UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Budapest, Hungary, 14-16 September 2015)

Topic (ii): Managing and supporting changes related to editing and imputation

Data collection optimization – first attempt

Prepared by Agnes Andics and Gergely Horváth, HCSO, Hungary

I. Introduction

1. A general aim of the business processes of National Statistical Institutes (NSIs) is to produce reliable estimations from the available data as soon as possible and for as low cost as possible. In an ideal world for official statistics all data needed for processing the outputs is available in time and statisticians have to work on estimates and releases only. Unfortunately, this is not always the case.

2. Despite the trend of using more and more administrative and other secondary data, every now and then NSIs encounter the problem of non-response. Looking around in European countries one can find several methods for the treatment of this problem (see Fabuel et al., 2013). In case of surveys a traditional approach to manage missing information is to gather data from the (non)respondents even after the deadline of sending back the questionnaires.

3. At the Hungarian Central Statistical Office (HCSO or Office) – beyond the fine the Office is allowed to charge data-non-suppliers – there is a reminder system in place to remind data providers to complete their assigned questionnaires in time. This system consists of many steps. Although the system reaches its goal automatically most of the time, i.e., using only the automated electronic reminders, from time to time many non-respondents have to be contacted via phone calls or letters which makes data collections still very costly. The question from cost-benefit point of view is whether it is worth making further efforts to collect information from the (non)respondents after the deadline or data quality will not improve significantly to justify these efforts and thus the work of employees could be rearranged for other data collections or other tasks.

4. Our aim is to create a highly automated procedure that enables the HCSO to track the progress of the estimations made at any moment of time during the data collection period and indicates whether the available data for a given survey is good enough both in quantity and quality. Using such a system results immediate gains in timeliness as well. This approach was very much motivated by the presentation of the Canadian Rolling Estimates on the EESW workshop in Nuremberg in 2013 (see Saint-Pierre, 2013).

5. This paper is organized as follows. Section II describes the theoretical framework of the planned procedure. In section III some pilot studies and their consequences are presented. Section IV gives a conclusion and raises some open questions.

II. Theory

6. Our aim is to model the processes (in our case process starts after data editing and includes imputation, outlier treatment and estimation), give daily early estimates by using the existing procedures of each survey process and check whether the current pre-estimates have already reached an acceptable threshold of accuracy. Let us analyse the task in details.

A. Basic information

7. For each survey *key variables* have to be determined for which the pre-estimations will be computed. These variables will guide our procedure since editing and imputation strategies have to concentrate on them as well. Subject matter statisticians are responsible for defining these key variables.

8. Expected data suppliers have to be known in order to use their data available in the statistical registers or recorded previously, for editing purposes. These data will later be used for our pre-estimations as well.

9. *Key respondents*, those data suppliers that have high impact on the estimations also have to be defined in advance. Information on the impact can be gathered either from previous survey data or from media or other sources. Each data supplier has to be assigned with an impact factor by the subject matter statisticians, and selective editing should be performed in alignment with the order based on this factor.

10. Responses need to be monitored. A daily evaluation of the respondents according to their answers is necessary to decide whether they belong to the target population or not, and the *rate of over-coverage* has to be determined. *Response rate* has to be computed as well in order to see whether the pre-estimation process can already be started.

11. Here the procedure may split in three parts:

(a) a separate system has to deal with those who don't belong to the target population (e.g., update the registers, give feedback to the field statisticians);
(b) another part has to investigate non-respondents' characteristics. The result of their analyses may influence the imputation techniques used later on;
(c) the rest of the responses with data has to be then edited.

B. Editing and imputation

12. In order to give daily pre-estimations by using the existing procedures, data have to be available in good quality already. This means that data editing has to be launched on a daily basis.

13. For reaching good quality of data it has to be checked first whether there are missing values for the key variables from the key respondents necessary for the pre-estimations. Then, again for the key variables, as many correspondence and constraints have to be evaluated as needed. Simultaneously, selective editing has to be performed for the predefined key respondents as well as for those data which turn out to make an influence on the estimations in the meantime.

14. Editing is in strong connection with item imputation. There is either a chance to correct the erroneous data by using automatic data editing procedures or data has to be deleted and imputed from a given – statistical or administrative – data source determined in the frame of the imputation method. Potential data sources for imputation have to be mapped thoroughly in advance.

15. There has to be an imputation policy for each survey process for unit non-response as well. Note that, unfortunately, currently existing procedures do not involve imputation methods in some cases, e.g., some full-scope surveys have no real imputation policy. Statisticians have a certain period after the deadline for returning the questionnaires and before the process starts – a so called "patience time" – when they post-collect as much data as possible from the non-respondents. Following this strategy they can achieve high response rates which does not make unit imputation practice necessary.

16. An optional but important part of the process of surveys using sample is the outlier treatment. Producing estimation from a sample, statisticians have to decide whether the sample contains any outlier. There must be then a treatment of outliers which should be built in the process depending on the imputation method used.

C. Accuracy

17. Accuracy is the crucial point of the system. First a decision has to be made on which measures shall work as quality indicators and how they shall be used. Then accuracy thresholds or targets can be set up based on analysis of previous data and quality requirements.

18. As *response rates* work as the starter of the process, *imputation rates* both on key respondents and on the whole set of respondents should act as indicator of the potential bias. *Coefficient of variation* (CV) of the pre-estimations is another obvious choice. Compute the CV of the pre-estimations for each day and check whether it changes significantly from day *d* to day *d+1*. Combination of these quality indicators using with flexible and well-determined thresholds may give answer to the question whether the follow-up can be stopped.

III. Practice and pilot studies

A. IT Systems

19. The Integrated Data Entry and Validation System (ADEL) in the HCSO supports the validation and correction of survey data. In 2014, there have been 130 surveys involved for which the data have been received through the Electronic Data Collection System (ELEKTRA). Although the ADEL system fully supports the validation by checking and scoring as many constraints for the data as reasonable (see Kómár, 2015), corrections are usually made manually.

20. Integrated Survey Management System for Establishment Surveys (GÉSA) supports the data collection management of establishment surveys where the data providers are enterprises having the official 8-characters ID number, primarily based on the Business Register of the HCSO. Its main tasks are the design and documentation of data collection management, selection of the data providers, producing, personalizing and sending the questionnaires, collecting and monitoring the questionnaires, support data process. As such it gives information on the over-coverage and reason of nonresponse via a coding system.

21. The HCSO has developed and is now implementing an Integrated Data Process System (EAR) for the process steps after data editing. Currently there are 63 surveys ready for using the system and the final goal is to involve the rest of the establishment surveys till the end of 2017 which means about 190 processes done by EAR. The system is able to handle the imputation, outlier treatment, weighting and estimation steps, can produce input files for Demetra (for seasonal adjustment), and outputs and aggregates at the end of the processes. It is required to compute quality indicators as well, this function is currently under development.

22. All systems are metadata-driven and provide great support for the regular process of the data. However, in order to give a daily estimation for each survey process, strong coordination is needed taking into account the IT capabilities (running times, processor capacities, etc.).

B. Step by step

23. For our purposes, the ADEL system provides the data and information on the data (metadata). The computations, programs have been built on these databases by using SAS and R software tools since the process for the surveys we picked up for analysis is not yet prepared in the EAR system, and on the other hand we planned to follow the trend in the ESS to work with IT tools easy to share. Surveys used for pilots at different stages were the Annual statistical survey of production (OSAP[1] 1039), the Questionnaire on the Current Environmental Protection Expenditures and on the Environmental Protection Investments (OSAP 1799) and the Monthly Statistical Survey of Industry and Construction (OSAP 2235) with their previous year's data.

[1] OSAP is the abbreviation for National Statistical Data Collection Program. Each survey has an identification number in the frame of the Program which we use as reference number.

24. Our first task is to monitor the responses according to their arrivals. This can easily be done by using the date stamp stored in the ADEL database which shows when the questionnaire was returned. On Figure 1 the cumulative distribution of questionnaires return for the Questionnaire on the Current Environmental Protection Expenditures and on the Environmental Protection Investments over time is presented for year 2014. The deadline for sending the questionnaires is 31st of May each year. One can see the following:

(a) about half of the questionnaires arrived before the deadline;
(b) within some days another quarter of the questionnaires was sent thanks to the first reminder two days after the deadline;
(c) then there is a slow, moderate increase in the arrivals.

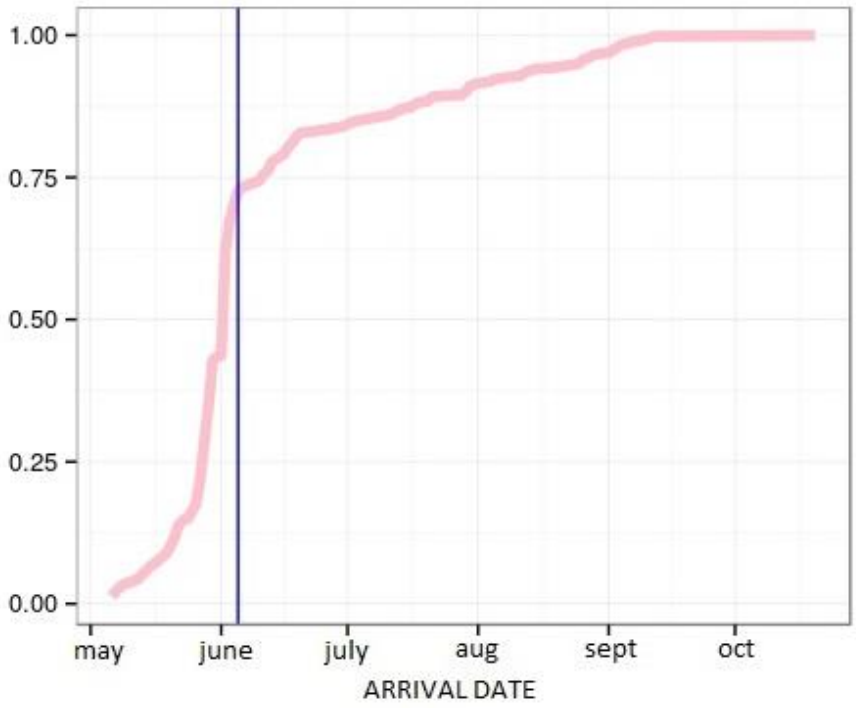

Figure 1
Cumulative percentages of questionnaire arrivals of survey OSAP 1799

25. The shape of the graph raised the question: when should the pre-estimation process be started? In case of many surveys the deadline is a feasible criterion for the start. Although deadlines were set up in such a way that most of the data are preferably available by that time, in some cases response arrivals may depend on other circumstances. Note here, that actively monitoring the questionnaire arrivals may give help in changing the deadline for some surveys.

26. Let's suppose we start our procedure immediately after the deadline. If we have very low number of returns the results given by the pre-estimations can be biased. This motivated that the response rate should be a built-in factor of our procedure. For computing the response rate it is necessary to know the over-coverage of the survey frame. There is a so called non-response code in the GÉSA system assigned to each data supplier which shows the reason of not sending data. The value of this code should be taken into account. Analysing past data could give a hint for determining a suitable threshold for response rates which may vary with the surveys. On Figure 1 the suggested date for starting the procedure is indicated with the vertical blue line.

27. The following goal is building up the editing phase to be as automated as possible. ADEL databases contain edited data. In order to model this step we can randomly modify these data simulating an erroneous data set for which existing automatic editing procedures (see e.g. in MEMOBUST, 2014) can be tested, at least for the key variables. This is one of our future tasks.

28. For monitoring the estimates we have investigated the Annual statistical survey of production. This is a full-scope survey and estimation was given for the total by summing up the data arrived till each particular day without any editing and imputation. Data arrivals may very much depend on the reminding procedure used and on other circumstances (e.g., receiving a package of administrative data later). On Figure 2 the sum of variable "Total production in natural unit" is presented as a function of days from the deadline for sending the questionnaires till the 50th day.

29. We expected that the estimation will converge after a while, i.e., after a sharp rise the line will slowly have a moderate increase. One can immediately see that there are actually two breaks on the graph. If we had stopped data collection at the first break we have had lost a great part of our data which have had caused a reasonable loss in the (estimation of the) sum. This figure indicates that it is highly necessary to develop an imputation policy for each survey.

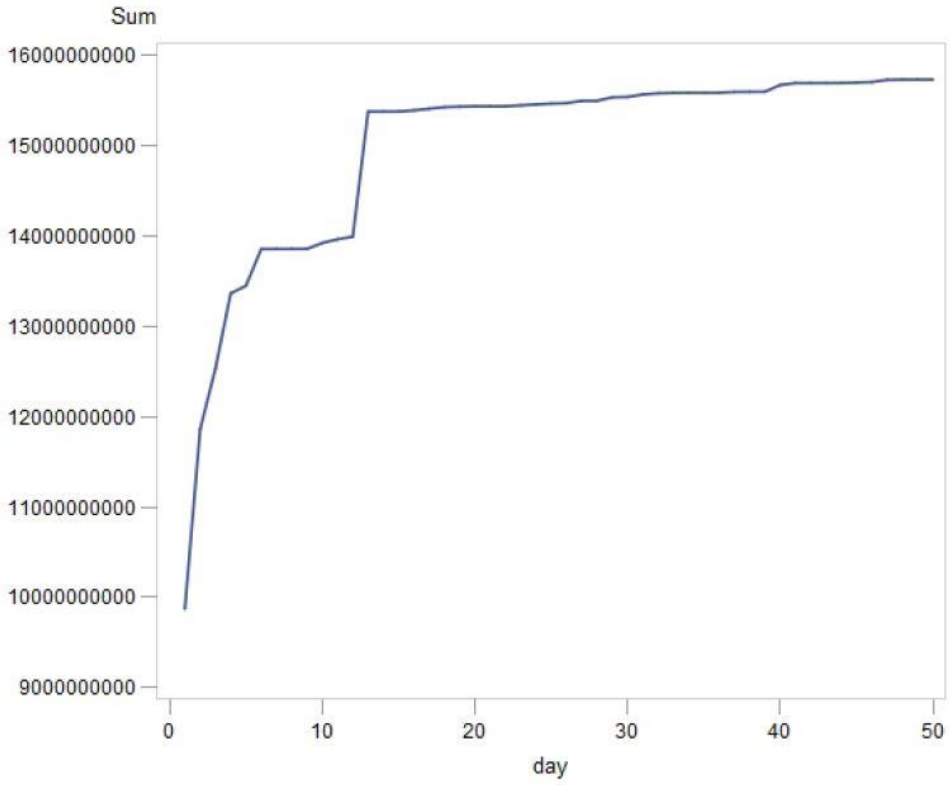

Figure 2
Daily sum of the "total product in natural unit" in survey OSAP 1039.

30. Imputation is also necessary in order to have good-quality microdata-sets for further research. Existing imputation methods have historical and methodological backgrounds. To come up with something new, the methodologist should perform regression imputation, nearest-neighbour imputation or other methods (see e.g. MEMOBUST, 2014 for a good collection of them) in the light of previous data, test their relevance and consult with the survey-owner statisticians. This is a time-consuming task which is also planned for the next years.

31. We have started to analyse the data of the Questionnaire on the Current Environmental Protection Expenditures and on the Environmental Protection Investments in 2014. First we checked whether data from the previous year can work as predictors but we have found no obvious connections. Next the differences between the respondents' and non-respondents' characteristics were investigated. According to their similarities or dissimilarities one can choose the appropriate imputation method. We could not yet find obvious solution as probably not enough information on the data providers has been used. The lessons learned from this analysis were that an IT tool which allows the investigation of the data from many aspects is worth using as well as gather as many information from the subject matter statistician as possible.

32. For checking accuracy in terms of CV daily distance, we have computed the standard deviation and the CV for the pre-estimation of one key variable for the monthly industry data. We observed the expected convergence of the CV. Remember, we would stop the follow-up when the CV does not change significantly from day *d* to day *d+1*. However, to give a limit to its change, i.e., to determine significance level of the change remained unsolved. This method needs further elaboration. Note here that a more precise estimation on accuracy could be given if we could compare the pre-estimations with target values based on a census.

V. Conclusion

33. For building up a highly automated monitoring system we need to take the following into account:

(a) For each survey process an in-depth study has to be performed which means getting to know the process, the data and their relations, and the desired output and the quality thresholds.
(b) Analysing the past is possible based on the data and metadata stored in the database of the ADEL system.
(c) Editing should be as much automated as possible, and the rest of the editing has to concentrate on the key variables – selective editing.
(d) An imputation policy for each survey process has to be worked out.
(e) Appropriate accuracy measure should be chosen with suitable thresholds for each survey.

34. Based on these aspects a long run project is scheduled for the next three years which involves the development of appropriate tools and the study of the existing literature (e.g., Godbout et al., 2011) more thoroughly.

35. Some of the concerns of our colleagues were the following: Suppose we stop data collection when we have reached our predefined limit/threshold. What will happen with our response rates? (Response rate is one of the quality indicators based on which NSIs are judged in their performances.) And what happens if this stop changes the behaviour of the data suppliers in the future? Note that using this method the response rate may decrease in the beginning since there is always a trade-off between timeliness and response rates. But this can also reveal that if the system works well we may change surveys from full scope to sample decreasing the response burden as well. This may influence the behaviour of data suppliers in a positive way. However, there should be other methods for keeping their interests in providing data – but this is beyond the scope of this study.

References

Francisco Fabuel, Pilar Martín-Gúzman, José Vila (2013). *Final report on the project "Analysis of Non-Response on the ESS System*, Ref. nr: ESTAT/B1/ARES-2012-1393082

Serge Godbout, Yanick Beaucage, Claude Turmelle (2011). *Achieving Quality and Efficiency Using a Top-Down Approach in the Canadian Integrated Business Statistics Program*, working paper for the SDE Work Session, Ljubljana, Slovenia

Erzsébet Kómár (2015). *Integrated Data Entry and Validation System in HCSO*, working paper for the SDE Work Session, Budapest, Hungary

MEMOBUST – Handbook on Methodology of Modern Business Statistics (2014), www.cros-portal.eu/memobust

Etienne Saint-Pierre (2013). *Rolling Estimates and the Dynamic Assessment of Quality: a Generic Approach and a Source of Efficiencies*, paper prepared for the European Establishment Statistics Workshop, Nuremberg, Germany

R Core Team (2015). *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

H. Wickham (2009). *ggplot2: elegant graphics for data analysis*, Springer New York, 2009.