 8.3. The issue of linking data at an aggregated level



Source: The Author.

Finally, differences may show up in the sociodemographic composition of various household groups. For example, the first income decile clustered on the basis of budget survey data may show a much larger number of single households and people below 25 that results according to the income survey. This may point to differences in (sample) populations and/or income definitions. The same may apply for other household groupings. If this occurs, compilers should investigate the main underlying reasons and make necessary adjustments to better align the results.

The benefit of this approach is that it is less complicated and probably less time-consuming than the first two approaches. The downside is that it may lead to less reliable results as the plausibility of the results can only be checked at an aggregated level and is normally only done on the basis of the variable relevant for the clustering. This is different when applying statistical matching, which relies on matching at the micro level and is done on the basis of multiple common variables, leading to closer matches and to one synthetic data set underlying the distributional results, ensuring consistency in income levels, number of households and socio-demographic characteristics for the various household groups across income and consumption.

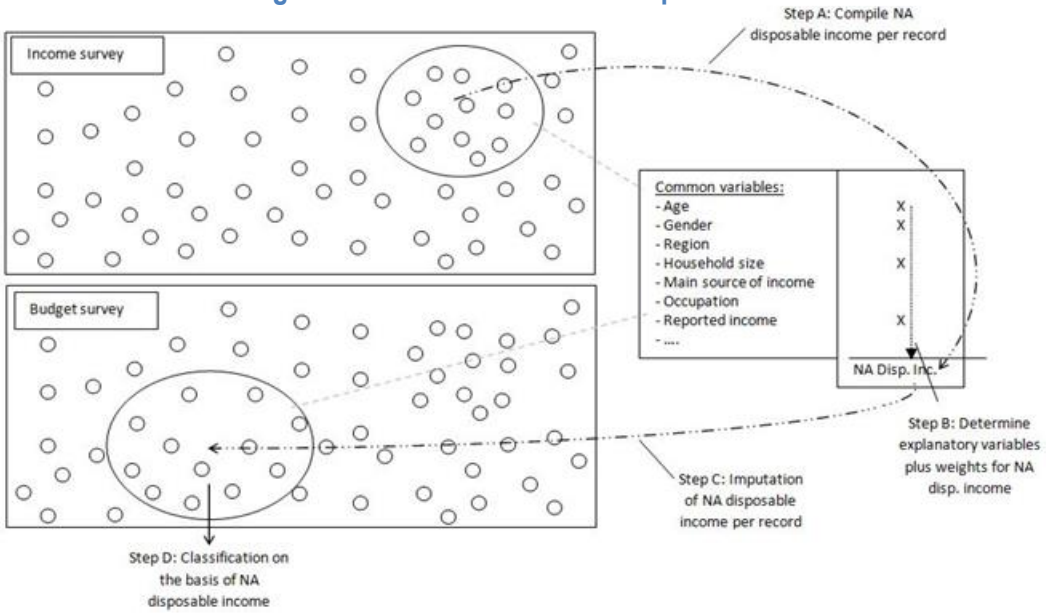
Furthermore, statistical matching enables analysing the plausibility of the results across income and consumption at the micro level. Plausibility checks are much more complex when linking at the aggregated level, as inconsistencies may be due to a larger number of factors. This also means that it is more difficult and may take more time to make accurate adjustments in case of any observed inconsistencies. In that regard, statistical matching is preferable over linking at the aggregated level.

8.5. Construct household groupings for each data source separately on the basis of an imputed variable

The last option to arrive at a coherent distribution of households across household groups is a variant of option 3, when the common variable to cluster households into household groups may not be available for all data sets. In that case, one may consider imputing the common variable in the relevant data sets and then cluster households accordingly for the various data sets separately.

For example, if disposable income is not available in all data sets, a disposable income variable could be imputed on the basis of common characteristics, via which households can be classified consistently into deciles. This method has some similarities to statistical matching in the sense that households are matched on the basis of similar characteristics, but it differs as in this option records are not matched at the micro level but at the aggregated level on the basis of an imputed disposable income item. As records are not fused individually, the various steps in the methodology can be processed independently, and at the final stage households can be classified on the basis of this imputed disposable income. Figure 8.4 presents a simplified example of how this technique works.

Figure 8.4. The issue of linking records on the basis of an imputed income variable



Source: The Author.

Looking at the specific steps, first, in the income survey, an income item has to be created according to national accounts definitions and in line with the national accounts totals (step A). This requires linking and aligning the relevant items from the micro survey to national accounts and imputing for any missing items. As a result, one arrives at an "*NA aligned disposable income*" per record. Subsequently, a regression analysis can be run on the basis of common variables in the various data sets to find explanatory variables (e.g. relating to households' characteristics, reported income and/or consumption, etc.) to explain these disposable income levels (step B). As these variables will be used to impute an "NA aligned disposable income" in all data sets, it is important to look at common characteristics available in all data sets. This may include "age", "gender", "marital status", "region", "household size", "main source of income", "occupation", "income", "type of labour contract", "country of birth", "level of education" etc. The regression analysis will lead to a model that can be used to assign NA aligned disposable income levels to micro records in the other micro data sets (step C).

In the final step, households in the other data sets can be classified into income groups on the basis of these imputed income levels (step D). The latter may be done using income boundaries defined on the basis of the imputed income results in the respective data sets or boundaries determined on the basis of the income part of the work. In the former case, one can make sure that the ten deciles consist of 10% of the households according to the results of the specific data set. However, income levels may deviate from the ones used for the classification of households in the income part, also implying that households with similar characteristics would not necessarily end up in the same deciles across all data sets. In the second option, the boundaries will match those used for allocating households in the income part (probably leading to a better match between income and consumption results), but as this may lead to different numbers of households per decile for the consumption part, this may require adjustment of weights for the underlying micro data. In adjusting the weights, one has to make sure that all deciles consist of 10% of the households and that the sum of the deciles still adds up to the national accounts totals.

As was the case with statistical matching, the approach depends on the coherence of information as reported for the common variables in the relevant data sets (i.e. homogeneous distributions) and that the common variables provide a good predictor of the target variable. Furthermore, it is important that the common variables are identical in terms of concepts and reporting, possibly requiring specific adjustments and/or edits if this is not the case.

This approach generally has the same benefits and downsides as the previous approach. The main additional benefit in view of clustering according to income groups, is that it ensures that the income concept for clustering households is consistent across different data sets and that – as it relies on multiple common variables to derive this variable - similar types of households will be assigned similar types of income. However, it may not be straightforward to run the regression analysis. Furthermore, given the downsides of only linking at the aggregated level (as explained in Section 8.4), the approach is still sub-optimal in comparison to linking data at the micro level as is done in the first two approaches.

References

Balestra, C. and F. Oehler (2023), "Measuring the joint distribution of household income, consumption and wealth at the micro level", *OECD Papers on Well-being and Inequalities*, No. 11, OECD Publishing, Paris, https://doi.org/10.1787/f9d85db6-en. [1]

Eurostat (2013), "Statistical matching: a model based approach for data integration", https://ec.europa.eu/eurostat/documents/3888793/5855821/KS-RA-13-020-EN.PDF.pdf/477dd541-92ee-4259-95d4-1c42fcf2ef34?t=1414780333000 (accessed on 30 August 2023). [2]

Notes

[1] These methods are discussed in detail in Balestra and Oehler (2023[1]).

[2] Please note that this may not always be possible (even if micro data may include unique identifiers) due to legal constraints (e.g. in view of general data protection regulations).

[3] As explained in Balestra and Oehler (2023[1]), the term "synthetic" is used as not all data for each household as included in the resulting data set have been directly observed but may have been obtained by combining information from different data sources.

[4] For example, if common variables in a specific survey are deemed more liable to reporting errors than in others, this may affect the quality of the matching. This may be particularly relevant in combing survey data with administrative data. In order to avoid incorrect matching results, it is important to first edit the micro data in the various data sets before applying the statistical matching.

[5] Compilers should avoid taking reported income from other surveys as a direct proxy for disposable income as defined in the System of National Accounts. Imputation for missing items and alignment to national accounts totals is often affecting different types of households in different ways, affecting income levels in different ways and altering the ranking. For that reason, using the reported income as a direct proxy will likely lead to incorrect matching.

[6] This will not be an issue for grouping by income as the relevant groupings are defined by number of households, e.g. 10% of households in each decile when breaking down by equivalised disposable income deciles.