6 Imputation for missing items

In the third step, any gaps between the micro data and the national accounts totals need to be bridged. To this end, first, imputations need to be made for elements not covered in micro data sources, such as for missing parts of the population, informal activities and items that are specific to the national accounts. This chapter presents specific methods to apply these imputations.

6.1. Introduction

Whereas step 2 as described in Chapter 5 foresees in the selection of micro data for each national accounts item, it has to be understood that some information may not be covered in micro data sources, requiring the need for imputations to include the relevant amounts in the distributional results. Because missing information in the micro data may explain a large part of the gap between the micro aggregates and the national accounts totals, imputing for this missing information constitutes the first part of the third step of the step-by-step approach. After making the relevant imputations, compilers can assess the remaining gap and allocate the amounts to the relevant households on the basis of the most likely underlying reasons for these remaining gaps. This second part of the third step is discussed in Chapter 7.

There are four types of missing information. First of all, some items will have no counterpart in the micro data. This is for example the case for items that are specific to the System of National Accounts (SNA). Secondly, it may be the case that certain items are covered in micro data sources, but that these are not (yet) available to the compiler. This may be related to the timeliness and the frequency of the data sources. Thirdly, it may be the case that the selected micro data source may not cover the whole population, for example in case of surveys that only target specific household groups or in case of reporting thresholds in administrative data. Finally, it may concern missing information related to the underground economy and/or illegal and informal activities. These are included in the national accounts but usually not covered in micro data sources. As this missing information may concern substantial amounts that may relate to specific households or household groups, their allocation may significantly affect the distributional results. For that reason, it is important to separately impute for these types of missing information.

This chapter discusses imputations in relation to these four underlying causes, presenting basic techniques for allocating the related amounts to the relevant households or household groups. Section 6.2 discusses the imputations for items for which no counterpart information is available in micro statistics. Section 6.3 discusses the case in which items are covered in micro data sources but are not (yet) available to compilers. Subsequently, Section 6.4 discusses imputations for missing parts of the population. Section 6.5 discusses the imputation for the underground economy, and illegal and informal activities. More detailed guidance on how to impute for missing information at the level of specific income and consumption items is provided in Chapters 10 and 11.

6.2. Imputation in case an item is lacking from micro data sources

The first type of imputations concerns those for items that have no counterpart in the micro data. This often relates to items that are specific to the SNA, such as *employers' imputed social contributions* (SNA codes D122 and D612), *investment income attributable to insurance policyholders* (D441), *investment income payable on pension entitlements* (D442), *financial intermediation services indirectly measured* (FISIM), and *social transfers in kind* (D63). As these items are specific to the SNA, no direct information will be available in micro data sources and the amounts will have to be allocated in a different way.

In general, three methods are available to derive an appropriate allocation in case no micro data are available, all making use of indirect information. The first method (defined as method B[1]) proxies the missing information by using the distribution of another component, assuming that the two are distributed in a similar way. The distribution for *employers' imputed social contributions* (D122) may for example be derived on the basis of the distribution of *wages and salaries* (D11), whereas the distribution of *FISIM* may be linked to *interest paid* (D41P) and *interest received* (D41R).[2]

The second method (method C) imputes missing distributional information according to exogenous data, e.g. socio-demographic information used for the distribution of *social transfers in kind*, available at the individual or at the household level. In both cases, it is preferred to employ the imputations at a level as detailed as possible as it enables classifying households into different groupings in the remainder of the

process. When imputations are made at the group level, this will need to be done for the various classifications that are needed.

If no information is available, a third method can be used (method D) in which the distribution of one of the balancing items is used as a proxy. In that way, the inclusion or exclusion of the component does not change the distribution of that balancing item. However, this should only be done as last resort. Naturally, this can only be done at the end of the process when the distributional information has been derived on the basis of the other variables. In applying this solution, it has to be decided to which balancing item to best link the specific item. For consumption items it will be best to link it to either *final domestic consumption expenditure* (P31DC) or *final national consumption expenditure* (P31NC) (excluding the item or items for which an imputation is still needed). For income components, the distributions may be linked to the *balance of primary incomes* (B5), *disposable income* (B6) or *adjusted disposable income* (B7). It will depend on the underlying item what aggregate will provide the best proxy. It may also be the case that one would like to use the distribution of a balancing item but excluding a specific item. Compilers should assess which item or combination of items they think will provide the best proxy for the relevant item.

Chapters 10 and 11 discuss the various income and consumption items in more detail including possible imputation techniques for the items that are most likely to be missing in micro data.

6.3. Imputation in case the micro data source is not (yet) available for a specific period

The second type of imputations concerns those for items that are usually covered in micro data sources but that may not (yet) be available to compilers for a specific reference period. This may be due to the fact that the data source only becomes available with a certain time lag or is only conducted every couple of years, as a consequence of which it is not available for the specific reference period.

In the case that data are not yet available, it may be relevant to assess whether results can be obtained by extrapolating results on the basis of historic data. The most simple approach is to just apply the distribution available for the most recent year (thus assuming no change in the relative distribution across households). A more sophisticated approach would be to look whether one can spot specific trends in the historic data that may assist in deriving more accurate estimates for the reference year. Alternatively, one may assess whether the results correlate to other data for which more timely information may already be available. This may be in relation to national accounts totals but also in relation to meso-information such as labour market data or sociodemographic information. Furthermore, in case of specific policy changes, one may try to assess how these may affect specific households or household groups. In this way, one may arrive at more accurate estimates for the reference year. These may then be revised once the actual micro data become available.

If a specific micro data source only becomes available every couple of years, the above techniques may be used to derive first estimates for the missing years. These can then be revised at a later stage when results become available for a more recent year. In that case, interpolation techniques could be applied to arrive at more accurate estimates for the intermediate years, therewith overwriting the earlier results.

For both the extrapolation and interpolation techniques, it is recommended to apply them at the micro level as this will lead to the most accurate results. In this regard, it will provide the opportunity to update the clustering of households according to the interpolated or extrapolated micro data, taking into account dynamics between household groups, which may not be captured if these techniques are only applied at the level of household groups.

For both techniques, it will be important to assess their reliability on the basis of the size and direction of the revisions for the various household groups. If needed, compilers may need to further improve the techniques to arrive at more reliable results. It is also important to look at the revisions for the various

household groups to assess at what level of detail to publish the estimates. If the revisions turn out to be particularly large at a specific level of detail, it may be decided to only publish the extrapolated results at a more aggregated level of detail.

6.4. Imputation for missing parts of the population in the micro data

A third type of imputations relates to specific groups of households that may be missing from micro data sources. With regard to surveys, this may relate to people living in overseas territories or in sparsely populated areas but also to other groups that may be difficult to capture, such as very rich households or people with no usual place of residence. With regard to administrative data sources, it may be the case that these only target specific parts of the population or use thresholds, which may exclude specific groups of households from the population.

In case specific groups of households are missing, it is important to assess whether their information can be obtained in other ways. A first solution is to impute on the basis of micro data available from other micro data sources. In that regard, survey data may be complemented with administrative data and vice versa. In that case, it is important to first check whether both micro data sets are based on the same underlying concepts. If this is not the case, the micro data from the "donor" data set will first need to undergo some adjustments in order to align to the concepts of the "recipient" micro data set. These adjustments may for example be done on the basis of patterns found for households that are covered in both data sets and that are deemed comparable with households for which imputations are needed.

An alternative solution is to look whether auxiliary information may be available on the households that are missing on the basis of which their results can be approximated. For example, if no information is available on property income for a specific group of households, information may still be available on their ownership of specific types of financial and non-financial assets. In that case, this may be used to derive estimates for the missing population on the basis of assumptions of a specific rate of return. It may also be the case that another item may provide a valid proxy to derive the results for the missing households. This is similar to the technique as explained in Section 6.2 under method B, but now only being applied to a part of the population. In that regard, it is also possible to impute for the missing part of the population by linking it to exogenous information, in line with method C as explained in Section 6.2.

A third solution is to look for comparable households in the micro data set on the basis of which the missing households may be imputed. In some cases, this may concern a simple adjustment of the sample weights, but in case the missing households have very different characteristics, it may be needed to link them to specific individual households in the sample or in the register. This may be done on the basis of one-to-one linking, searching for a specific household record with similar characteristics, but it may also involve looking at a group of households with similar characteristics and taking the average amount of this group. Finally, it may involve regression analysis in which the value for a specific household is explained on the basis of a set of underlying characteristics derived on the basis of analysis of data of other households included in the data set. This is explained in more detail in the OECD Framework for Statistics on the Distribution of Household Income, Consumption and Wealth (OECD, 2013[1]).

Finally, parametric tail adjustments are often used to assess the impact of missing very wealthy households. This can be done by using external benchmark data to assess the size of the measurement error, such as done by Vermeulen (2014[2]) who uses the Forbes list of extremely wealthy to improve estimates of wealth survey micro data. This technique can also be applied to income. In this regard, Lakner and Milanovic (2013[3]) proxy the missing top incomes on the basis of the discrepancy between survey and NA consumption data and allocated this to the top using Pareto fitting.[3] The latter is, however, not preferred as multiple reasons may underlie the gaps between the micro and the macro aggregates, so taking this as a proxy for the missing top incomes may lead to incorrect distributional results.

Alternatively, Törmälehto (2017[4]) suggests, in the absence of external data on top incomes, to replace the whole tail of the outliers in survey data with estimated Pareto distributions, using hypothetical Pareto coefficients. Furthermore, Grilli et al. (2022[5]) provide a specific application of a Pareto-tail adjustment for income, using the available micro data to explore the existence of a Pareto-tail for specific items and providing guidance on how to make adjustments to the micro data in case the top-tail appears to be missing from the micro data source.

In analysing the possible need for top-tail adjustments, compilers are also encouraged to assess the distribution of the top tail in other countries and to compare survey-based results with register-based results. Furthermore, a comparison over time may provide useful insights into whether information at the top (and at the very bottom) may be missing for specific years.

As the group of households that may be missing from the micro data source may concern households with different characteristics, it may require different techniques to impute for the missing information. In that regard, it is recommended to try to derive more or less homogeneous groups of households for which a specific technique is deemed to provide the best results. This grouping can be done on the basis of socio-demographic information as well as on the basis of values obtained for these households in other parts of the work. For each household group, amounts should be derived on the basis of the technique that is deemed most reliable. This may for example imply that auxiliary information is used to impute values for unemployed persons that are not captured in the survey, an adjustment of the survey weights is applied to include households living in sparsely populated areas, and a Pareto-tail approximation is used to derive results for the very high-income households. Results on the basis of the different techniques may also be compared to see whether they show large differences and whether adjustments may be needed to some of the results before incorporating them in the distributional analysis.

It is recommended to select the appropriate imputation techniques for the relevant underlying household groups in close cooperation with the responsible experts from the relevant micro data source. They have the best overview of what is covered in the micro data source and what imputation techniques may lead to the best approximation for specific groups of missing households and to comparable results with the data included in the micro data source. Moreover, they may be able to process (some of) these imputations as part of their compilation process, providing the compilers of the distributional data with a consistent, comparable and comprehensive data set at the micro level.

6.5. Imputations for the underground economy, and illegal and informal activities

The fourth group of imputations concerns those for economic activities that are deliberately concealed to avoid tax payments (underground production) or are not captured because of their illegal or informal nature. As these activities are usually not captured in micro data sources (for that reason often referred to as the non-observed economy), the related amounts will have to be estimated indirectly in order to include them in the distributional results.

As national accountants often make explicit estimates for these activities, this will normally provide the starting point for allocating the relevant amounts to the underlying households or household groups. Ideally, the national accounts provide information on the imputed amount broken down into the three underlying types of activities (i.e. underground, illegal and informal activities), so that the amounts can be allocated accordingly. In that regard, it is not only important to obtain information on the specific values, but also on how these amounts have been derived. It may then be assessed whether the underlying assumptions for calculating these amounts may also provide input to allocate the relevant amounts to underlying households. For example, if part of the underground economy is imputed on the basis of the assumption that specific types of jobs are more likely to be involved in such types of activities, this may be used to link the amounts to specific groups of households. Of course, these assumptions can be further tuned to take into account specific characteristics that are available at the micro level on the basis of which

it can be decided which households should be assigned what amount and whether some specific groups of households should be excluded. For example, assumptions may be made with regard to the background (sex, age, employment status and living location) of drug dealers, prostitutes and traffickers. For some of these activities, information may also be available from police records.

If no specific information is available from the national accounts on the size of the underground economy, illegal and/or informal activities, it is important to separately estimate the related amounts and to separately allocate them to the relevant households, as the amounts are likely to involve (partly) different groups of households. A first step would be to look at the micro-macro gap and to assess what part may be explained by these three types of activities. In the second step, the amounts should be allocated to the households that are most likely to be involved in them. As mentioned above, in some cases information may be available on what type of households are more likely to be involved in what type of non-observed activities. In that case, the related amounts can directly be allocated to relevant households or household groups on the basis of their specific characteristics. In other cases, assumptions will need to be made, for example looking at the likelihood of households to be involved in these activities on the basis of their reported data (see below).

In looking at which households may possibly be involved in underground activities, illegal or informal activities, one may look at the plausibility of the overall results at the household or at the household group level to see whether specific amounts may be missing. For example, if for some household groups consumption by far exceeds their income, it may be the case that they are actually running a deficit and sell off assets or engage in liabilities,[4] but it may also be the case that part of their income is not covered in the micro data source. In that case, this may require an imputation, the specific item depending on what item is most likely to be underreported by the specific households. The latter may be based on the items that are most likely to be underreported in general and show the largest micro-macro gaps (e.g. mixed income, property income and social benefits) or which are most likely to be underreported for specific groups of households.

In that regard, it is also interesting to cross-check results for households with similar characteristics. It may of course be the case that they report different amounts for specific items, but in case these are much larger in a specific year or for a specific group of households, this may point to possible outliers or errors in the data. If on the basis of such analysis it is indeed concluded that the micro results are likely to be incorrect due to underreporting in relation to underground economy, illegal or informal activities, an imputation may need to be made, looking at a more plausible value in relation to previous years or comparable households.

Of course, some of these imputations will be very sensitive to assumptions on the plausibility of the micro data. For that reason, it is very important that this analysis and allocation are done by or in close cooperation with the responsible experts from the relevant micro data source. They are best equipped to assess the plausibility of the results for the various groups of households and best suited to assess where an imputation for non-observed activities may be most valid.

6.6. Conclusions

This chapter discussed general techniques how compilers may deal with elements for which micro data may be lacking. As explained, the imputation technique, which may differ across households or household groups, will depend on whether there is no micro data available at all, whether this may not (yet) be available for the specific recording period, whether only part of the population is covered, or whether information on specific activities may be missing. The micro and macro experts should discuss which technique is deemed to provide the most reliable estimates for which specific households or household groups and carefully check the results, also in relation to data that are available in the micro data sources.

Ideally, imputations are made at the micro level. This provides the opportunity to check the reliability of the results at the micro level and also ensures that the next steps in the process can start from underlying micro data. In that regard, it has to be borne in mind that the alignment of the micro data to the national accounts totals should also be done on the basis of the micro data, after which the households can be ranked at the micro level according to their income levels including the imputed amounts. This also ensures that results can be aggregated into multiple household groupings, all arriving at consistent results in line with national accounts totals.

References

Grilli, J., P. Engelbrecht and J. Zwijnenburg (2022), *Pareto tail estimation in the presence of missing rich in compiling distributional national accounts*. [5]

Lakner, C. and B. Milanovic (2013), "Global Income Distribution - From the fall of the Berlin Wall to the Great Recession"*, Policy Research Working Paper*, No. 6719, World Bank, https://openknowledge.worldbank.org/bitstream/handle/10986/16935/WPS6719.pdf?sequence=1&isAllowed=y (accessed on 27 October 2017). [3]

OECD (2013), *OECD Framework for Statistics on the Distribution of Household Income, Consumption and Wealth*, OECD Publishing, Paris, https://doi.org/10.1787/9789264194830-en. [1]

Törmälehto, V. (2017), "High income and affluence: Evidence from the European Union statistics on income and living conditions (EU-SILC)"*, Eurostat Statistical Working papers*, http://ec.europa.eu/eurostat/documents/3888793/7882117/KS-TC-16-027-EN-N.pdf (accessed on 9 October 2017). [4]

Vermeulen, P. (2014), "How fat is the top tail of the wealth distribution?"*, ECB Working Paper*, No. 1692, ECB, http://www.ecb.europa.eu/pub/scientific/wps/date/html/index.en.html (accessed on 27 October 2017). [2]

Notes

[1] Please note that method A is reserved for deriving distributional results on the basis of actual underlying micro information.

[2] Please note that ideally this should only include interest paid to and received from banks, but the total amounts can still provide a good proxy.

[3] Pareto-tails are based on the observation that in many populations the income distribution at the top is distributed in a similar way. As explained by Lubrano (2017[6]) it assumes that the number of individuals whose income exceeds a given level x can be approximated by Cx^{α} for some choice of C and α. This approximation seems particularly accurate for large incomes, i.e. for x above a certain threshold. Therefore, Pareto tails approximations are often used to check the plausibility of survey results for higher income households. In that regard, they can also be used to derive estimates in case very high-income households are deemed to be missing. For more information on Pareto tails, please see Vermeulen (2014[2]), Lakner and Milanovic (2013[3]), Armour et al. (2014[7]) and Chakraborty and Waltl (2018[8]).

[4] For this purpose, it would be very useful if the information could be combined with information from the capital and the financial accounts.