



Session 4: Best practices in imputation and estimation

Imputation in UNECE Statistical Databases Principles and Practices

Steven Vale and Heinrich Brünger, UNECE

Introduction

The issue of imputation by international statistical agencies as a means to produce harmonized and comprehensive statistical outputs was given prominence in the adoption of Resolution 2006/6 by ECOSOC (the United Nations Economic and Social Council). Paragraph 5 of this resolution sets strict limits for the use of imputation in the production of data on progress towards the Millennium Development Goals, but in doing so, it acknowledges that, within these constraints, imputation is a valid technique for data production within international statistical agencies¹.

The UNECE currently makes only very limited use of imputation techniques during the production of statistical data series. This affects the completeness of time series, and also places serious constraints on data production and dissemination activities. For example, many aggregate data for groups of countries are missing, because, as a general rule such aggregates are only currently produced if reliable data can be obtained for all of the countries within the grouping.

This paper describes how the UNECE policy on the appropriate use of imputation techniques within our statistical database is evolving to overcome these constraints. It starts by taking stock of the rather limited current practices regarding imputation, then sets out a series of basic principles in line with the ECOSOC resolution, and considers how these principles provide a framework for extending the use of imputation in the future.

The proposed approach is to increase the use of imputation step-by-step, without compromising on data quality. This allows the opportunity to pause and reflect, and to properly assess the impact and risk, each time an additional step is implemented. The implementation will take place in the context of the current re-engineering of the UNECE statistical database system, and will be automated as far as possible, based on a set of methods and algorithms designed to give statisticians a certain degree of choice concerning the most appropriate methods for different circumstances.

¹ The implications of this resolution are examined in more detail in the UNECE paper prepared for the September 2007 meeting of the CCSA (Committee for the Coordination of Statistical Activities): <http://unstats.un.org/unsd/acsub/2007docs-10th/SA-2007-11-ECERep.pdf>

Current practices

We define imputation as “a procedure for entering a value for a specific data item where the response is missing or unusable”². Imputation is currently only used in the following four specific circumstances:

1) Computing data according to definitional identities

In the context of National Accounts, missing values can sometimes be derived from the available data through the use of logical account relationships. For example, a sum can be calculated if the components are present. Whilst it could be argued that this is not strictly imputation if one source is used, sometimes it is necessary to use two or more sources, and any discrepancies between these sources can introduce a probabilistic element to what would otherwise be a deterministic process. The size of any such discrepancies is used as a quality criterion in deciding whether or not to publish such imputations.

2) Imputation of regional aggregates

In a very limited number of cases, regional aggregates are produced where not all of the data for the countries within the grouping are considered fit for publication. This is only done where the poor quality data have such a low weight in the calculation of the aggregate, that their potential lack of accuracy has no real impact on the aggregate at the level of precision to which it is published. A practical example is that estimates for some National Accounts series for Uzbekistan are used only for the compilation of aggregates for the Commonwealth of Independent States (CIS), and for UNECE totals, and are not published in their own right. In the future, a threshold of 10% of the total weights will be used to determine whether such cases have a sufficiently low impact on the resulting aggregates, but in present cases the proportion is much lower (typically 1-2%).

3) Reclassification

Some countries do not use standard international classifications (e.g. ISIC/NACE classifications of economic activity), or do not provide historical data on this basis. In such cases, data are converted according to a mapping of the different classification systems. Where such a mapping is not on a clear-cut one to one basis, this introduces an element of imputation within the definition above.

² Source: UNECE Glossary of Terms on Statistical Data Editing, and the SDMX Metadata Common Vocabulary

4) Using imputations made by others

Where data are sourced from other international statistical agencies, these often contain imputations, particularly for totals. If the source is considered to offer sufficient quality, the imputations are considered to have the status of official data, and are therefore suitable for publication. One specific issue is that such data are not always flagged as imputed in the source files, so it is not always possible to identify them. In such cases, the source data do not comply with the transparency requirements of the Principles Governing International Statistical Activities³. Full compliance with these principles would be our preferred indicator of quality for data sources.

Basic principles for extending imputation in UNECE databases

The following basic principles have been derived in line with our interpretation of the ECOSOC Resolution:

- National data that are imputed to allow the production of aggregates for groups of countries will not normally be published in their own right, except where these imputations are already published by other international agencies, or the country concerned gives explicit consent to publication.
- Only “official” data sources will be used for the purposes of imputation. These are defined for this purpose as data from national or international statistical agencies.
- Imputation should be based on real data from the same country, rather than data from other countries, unless there is a strong reason to assume that countries will have very similar values (e.g. certain financial variables for countries in a monetary union).
- A clear distinction between “real” and imputed data will be made throughout the statistical database system.
- Any data that are imputed, or are derived, wholly or partly from imputed data will be clearly flagged as such to users, and full documentation of imputation methods will be provided.
- Aggregates for groups of countries will only be published where at least 90% of the total is based on real data (or 90% of the total of the weights for indices and rates), and if real data for more than 50% of the countries within the grouping are present. (This last clause prevents data for “North America – 2” being calculated if Canadian data are missing).
- To allow for the possibility of revisions to earlier data, all automatically imputed values will be deleted periodically, and re-calculated. Statisticians will be responsible for maintaining manually imputed data following revisions.
- The method to be used for automatic imputation will be defined at the level of the variable, and stored as an attribute of the variable within the database system. The default will be “no imputation”.
- Decisions on whether to apply imputation in specific cases, or groups of cases, should be taken with strict regard to the quality framework.

³ See:

http://unstats.un.org/unsd/methods/statorg/Principles_stat_activities/principles_stat_activities.asp

Applying the principles to extend imputation in a step-by-step approach

The intention is to develop automatic imputation routines based on these principles in a pragmatic and step-by-step approach, starting with the routines that are simplest and give the greatest expected benefit in terms of a trade-off between quality and quantity of imputation. After each step, there will be a pause and review period, during which the quality of the outputs will be examined, and the operation of the routine may be improved. Strict attention will be paid to costs and benefits, where costs will be measured in terms of the opportunity costs of not using our limited resources for other database system enhancements, and benefits will be quantified in terms of additional data cells published.

The proposed imputation routines will be implemented in the context of the re-engineering of the UNECE statistical database systems, introducing options for automatic imputation alongside existing calculation routines. It is envisaged that over time several methods will be made available, allowing statisticians to choose the one that is most appropriate for each data set. An option to input manually imputed data will also be made available, though this should be used only in exceptional cases, and the manual imputation methods must be fully documented.

The first method to be implemented will consider national time-series data, and will construct a linear trend, which can then be used to impute any missing values. This method clearly works best with time series that closely follow linear trends, i.e. the coefficient of determination (R^2) is close to one. A quality criterion for using such an approach is therefore that R^2 must be greater than a certain threshold (to be determined). The draft rules to apply this method in practice are set out in the Annex.

In practice there may be some quality issues in deriving R^2 values for a trend based on as few as three observations (as proposed in the Annex). This limitation is imposed by the shortness of some time series. Another limitation is that, to simplify the programming requirements the method proposes only to construct trends using data for periods before the missing period (for example, if data for 2003 are missing, the trend would be constructed using the available data from the periods 1998 to 2002, and not data for 2004 or later periods). In practice, this is not a significant limitation, as the vast majority of cases where we would want to apply imputation concern missing data for the most recent periods. Obvious future enhancements would therefore be to add the flexibility to use longer time series where they are available, to use data for periods after the missing period, and to use non-linear trends. These options would add significantly to the complication of the system, so would have to be justified in terms of the value of the outputs produced.

Other future steps could include a routine to impute data based on values for the same period for “similar” countries, however to comply with the third basic principle above, this would need to be done under strictly controlled conditions. The limited benefit therefore means that this is not a priority for automatic imputation. Cases using this method will be treated manually, at least initially.

Conclusions and open questions

The practice of imputation is very strongly linked with the concept of quality. Imputation can be seen as improving accessibility by allowing the provision of additional data, but is likely, in most cases, to have an adverse impact on accuracy. Quality constraints should therefore drive any imputation process, setting clear limits as to what is acceptable. Publishing poor quality imputations, particularly if they are not flagged as such, risks serious damage to user confidence and the public perception of an organisation. The step-by-step approach to implementing imputation, with quality reviews after each step, is strongly recommended, (particularly for automatic imputation).

Transparency is vital for imputation, methods and decisions must be clearly and fully documented, and made available for scrutiny by users. In this context, it is also clearly desirable that imputation methods and practices should be standardized between statistical organizations, as far as possible. An inventory of methods and tools would be a useful first step. To make further progress, a task-force or similar group may need to be convened.

This gives rise to the following open questions, which are posed with the intention of stimulating debate on imputation policies and practices. Some questions are carried forward in part from the paper presented by the UNECE at the September 2007 meeting of the CCSA:

- 1) To what extent are international organizations interested in defining a common policy on the use of imputation, in response to the ECOSOC Resolution?
- 2) Could we go further and consider harmonization of methods and tools?
- 3) How should this be done? Is a specific forum needed, or can imputation issues be sufficiently dealt with in combination with work on data quality?
- 4) Have other organizations modified their policies on imputation in the light of the ECOSOC Resolution, and if so, how?

Annex – Proposed rules for implementing automatic imputation using linear trends

Step 1 – Do real (i.e. non-imputed) data exist for at least three of the five previous consecutive periods? If not, the automatic imputation routine stops.

This rule has three effects; firstly data can only be imputed if the series itself already includes data for at least three periods. Secondly, data cannot be repeatedly imputed based on automatic imputations from the previous years. Thirdly, data are less likely to be imputed if the time series has not been compiled for each of the last five periods.

To illustrate this last point, consider a series for which values are held for 1990, 1995, 2000, 2004, 2005, 2006 and 2007. Data for 2006 could not be imputed, because out of the previous five consecutive years (2001, 2002, 2003, 2004 and 2005), it is only possible for data for two years (2004 and 2005) to exist. However data for 2007 could be imputed if real data for 2004, 2005 and 2006 are present. In this way, older data values (e.g. for 1990, 1995 and 2000), which may reduce the quality of imputation, cannot be used.

The logic behind the programming of this step should therefore be:

- If a regional aggregate (RA) is missing for period t , retrieve data for that aggregate and the country data used to compile it, for the previous five consecutive periods (RA_{t-1} , RA_{t-2} , RA_{t-3} , RA_{t-4} and RA_{t-5}).
- Determine the countries within the grouping for which data are missing.
- If real data exist for all of those countries, for at least three of the previous five consecutive periods, continue, otherwise stop, and do not impute a value for the aggregate.

Step 2 – Establish a benchmark period to determine whether the missing country or countries contribute less than ten per-cent to the value of the regional aggregate.

The logic behind the programming of this step should be:

- If data are available for all countries in the grouping for at least one of those five periods continue, otherwise stop, and do not impute a value for the aggregate.
- Determine the most recent of those five periods for which data are available for all countries in the grouping ($t-n$).
- Compute the percentage contribution to the total in period $t-n$ for the countries for which data are missing in period t .
- If that figure is less than or equal to 10% continue, otherwise stop, and do not impute a value for the aggregate.

Step 3 – For those countries with missing data in period t , calculate the linear trend for the time series over the five previous consecutive periods, and calculate the

coefficient of determination (R^2) for that trend. If R^2 is greater than x (value to be determined), continue, otherwise stop the imputation process.

Step 4 – Use the linear trend to derive the imputed value for that country for period t .

Note: If data for $t-1$ are missing, the value obtained from step 3 should be doubled and added to the value for $t-2$ to get the imputed value for period t . If data for $t-1$ and $t-2$ are missing, the value obtained from step 3 should be tripled and added to the value for $t-3$ to get the imputed value for period t .