Use of non-official sources for transforming national data into an international statistical product – UNIDO's experience

Shyam Upadhyaya
UN Industrial Development Organization (UNIDO)
s.upadhyaya@unido.org

When a research study covers a single country only, the required data can be obtained from the country's national statistical office (NSO). However, if the study encompasses several countries, it may be more practical to derive the necessary data from international agencies, considering that the data they make available are internationally comparable in terms of statistical coverage, classification standards, valuation and computation methods of major variables. International data producing agencies face two major challenges. First, the data reported by NSOs may not be distorted when presented in international publications, and secondly, quality assurance of data in terms of accuracy, coherence and international comparability requires a reasonable degree of adjustment of official data. UNIDO applies five stages of transformation. In the first two stages, official data are fully preserved. In the subsequent stages, official data are adjusted and supplemented by UNIDO's own estimates based on external sources including non-official sources. This paper describes UNIDO's efforts to retain and utilize official national data, but to simultaneously also meet users' requirements for more comprehensive and coherent statistical data.

Key words: data transformation, comparability, non-official sources

1. Introduction

UNIDO Statistics maintains an international industrial statistical database and disseminates global statistics through the publication of the *International Yearbook of Industrial Statistics*, *World Statistics on Mining and Utilities* and online access to INDSTAT and IDSB databases[1]. The primary sources of these databases are results of industrial surveys conducted by national statistical offices (NSOs). National data are transmitted to UNIDO by returning the general industrial statistics questionnaire which contains eight indicators related to employment, wages, gross and net output and capital formation. National data undergo scrutiny and transformation in UNIDO's data production process with the purpose of converting national data into an international statistical product.

The transformation process is part of UNIDO's data quality assurance framework. National data represent the official statistics of a member state; thus, it is necessary to preserve its original nature. Statistical products of UNIDO are freely shared with NSOs and their general conformity with national databases and statistical publications is greatly

[1] The Industrial Statistics – (INDSTAT) database contains data on major indicators of industrial statistics for around 160 countries at the 2- and 4-digit level of ISIC. The database can be obtained in CD Rom or accessed online on UNIDO's Statistics webpage. The Industrial Demand and Supply Balance (IDSB) database contains production and external data by country at the 4-digit level of ISIC.

appreciated. At the same time, the quality assurance of data in terms of accuracy, coherence and international comparability requires a reasonable degree of adjustment of national data. Data transformation in the UNIDO context implies the improvement of data quality, but by no means a replacement of reported data. It refers to the entire process from detecting and correcting obvious reporting errors to nowcasting for the most recent years. Any corrective measure poses a certain degree of intervention to the original data. Data transformation at UNIDO is therefore carried out in stages, with the degree of intervention increasing from lower to higher stages. Consequently, the official status of data is fully preserved at lower stages, while data from non-official sources are used at higher stages.

This paper describes the different stages of data transformation at UNIDO Statistics and the use of non-official sources in this process. To preserve the official status of data, the use of non-official sources is limited to higher stages, mainly for the imputation of missing data.

2. Stages of data transformation

The main objective of data transformation is to convert national data into an international statistical product. National data inherently differ by currency, national adaptation of industry classification, reference periods, etc. Even when the country follows the international statistical standards for classification of economic activities, there is a regional or national adaptation in most cases that adds to the deviation from international standards. Occasionally, data are reported with a certain degree of deviation from own national standards. Quite often, NSOs carry out split or combination of industry groups to adjust a smaller number of observations that cannot be reported separately for confidentiality reasons. Sometimes, adjustments are made to maintain the historical series of data initially reported in different versions of industry classification. The volume and sequence of work necessary to transform data is determined by the number of incompatible, missing and dubious values and usually delays the report. Obvious errors and discrepancies are immediately detected in UNIDO's screening process once the data transfer has been completed. More complex problems are encountered in the process of data analysis.

Eliminating inconsistencies and imputing for missing values is mostly based on reported figures. Economic variables, such as the number of employees and wages and salaries paid, output and value added, are highly correlated. Thus, the ratio derived from the reported variables often serves as a predictor for missing values. However, there are cases in which one of the variables required to obtain an appropriate predictor for a missing value is not available in reported data. In that case, data from non-official sources come to play a crucial role. Another problem relates to varying time lags in data reporting. In some countries, industrial surveys involve extensive and time-consuming field operations which delay the publication of results that are subsequently reported to UNIDO. To bring the national data in line with the most recent single year, data need to be extrapolated.

Data transformation at UNIDO Statistics is carried out in five stages:

1. In the first stage, only obvious reporting errors are corrected. At this stage, data fully retain their original form. These data are used to pre-fill the questionnaire that is submitted to NSOs in the following round of data collection.

2. Any inconsistencies found at this stage are corrected with official data that are available in NSO publications or websites. Estimates are generated to correct obvious inconsistencies or to replace the missing values. Data contained in the survey reports conducted by NSOs under UNIDO-funded projects are also considered to be official.

   Stages 1 and 2 fully preserve the official status of data. These data are published in the *International Yearbook of Industrial Statistics* and the *World Statistics on Mining and Utilities* with a brief description of the data source – such as the name of the data supplying institution, coverage and method of the survey and other information.

Estimated figures may be presented in these publications in relative or aggregated form only. These publications provide quick reference for policymakers and other users to the latest statistics on the general trend of global industry. Researchers and development analysts who prefer to carry out their own analyses using longer time-series data can contact UNIDO for a database in electronic media.

The databases disseminated in electronic media through CD Rom or online access to UNIDO's website are further transformed. The database in electronic media has wider coverage in terms of the number of countries reported and time periods.

3. Stage 3 resembles stage 2 in terms of process, but the difference is that non-official data can be used to make any necessary adjustments to eliminate the deviation of reported data from international standards.

4. Most of the imputation for missing data is done at this stage. It involves automatic interpolation as defined in the imputation guide as well as any remaining disaggregation due to the lack of supplementary information.

5. At this stage, extrapolation is carried out whenever applicable in order to bring the data in line with the most recent single year. A time lag of two years is considered normal for structural business statistics. However, many countries have longer time lags; thus, missing data for the latest years have to be estimated using the extrapolation method. Such estimates are considered provisional and are replaced as soon as survey data become available.

Data become available to update the database once it has undergone all stages of transformation. In general, data collection, transformation and updating is a live and ongoing process. Data are scrutinized as soon as they are received. However, UNIDO

prepares a set of data products in CD or as an online version for dissemination purposes. These products are released once annually, a few months after the printed version is released.

3. Non-official sources

There is no common understanding among statisticians about the distinction between official and non-official sources. In the UNIDO context, non-official sources generally refer to data that were not officially reported or published by the national statistical organization (or any agency responsible for statistics). The substantial part of the UNIDO database is fed by data that is officially reported by NSOs on account of the mandate of the Organization in the area of industrial statistics. UNIDO also regularly receives data from other organizations such as OECD and UNSD under the data exchange programme, which are first reported to these agencies and then transferred to UNIDO. Naturally, such data are considered official data.

In some countries, industrial surveys are conducted under UNIDO-funded technical assistance projects. The project report usually cleared by the government contains the results of the survey. In case there is a delay in official reporting, the data in the project report are considered official and are entered into the database.

Hence, data from the following sources are considered non-official:

- Data compiled and disseminated by international agencies without direct reference to the national sources;
- Data obtained from commercial data providers or knowledge institutions;
- Penn World Tables, Economic Intelligence Units, etc.;
- Estimates using a combination of official and non-official sources;
- Imputed data;
- Estimates generated from time-series models – forecasts, nowcasts.

It is quite often difficult to distinguish official from non-official sources with regard to international agencies. Most international agencies collect data from national sources, however, these data are often supplemented with their own estimates, which cannot be considered official data of the given country. Therefore, not all data from international agencies can be regarded as official sources.

4. Imputation strategy

Non-official data sources are predominantly used at UNIDO for imputation of missing data in business structure statistics. In the case of macroeconomic variables, such as GDP and MVA, non-official sources can be directly used for the compilation of the final tables intended for publication. However, such data are mostly presented either in relative or aggregated form. Structure business data imputation can be carried out for a missing data

item, missing period or missing section (entire country). UNIDO uses both the single imputation and multiple imputation method. However, in practice, single imputation based on the economic relation of variables is widely implemented.

A study was carried out at UNIDO to determine the extent of missingness in databases. For this purpose, data were scanned using the R-based VIM package – VIM stands for "Visualization and Imputation of Missing Values" (Templ, Alfons, & Kowarik, 2010). The visualization tools became particularly useful to explore the data and structure of the missing values, which helped in selecting an appropriate imputation method. VIM is applied only to identify missingness, i.e. it is used before imputation is performed.

The results of the study on time series evolution of missingness in UNIDO business structure data have shown that missingness increases in later years. As mentioned earlier, a normal time lag between the survey and data reporting period is two years. However, this time lag is longer for more than 2/3 of the countries that are reporting data. For many developing countries, the time lag is longer than 3-4 years. Imputation may significantly reduce the extent of missingness, however, it cannot be completely eliminated due to the lack of auxiliary information required for imputation. Figure 1 illustrates how the extent of missingness changes after imputation is performed.

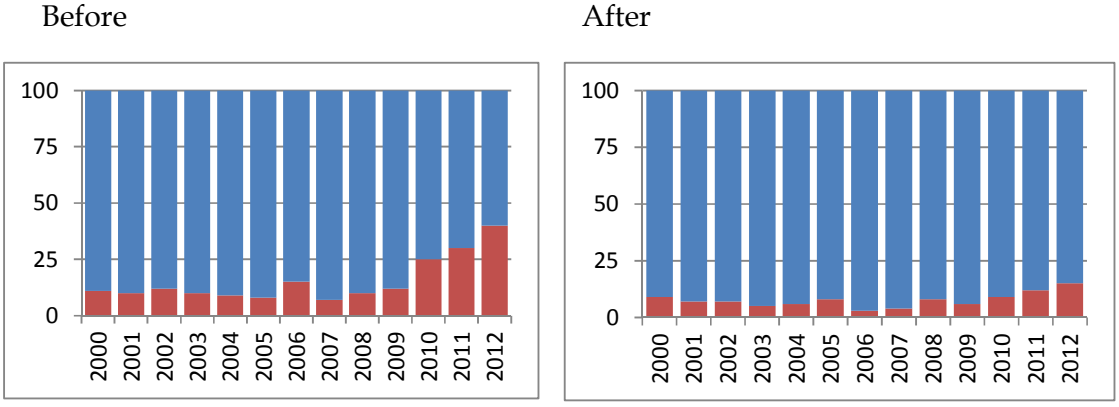

Fig 1: Extent of missingness over time before and after imputation

While for interpolation, officially reported data might be available for any auxiliary variable to be used as a basis for imputation, the same does not apply to extrapolation. Suppose the estimate of gross output (EGO) is to be determined for any year *t,* which can be obtained from its value for the past year t-1 multiplied by the ratio of volume (IIP) and price (PPI) changes during the observation period. The estimation is done as follows:

$$EGO_t = GO_{t-1} * \left(1 + \frac{IIP_{t:0} * PPI_{t:0} - IIP_{t-1:0} * PPI_{t-1:0}}{IIP_{t-1:0} * PPI_{t-1:0}}\right) \qquad (1)$$

In this example, data for GO_{t-1} and *IIP* for period $t-1, 0$ *and* t are available in the UNIDO database from official sources, but PPI data are obtained from external non-official sources.

In other cases, the total data industry value added (IVA) is available from official sources, however, it has to be split to manufacturing value added (MVA) and mining and utilities value added (MuVA):

$$\begin{aligned} EMVA_t &= s_t^{\,1} * IVA_t \\ EMuVA_t &= s_t^{\,2} * IVA_t \end{aligned} \quad \text{where} \quad s_t^1 + s_t^2 = 1 \qquad (2)$$

The notation s^1 and s^2 denote the share of MVA and MuVA in IVA. Data for these shares are obtained from non-official sources.

UNIDO Statistics has carried out several studies on imputation schemes applicable to different data sets. This scheme is gradually evolving into a manual that will guide the entire imputation process using both official and non-official sources. The implementation of an imputation scheme has not yet fully materialized.

5. Conclusion

Non-official data sources are an important part of statistics disseminated by international agencies, however, there is no consensus among statisticians on the content and use of non-official data. The main concern is quality rather than type of data source. It is generally assumed that the methodology applied to produce official statistics is transparent and meets quality assurance criteria set by the national statistical institution. However, the mean of verification of such an assumption is not particularly strong. At the same time, international agencies can only assist NSOs in improving their data, but cannot run their own data collection programme.

Based on experience, UNIDO seeks to maintain the official status of the reported data as best as it can. Non-official sources are used for imputation purposes, thereby improving the quality of data in terms of coverage, timeliness and international comparability.

6. References

1. Boudt, K., Todorov, V., Upadhyaya, S. (2009). Nowcasting manufacturing value added for cross-country comparison. *Statistical Journal of the IAOS: Journal of the International Association of Official Statistics* 26:211-252.

2. Boudt, K. (2010) UNIDO Statistical Database: Methodological Notes Regarding Stage 4 and Stage 5; Study Report

3. Templ, A. M., Alfons, A., & Kowarik, A. (2010). Package " VIM "

4. Todorov, V., Hamel, N. UNIDO Statistical Database: Handling of missing values and imputation, Working paper, UNIDO, 2013

5. UNIDO: UNIDO Industrial Statistics Database: Methodological notes, 1996

6. Upadhyaya, S., & Todorov, V. (2008). UNIDO Data Quality: A Quality assurance framework for UNIDO statistical activities. UNIDO Staff Working Paper.