

UN/ECE Work Session on Statistical Data Editing
(Ottawa, Canada, 16-18 May 2005)

Topic (v): Quality indicators and quality reporting

THE ROLE OF METADATA
IN THE EVALUATION OF THE DATA EDITING PROCESS

Contributed paper

Submitted by Office for National Statistics, U.K. ¹

I. Introduction

1. The Office for National Statistics (ONS) has embarked on an ambitious modernisation programme out of which it aims to deliver a standard technical infrastructure, methodologies and statistical tools.

2. The aim of this programme is to apply common recognised standards and practices in a highly efficient way. It will enable the ONS to:

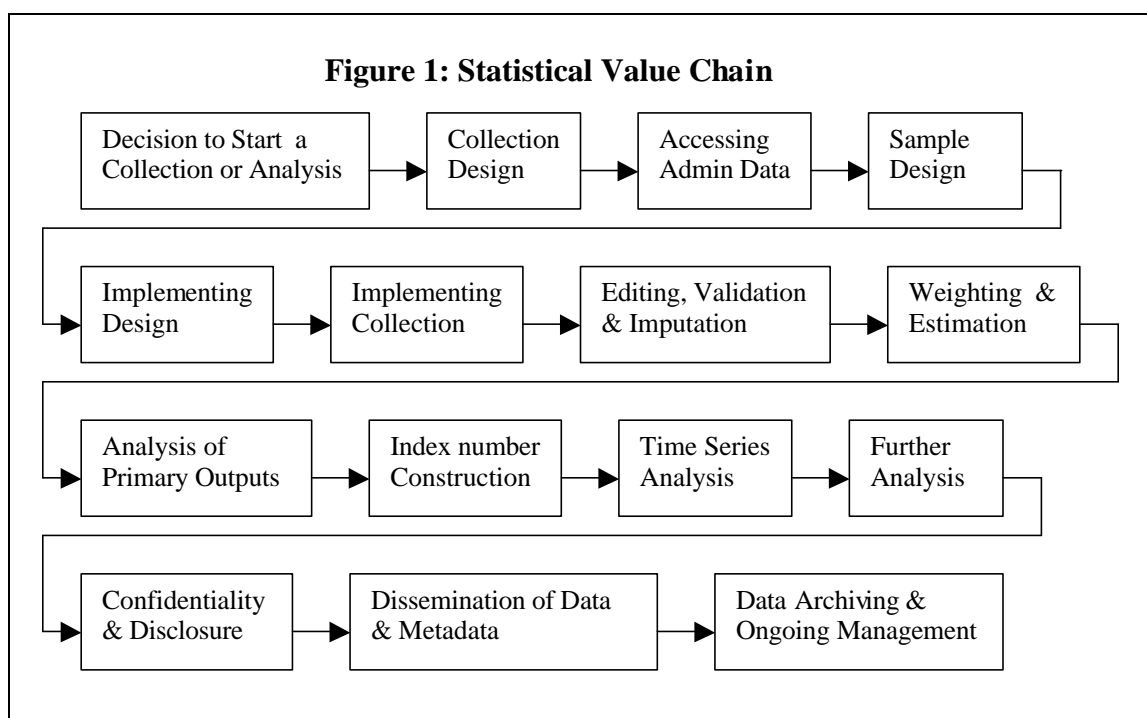
- be expert in using a smaller range of standard software;
- use simpler, more automated processes that provide greater efficiency;
- use an integrated information management system, that will expand the opportunities to assemble information from a wider range of sources;
- use the web as the main medium for our day to day business.

3. One of the key elements of this programme is a series of statistical infrastructure projects to provide corporate statistical tools and methods for use across the ONS. These projects are based on the statistical value chain, (SVC), shown in Figure 1, which describes statistical components from survey design to data dissemination.

4. The project for the editing and imputation component of the SVC has the objective of developing and implementing cost-effective, standard editing and imputation tools for all ONS data sources, incorporating methodological best practice, and operating within the new Information Management environment.

5. This paper discusses the place of the data editing process within the new statistical infrastructure, and its relationships and interfaces with the other statistical processes. It considers particularly the role of metadata in these interfaces, and its contribution to the management and evaluation of the processes themselves, and to the measurement of the quality of the statistical outputs.

¹ Prepared by Pam Tate



II. Managing the editing and imputation process within the Statistical Value Chain

6. As one would expect, data editing and imputation is, in the SVC, most closely linked with data collection on the one hand, and with weighting and estimation on the other. In most circumstances, the editing and imputation processes and tools would be applied after data collection, and before weighting and estimation.

7. The traditional model of statistical data processing is to apply the various individual processes in sequence to a dataset deriving from a specific data source. However, increasingly we may expect to see multiple modes of data collection deployed within a single data source, some modes being able to apply editing at the point of collection, and collection instruments tailored to subgroups or even individual respondents.

8. Thus, although some processes will by their nature continue to be applied to whole datasets, some may be applied to different subgroups or individual cases at different times, or in different ways, or not at all. To avoid too much complication of the language, the discussion that follows refers generally to datasets in the traditional way, but the reader is asked to bear in mind that more complex sequences of operations may often be applied in practice, and that these may demand more complex structures of data and metadata.

9. In the new statistical infrastructure, the methods employed should incorporate best practice, and should be applied in a standard fashion. One implication of this is that there should be no need for any human intervention in deciding which process should be applied to a dataset when, or in what order. This therefore needs to be determined by metadata accompanying the dataset, and interpretable by the data management systems and the statistical tools.

10. In order for the transition from data collection to editing and imputation to operate correctly, the editing and imputation tool must be able to: recognise a dataset which is due to be subjected to editing and/or imputation; recognise which editing methods should be applied to it, and with what parameters.

11. Furthermore, in order for the transition to weighting and estimation to function correctly, the editing and imputation tool must be able to indicate that the dataset is next due to be subjected to weighting and estimation, and with which methods and parameters.

12. Some of this information will derive from the results of the editing and imputation process. For example, imputed values may need to be treated differently from reported ones in the application of the estimation method, so they will need to be flagged.

13. Some of the information however will derive from processes applied at an earlier stage, and will have to be carried along with the dataset in some manner. This may be directly, or more probably by means of a dataset identifier pointing at a separate repository of information on the dataset as a whole, and on the processes it is to undergo.

14. We therefore need two kinds of process management metadata for each dataset: those relating to its progress through the processes in the SVC; and those relating to the options and parameters which need to be applied to it within each individual SVC tool.

15. These metadata depend primarily on what outputs are to be produced from the dataset, and what quality attributes the outputs need to have, as is discussed in more detail below.

III. Evaluation of the quality of the editing and imputation process, other elements of the survey process, and the survey outputs

16. There is a close relationship between the editing process and the quality of the outputs, in several different ways. Editing changes to the data affect the accuracy of the outputs, as does the extent to which imputation is used. The time taken for editing affects the timeliness of the outputs. The nature of the edit checks affects the comparability and coherence of the outputs, as does the imputation methodology used.

17. Some of these processing measures therefore contribute, either directly or as proxies, to the quality indicators for the outputs.

18. We also need measures of the quality of the editing and imputation process itself. Some of these measures can suggest possible ways of improving the performance of the process. For example, information on the number of cases failing each edit check, and on the number of these for which changes in the data resulted, indicates whether the checks are working efficiently in detecting errors.

19. Yet other editing process measures may be able to suggest ways of improving other elements in the survey process. For example, information that a particular variable is frequently changed as a result of failing edit checks may indicate that the question on which it is based should be assessed for quality of concept and wording.

20. Finally, management information on the operation of the editing process can contribute to the management of the survey process as a whole. In particular, up to date information on the progress of data through the editing process, and other individual processes, can enable resources to be switched between the different processes and datasets in the most productive and efficient way, thus improving the quality of the overall survey process.

IV. The role of metadata in the evaluation of quality

21. These measures of the quality of the editing and other processes, and of the survey outputs, are dependent on the creation and use of various kinds of metadata.

22. Working through the SVC from the beginning, the evaluation of the Collection Design stage needs information on how effectively the data collection process has functioned in the past. Some of this comes from the past results of the editing process in identifying errors in the data. Metadata are thus needed on what edit checks were applied, what proportion of records failed each check, and what changes were made to the data in response to edit failure.

23. In Implementing Collection, metadata need to be gathered on the mode of collection, and whether computer-assisted methods were used, since these factors may affect the processes applied later. Where computer-assisted methods are used, information should be gathered on the performance of the editing element of the process, to be used for evaluating its efficiency and effectiveness and identifying improvements for the future.

24. The metadata needed for evaluation and future improvement of the Editing, Validation and Imputation process itself include what edit checks were applied, what proportion of records failed each check, and for what proportion of these failures the data were subsequently changed.

25. For Weighting and Estimation, metadata are needed to identify records that may need special treatment in this process. This includes whether data have been imputed; and also whether data have been identified as implausible by statistical edit checks and then confirmed – these may be outliers.

26. Also, there may be circumstances in which data are considered to be unsuitable for use in imputation, for example if they have been left unedited through a selective editing procedure. This needs to be indicated through metadata.

27. In Analysis of Primary Outputs, metadata are needed to support assessment and evaluation of the quality of the outputs. Where data have been identified as implausible by statistical edit checks and then confirmed, the metadata should include the reasons for the implausibility of the data.

28. For each key output, the quality indicators need to include the proportions of records which had data changed during editing, the proportion which had imputed data, the difference made to the output by editing, and the proportion of the value of the output which derived from imputed data.

V. Implications for data and metadata structures

29. There is thus a wide range of metadata needed for evaluation of quality, relating to data at different levels of aggregation, from an individual variable in an individual record to a complete dataset.

30. Before discussing the implications of this for data and metadata structures, we need to consider the ways in which data and metadata are to be stored and managed in the new ONS infrastructure.

31. The Central ONS Repository for Data, or CORD, was mentioned earlier as an element of the modernisation programme. It is proposed that it will hold all forms of ONS data, cross-sectional and longitudinal, from surveys and administrative sources of all kinds, at all levels of aggregation.

32. It is also proposed that within CORD there be a Central ONS Repository for Metadata, or CORM. It is envisaged that this will contain metadata about entities, such as method, survey, dataset, data item, classification, question. These metadata will be updated at defined trigger points of the SVC.

33. The CORM will then be a convenient vehicle for ensuring that metadata are automatically made available to users of data aggregates and other outputs, which are disseminated through the web tools. Additionally, in parallel, the microdata and associated unit level metadata will still be available internally for analysis.

34. When considering the structure of the metadata needed for the interfaces of the editing process, we therefore need to distinguish between unit or record level metadata, sometimes called micrometadata, and summarised or aggregated metadata.

35. Unit level metadata includes for example the failure of a record to pass an editing check, a change in the value of a data item, the reason for the change, and so on. Summarised metadata includes the number of records that have failed a particular edit check, the number that have been changed, the proportion of an estimate that derives from data changed through editing, and so on.

36. Micrometadata are created as each individual record passes through the survey process. They describe the characteristics of the data in that record as identified by the survey process, and the interaction of the data with the survey process.

37. The summary level metadata are derived from the micrometadata, but describe for example the characteristics of the dataset as a whole, or of an estimate derived from that dataset, or an edit check applied to that dataset. They relate to a variety of higher level entities, in contrast with the micrometadata which relate to an individual data item.

38. The unit level metadata need to accompany or be linked to the unit level data in the data repository. They are needed for monitoring the operation of the process itself, (providing *inter alia* management information), and for monitoring the performance and quality of the process.

39. The summary level metadata are more appropriately held in the metadata repository, together with other information about the dataset as a whole.

40. This implies that the design of the data repository needs to take account of the needs for unit level metadata; that the design of the metadata repository needs to take account of the needs for summary level metadata derived from the unit level metadata; and that there need to be (automatic) processes for deriving the summary level metadata from the unit level metadata.

41. In addition, there are some items of unit level metadata that need to be accessible across various data sources, for example information gathered from a particular respondent about the reasons for an implausible but confirmed piece of data - this may well explain other implausible data gathered from that respondent in another survey, and may also be of use to compilers and users of more aggregated data.

42. This category of metadata, which relates more to the unit in general than to the specific data item, sometimes has implications for the survey frame, and sometimes just for other operations on that unit. In either case, the most convenient location for it is likely to be in or linked to the frame or register in which the general data about the unit are held.

VI. Managing the process interfaces through metadata

43. The interfaces between the editing process, and the adjacent processes of data collection and weighting and estimation, are managed by two types of metadata. One consists of the micrometadata that accompany the data, and include information about the history of the data, at unit level, as it passes through the various processes in the SVC. An example of the contents of this, at the points of entering and leaving the editing process, is sketched in Figure 2.

44. The second type of metadata needed for managing the interfaces between processes is information relating to the whole dataset on the processes which are to be applied to it, and the options and parameter settings within those processes that are applicable to this dataset. These need to be held in a repository of process control settings, as part of the operational management of the survey process.

45. For the editing and imputation process, the options and parameters need to define such things as the edit checks to be applied, the actions to be taken in case of failure to pass an edit, the automatic correction procedures to be applied, the imputation methods to be used, and in what circumstances, and so on.

46. The choice of these options needs to be based on a thorough understanding of the subject of the survey, analysis of past and related data, and up to date knowledge of best practice in editing and imputation methodology. It also needs to be determined in co-ordination with the other elements of the survey process, and informed by assessment of the interactions between them.

Figure 2: Example of micrometadata input to and output from the editing process

INPUT

Unit level

Unit identifier

Collection mode

Response category

Capture date/time

Explanatory commentary on implausible data obtained from computer-assisted collection

Process history (which tools applied at what date/time)

Variable level

Variable identifier

Whether data obtained, data missing or variable not applicable to unit (may be separate metadata variable or special data values)

OUTPUT (in addition to input metadata)

Unit level

Updated process history

Explanatory commentary on implausible data from scrutiny or respondent follow-up

Whether each edit check failed

Selective editing category

Whether any data changed during editing

Whether any data imputed

Whether unit excluded from providing imputed values

Variable level

Whether data failed each relevant edit check

Selective/priority editing score

Whether data confirmed

Whether data changed

By what/whom data changed or confirmed (e.g. automatic correction, editor scrutiny, respondent follow-up)

New data (in addition to original data)

Date/time of data change

Whether new data imputed

Imputation method

VII. Conclusions

47. Data editing is linked to other elements of the survey process in many and various ways. Some relate to the management of the editing process itself within the Statistical Value Chain; and some to the ways in which information about the editing process and its effects contributes to the operation of that and other processes, and of the survey process as a whole.

48. These relationships can contribute greatly to evaluating and hence improving the quality of the survey outputs, and the efficiency and quality of the survey process. But the achievement of these improvements depends very much on creating the right metadata, and being able to use them effectively in conjunction with the survey data to evaluate the outputs and processes.

49. This involves three elements. Firstly, the necessary metadata must be specified, at both unit and aggregate levels, to support the evaluation of quality. Beyond that, it is essential to specify data and metadata structures that can facilitate the use of the metadata in managing and evaluating the survey process. And lastly, these structures must also support the analysis of the metadata together with the survey data, in order to determine how to improve the quality of both processes and outputs in future.
