

Improving imputation methodology in the Hungarian Central Statistical Office (HCSO)

Zoltán Csereháti

Hungarian Central Statistical Office, e-mail: zoltan.cserehati@ksh.hu

Abstract

To have sound imputation methodology in the official statistics it is important to collect, categorize and assess the methods which are currently in use.

In the HCSO there is a project currently running with the aim of Creating an Integrated Data Processing System “IDPS”. The IDPS system will enable us to build more complex and tailor-made imputation methods for different survey processes.

A documentation scheme for imputation has been elaborated to be filled in for all the surveys in the HCSO.

To stimulate the improvement of this field, an internal training course is being organized regularly on imputation methods in the HCSO, presenting not just the basic methods and recommended practices, but some more sophisticated case studies as well.

Keywords: Imputation, documentation scheme, integrated data processing system

1. Introduction

The Methodology department in the HCSO is gradually widening its scope of work and responsibility on the coordination and management of the applied methods in different subject matter areas in two dimensions: in the scope of phases and tools offered.

Doing sample selection, estimation or standard error calculation one needs to compile procedures which are typically quite sophisticated and therefore almost exclusively done by methodologists. Concerning the processing phases our central methodological unit is dealing with sampling and estimation, imputation, seasonal adjustment, data confidentiality, gradually widening the list. Concerning the tools offered, we developed quality guidelines for the elements of value chain, and also a methodological documentation scheme, good practices, quality indicators, methodological support, training course material, quality assessing tools and handbook for several phases.

Each statistical survey face increasingly problems related to missing data. Non-response errors badly affect the quality of the data, as it introduces a bias which is hard to reduce. Selecting larger samples is not a solution to this. Coping with missing data and reducing non-response bias is a central methodological problem.

It is important to take into account the possible solutions. The main alternatives are using different re-weighting or imputation methods. In some areas no effective imputation method can be used because the nature of the survey. To solve these problems one can apply only reweighting methods. However, in most cases implementing some kind of imputation method yields the best solution.

There is a huge variety of imputation methods. Many of them are quite simple and easy to implement. This is the reason why unlike the above-mentioned methodological areas imputation is a processing phase which is often conducted by subject matter statisticians without the supervision of methodologists.

From the methodologist point of view it is crucial to have an overview of all the implemented imputation methods, especially of those, which are applied without the assistance of the methodological department.

Supposedly many of these methods could be improved. Surveying the currently applied methods must be the first step towards the aim of building tailor-made imputation methods for all the statistical surveys.

A documentation scheme for imputation (Chapter 3) has been elaborated as a uniform basis for conducting this survey. Similar surveys for other methodological areas are also being conducted.

In the past years an internal training course on imputation (Chapter 4) has been conducted in the HCSO. The content of the course and the gained experiences during the lectures will form the basis to build the thematic structure of the handbook.

In the near future we plan to compile a handbook on imputation (Chapter 5)

Hopefully all these efforts will help us to apply sufficiently sophisticated imputation methods all over the office, which comply the corresponding methodological guidelines.

2. The "IDPS" (Creating Integrated Data Processing System) project

In the last year, HCSO launched an ambitious new project called IDPS.

The objectives of this project are the following:

—to develop user-friendly integrated data processing system based on standard logic covering the widest range of surveys accessible via a standard user interface and providing a clear and efficient tool for the statisticians, which include data quality requirements and data processing procedures documented in the meta-database and integrated with other general purpose systems such as data entry, dissemination;

—to develop applications or frame systems allowing coordination and quality management in the control of processing and direct access to data for the purpose of verification and analysis;

—by developing these data processing systems to restructure the division of labour with the IT staff focusing on innovation, development and production quality data faster through direct data processing.

After a public procurement procedure an appropriate IT company has been chosen to build the desired system in strong collaboration with our IT experts and methodologists. On behalf of the HCSO, the project leadership is the responsibility of our IT Department. It is important to emphasize that the main aim of the project is to build an integrated, well-performing IT system. We do not want to settle strict methodological standards. The so-called “standards” of the IDPS system will be optimally designed software components for implementing different algorithms and procedures which are useful as building blocks to compile the IT version of different methods.

We anticipate having a working system by the end of 2010.

The potential benefits of the new system are:

- A common, integrated platform for all of the statistical surveys.
- Less redundancy. Now, there are multiple IT solutions for some basic methods. In the new system these methods will be implemented as basic routines which can be referenced from a data processing stage of any survey, and can be appropriately parameterized.
- A more transparent system. All the process flows will be documented in a standard way. Both the IT experts, the methodologists and the subject matter statisticians will have an access to the system functionalities which will help to gain a good overview of the process plans and the actually running process flows as well. Thus, more functionality will be by the hand of the subject matter statisticians.
- The system will enable the statisticians to build new data process flows more easily.

Main steps of the preparatory work already done:

- Documentation of the data process flow elements in the different subject matter areas.
- Designing a general scheme for a universal data processing flow.
- Identifying process stages such as editing, imputation, outlier filtering, consistency checking, etc.
- Identifying basic methods currently in use in the different stages.
- Identifying process steps from which the individual implementations of the methods are built from.

The work will be done by collecting, analysing, categorizing, assessing and synthesizing the methods currently in use all over the office. The assessing phase is being done by the overview of the processes from the ITs point of view. Synthesizing means to build a coherent and well-performing IT solution, which integrates all the informatical components needed for the management of all the data processing flows. The system logic plan is being elaborated right this time. The work on the physical implementation of the system plan will be done later this year.

Although the aim is to build an integrated IT system, the methodological support is crucial, by contributing to a handbook on standard methods, to help process documentation. Besides the documentation task it is important to identify common standard methodological modules in use, which are not included in the handbook.

This documentation gives the opportunity to perform methodological tasks in the process of imputation too. Specially, we have surveyed the existing imputation methods all over the office. We are now assessing the currently used methods.

We pay an extra attention to the composite methods in use. In many cases a combination of different imputation methods yields an optimal solution for a special non-response problem. The IDPS system has to support the compilation of such methods besides implementing the existing ones. It will be a critical point of the system development to assure an optimal solution.

In the next step we will try to identify standard processes which will serve as the building blocks of the IDPS system.

What does the ideal standard process look like?

- It is small and special enough to serve as a building block.
- It is flexible and general enough, having a number of parameters for fine tuning, to be suitable for various purposes.
- Thus, building standard modules one will inevitable face a difficult trade-off situation.

3. A documentation scheme for imputation

In the past years a number of documentation schemes have been elaborated with the purpose of having a thorough view of the different methodology related features of all the survey processes in the HCSO.

The affected methodological areas are sampling, imputation, estimation and standard error calculation, seasonal adjustment and confidentiality.

The main aims of these documentation schemes are to build a uniform structure to gain a better overview of the methods used by various surveys and to improve process quality.

These schemes are formulated as a special type of questionnaires. It is the task of the appropriate subject matter area experts to fill in these questionnaires. This task is currently in progress.

As mentioned above, one of the affected methodological areas is the processing phase of imputation. The main questions of the scheme are the followings:

First some general information: name and code of the survey. Purpose, frequency, main variables observed. Data collection mode. Main population and sample characteristics.

Then it is being asked to name the most important variables.

The following questions are related to the applied method in general. First we ask whether there is a working imputation method, and if yes, whether it is done regularly, automatically.

It is important to ask if the imputed values are flagged or not. From a systemic point of view it is crucial to clarify whether the applied method is reproducible.

Sometimes there is a guideline of some international organization which contains regulations concerning the imputation methodology to be applied. Therefore we ask the subject matter statistician to give sufficient information related to this question.

It is a crucial point to clarify whether the imputation procedure is documented in detail, and if yes, where it can be accessed.

There are questions relating non-response. First, it is important to ask if the cause of non-response is coded and recorded, and if yes, whether this information will be used during the imputation process.

An important related question is: What are the non-response rates for the main variables?

Concerning the imputation process itself first we ask whether the imputed data are micro-data or some kind of aggregates

In which phase of the data processing chain is imputation done? For example logical imputation in the data capturing phase, or after deleting some data in the consistency checking phase, or after comparing the data to another datasets.

The next few questions are related to the software part of the implemented methods. First we ask if there is any software in use to assist the imputation. If yes, is it a specially designed software solution or a general system? Are there any other self-programmed algorithms in use?

Non-response problems arise in a different manner in special cases: some surveys are suffering mainly from unit non-response, while others are typically affected by problems arising from sporadic item non-response cases. Therefore it is important to ask how different non-response patterns (item-, partial, or unit non-response) are dealt with.

To clarify the severity of the problem situation a basic question is what proportion of the records and values is affected by the imputation. Another important issue is what the typical reasons of the missing data are.

The next question is related to the auxiliary data sources: what is the information used for imputation? For example donor records from the same time period, donors from past periods, data from relating datasets from the same or past time periods, data from registers, data from external data sources (TAX authority, National Bank, Ministries, other authorities), or mixed data sources.

As more sophisticated methods need special attention, it is important to ask whether the implemented imputation method is a simple one or it is a composite method. (If it is a composite one, which are the building blocks of it? How is it built from these blocks? What does the appropriate flow chart look like?)

In the next section of the questionnaire we give an overview of the most important imputation methods, and ask to choose those ones which are being used.

As an example, here is the corresponding part of the questionnaire:

Indicate the method used for imputation!

(If you apply different methods, please indicate all of them.)

- Rule based Imputing methods
 - Deductive imputation (logical reasoning)
 - Using external rules for imputation
- Model based methods
- Explicit models
 - Imputing with mean
 - Imputing with population mean
- Imputing with class means
 - Imputing with the median
 - Imputing using random numbers
 - Regression based imputations
 - Impute values using simple linear regression
 - Imputation based on multivariate regression model
 - Predictive mean matching
 - Other (please specify):
 - Implicit models
- Hot deck imputation
 - Sequential hot deck method
 - Hierarchical hot deck method
 - Nearest neighbour method
- Cold deck imputation
- Other (please specify):
 - Time series models
 - Sophisticated models
 - ARIMA models
 - Kalman filter models
 - Simple models
- Imputation methods based on population dynamics considering consecutive time periods
- Imputation methods based on simple trend analysis of longer time series
 - Imputing using neural networks
 - Other (please specify):

Imputation methods are mainly used to reduce non-response bias. However these methods themselves may introduce some amount of variance. We must be aware of this fact, so it is important to ask whether there is a multiple imputation method applied, and whether an estimate for the variance induced by the imputation itself is being calculated.

4. Internal training course on imputation

Some basic features: everyone uses imputation methods but in many cases these are simple heuristics based solutions lacking in-depth studies.

Many of the subject matter statisticians are not aware of the fact that a well-performing, sufficiently tailor-made imputation method may be quite complex and sophisticated.

As for the main points of the course, before all it is important to clarify what exactly the concept of imputation covers.

Then we must pose the question: Why imputing at all? What are the problem situations which can be solved by imputing?

To meet the needs of the different subject matter areas it is important to point to the drawbacks and benefits of different methods and the specialities which determine the frames of the application areas.

It is important to emphasize that the main reason for imputing is to reduce non-response bias. To make this clear it is crucial to present the main differences between variance and bias, and why the latter is much harder to treat with.

Imputing fills in the gaps of the datasets. There are obvious benefits of complete datasets which are important to emphasize. (Several analysing techniques do not work on incomplete datasets.)

Once again, after getting acquainted with some methods it is important to clarify, what is an imputation method good for, and which kind of problems can be solved, or at least handled by it.

To get answers to these questions it is worth to deal with problems relating non-response in details. What are the different characteristics of unit- and item non-response, which problems arise relating non-response bias and how to compensate it for are the most important issues.

It is worth to mention that there are alternative solutions to imputation methods. A number of weighting techniques are being presented briefly during the course.

A much more practical, still very important issue is how to build a strategy, how to organize a method building process. A subject matter statistician may need the assistance of experienced methodologists and IT experts, too. It is a long iteration process of subsequent negotiations to get a sufficiently sophisticated method.

As auxiliary information play a key role in many imputation methods, it is important to clarify, how to choose appropriate data sources for imputation. Solely having a full scope view of the potentially applicable data sources is not an easy task.

To make the concept of imputation more clear it is worth to get back to the question: what is different in editing and imputation? There are some logic-based methods which are sometimes classified as editing methods, while elsewhere mentioned as imputation.

The next part of the course deals with basic imputation methods. During this part of the course the most important basic methods are being presented, with small artificial data examples to make easier to comprehend.

To have appropriate documentation is a crucial point in all methodological areas. It is the case by the imputation methods, too. One needs detailed description of the methods used: flow charts, algorithmic descriptions. It is important to flag the imputed values, and indicate when, why, how, by whom it has been imputed. A strongly related problem is that one has to try to avoid the use of imputed values as donors in consecutive time periods. To avoid the related problems we need to have reliable documentation of the data. The flagged values help us to avoid the usage of low-quality imputed values of past periods.

To follow a systemic approach it is crucial to determine the place of imputation in the whole data processing flow.

Interconnections between imputation and outlier-filtering are also important to clarify. For example we must not use an outlier as donor. Equally important is that a known outlier should not be imputed by a regularly behaving donor.

Then we try to give some usable advice, how to plan and assess an imputation method. To do this, a useful tool may be implementing simulation studies.

In the final part of the course we present some case studies.

Two special examples are the multiple donor imputation method applied by the survey of internet service providers, and the multi-stage donor selection method used by the survey of the non-profit sector. These are examples for tailor-made composite methods.

The course material has been complemented by a teamwork session. We select a practical problem and try to solve it together in teams. During this session we share the experiences and ideas related to the course material, too.

5. Conclusion, future work

Concerning the methodological area of seasonal adjustment as an example, the experiences of regularly held internal training courses, and the collected and analysed information gained from the filled in documentation scheme formed the basis of compiling the list of current best practices, and writing a handbook. This handbook will be in line with the Quality Guidelines of HCSO and other subject related international documents.

As a plan for the near future we are thinking about compiling a similar handbook on imputation which will contain all the recommended methods and detailed guidelines with specific areas of use. This handbook will serve as a guideline for internal use in the HCSO.

There are several international projects, guidelines, and handbooks on imputation. Some examples: the ONS paper Report on the Task Force on Imputation, The Statistics Canada Quality Guidelines, the results of the EUREDIT project, and the EDIMBUS project. We take into consideration the results of these projects and try to implement these to the special needs of the HCSO. The experiences gained during the work on the IDPS system, the feedbacks from the training course and the information collected by the documentation scheme together will form the basis of this work.

The handbook will contain guidelines on how to build an imputation method. It will list the basic methods with possible application areas, highlighting current best practices. It will formulate practical advices, focusing on issues related to Hungarian specialities.

References

The results of the EUREDIT project:

<http://www.cs.york.ac.uk/euredit/results/results.html>

The results of the EDIMBUS project:

<http://edimbus.istat.it/EDIMBUS1/>

The ONS paper: Report on the Task Force on Imputation (June 1996) GSS Methodology Series

Statistics Canada Quality Guidelines (Fourth Edition 2003)

Quality Guidelines of the HCSO (Legal Act 2007)

Hungarian Central Statistical Office: Strategy 2005-2008, pages 26-27.

Cserehádi, Z. (2006) *Multiple Donor Imputation Techniques*, Paper for the European Conference on Quality in Survey Statistics, Cardiff, 24-26 April 2006