

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Data Editing
(Rome, Italy, 2-4 June 1999)

Topic (i): Measuring the impact of editing in various phases of statistical survey processing

**MEASURING AND ANALYSING THE DATA EDITING ACTIVITY IN ISTAT
INFORMATION SYSTEM FOR SURVEY DOCUMENTATION**

Submitted by ISTAT, Italy¹

Invited paper

I. Introduction

1. Documenting and improving data quality is becoming one of the main goals of National Statistical Institutes. In fact, the users of statistical data are showing an increasing awareness of the importance of data quality. Moreover, international organisations are recommending evaluating data quality and are asking for information on data quality and on the transparency of the production process. For example, Eurostat is requiring its Member States to provide extensive and harmonised information about data quality, especially for business surveys. In particular, the Member States are asked to periodically send standard quality reports.

2. Therefore, it becomes important to provide the survey managers with useful tools to support the data quality control activity. In this respect, different solutions can be devised. In particular, two different typologies of instruments can be defined. The first one includes procedures, techniques or instruments planned for improving or optimising the execution of a single operation or phase of the production process. Examples of this typology are a generalised editing and imputation procedure or computer assisted data collection (CATI or CAPI). For example, ISTAT has developed a generalised edit and imputation procedure, named SCIA, which is being used for correcting qualitative variables. In the second one, we include information systems devoted to support the production process. For example, ISTAT has developed a unique information system containing all the information related to the interviewers. Such a database is currently used and updated by all the surveys, which adopt a face-to-face data collection technique. For business surveys, ISTAT has set up a database called ASIA, containing information related to all the Italian businesses. ASIA is used as a frame for all sample business surveys carried out by ISTAT. ASIA has been set up by using several administrative databases and it is updated yearly.

3. In this paper we describe in detail the **information system for survey documentation (SIDI)**, which has been designed and developed to support the quality control activity (Brancato, D'Angiolini, Signore, 1998). SIDI aims to monitor the survey production process, to document the activity of data production and quality control and also to disseminate suitable information on data quality to the final users. To this purpose, SIDI manages qualitative information (metadata related to the survey production process) and quantitative information (quality measures) in an integrated way. In our approach, we have focused on the production process since it is widely recognised that the quality of final data benefits of improvements in the process.

¹ Prepared by M. Fortini, M. Scanu and M. Signore.

4. In the paper, we first analyse how data editing relates to other phases of the production process. In particular, great attention is given to analysing the editing activity with regard both to the performance of the editing procedures and to the relationships existing between data editing and the errors coming from other phases of the production process (section II). After having described the general purpose and features of SIDI (section III), we illustrate how qualitative and quantitative information related to editing is managed into the system (section IV). We also emphasise the importance of combining quality indicators derived from different phases of the production process (e.g. editing and data collection) in order to analyse a particular aspect of data quality (e.g. interviewer effect), (section V).

II. Data editing procedures in the survey production process

5. The editing operations are complex to describe since they are often not performed in a simple sequence. In fact, generally they are included into or preceded by other phases of the production process (e.g. manual editing before the data capture operation). In general, we can organise the process in a sequence of sub-operations where each sub-operation may consist of a particular editing approach or imputation method.

6. Moreover the imputation, which does not mean “to replace a mistake with the true value”, is only one of the three purposes of the editing and imputation process. According to Granquist and Kovar (1997) the other two important objectives are: “gathering intelligence related to significant difference in the data for analytical purposes” and “providing feedback that can lead to improvements in data collection and processing”.

7. It is commonly accepted that the editing and imputation operations are performed mainly with the purpose of cleaning up the most important non-sampling errors from the data. Nevertheless, it should be considered that any chosen editing method, when applied, would affect the data with new errors. This sort of errors, even if they are not reported by the editing rules, might affect the survey estimates possibly even more than the unprocessed data could do. As an example, overediting is one of the worst effects that an editing and imputation process can cause.

8. The editing and imputation process’ “side effects” are a matter of concern that has led to a less invasive class of methods, called selective editing. These methods process only the most influential statistical units, with respect to the estimates, while all the others remain unchanged. The editing and imputation activity, apart from the error reduction, also allows the investigation of the non-sampling errors sources in the survey production process. This can be done analysing the type and the number of corrections done with respect to the units and studying the connection of the corrections with the characteristics of both the units themselves and the operations performed over the units during the collection phase.

9. It is worth mentioning that the rationale behind the adopted editing rules and the methods used to make imputations should always be considered in the analysis of the error sources. In fact the chosen editing rules and imputation procedures can affect the relationship between the observed data and the characteristics to be studied. For this reason appropriate statistical techniques and other relevant information should be considered in order to remove any possible confounding effect from the analysis of error causes.

10. In particular data coming from survey phases other than the editing and imputation ones should be studied in depth. For example, the incidence of frame errors and non-respondent units, the information about the follow-up activity and number of proxy respondents and the information regarding the incoherence showed by respondents in filling in the questionnaires recorded by a CATI, CAPI data collection mode, represent useful sources for improving data regarding the editing and imputation phase. We therefore emphasise the importance to store all the computed quality indicators about both data and process quality in order to compare the indicators with each other and their modification over time. Moreover, it might be interesting to compare the level assumed by the quality indicators between similar

surveys. It could be interesting to study, for example, the differences in quality indicators for different surveys on the same population, which use different editing and imputation techniques.

11. In the next section we show how ISTAT is dealing with the issue of storing and managing both information about survey quality measures and meta-information about the survey production process. In particular, we describe the information system for survey documentation (SIDI) used by ISTAT and its relationships with other databases used by single or small groups of surveys to store and manage data and metadata about topics such as the phenomena under study, both survey instruments and procedures, and quality of data.

III. Documenting and monitoring the survey production process

12. As we have seen, editing is strictly related to the other phases of the production process. It is, therefore, important on the one hand to **analyse the quality of editing** by taking into account the problems and the errors occurring **in the various phases of the production process**. On the other hand, the **impact of editing on survey data** should be carefully analysed. In fact, what are needed are sets of quality indicators, which enable the survey manager to monitor each phase of the production process. In addition, it might be useful to jointly analyse the information on data quality coming from different phases of the survey process.

13. For this purpose, we have defined a system of quality indicators that allows the survey manager to monitor the execution of the survey operations and to evaluate the quality of final data. Specifically, a **set of standard quality indicators** has been defined **for each phase of the production process**, namely frame design and updating, data collection, data entry, data editing and imputation. Two additional sets of standard indicators describing the entire survey production have been defined, i.e. indicators on timeliness and punctuality and indicators on the survey costs.

14. However, further information is needed to properly interpret the quality indicators. Such information is represented by metadata concerning the survey characteristics. In particular, both the methodological features and the operational and organisational aspects of a survey can affect the quality of the final data. For example, non-response rates are highly affected by the data collection technique. The rate of imputed variables and/or records depends on the efficacy and completeness of the set of edits of the applied procedure. Therefore, to know whether the editing procedure has been checked for completeness of the set of edits might be very useful to better understand the values taken by the quality indicator “rate of imputed variables”.

15. As already mentioned, the evaluation of the quality of survey data is becoming a very important issue. Moreover, survey managers need to be supported in such a demanding task. ISTAT has decided to develop an information system, named SIDI, especially dedicated to documenting and analysing the quality of the survey production process. In fact, SIDI allows the survey managers to document in a standard way the relevant aspects and the methodological characteristics of their surveys. SIDI also helps the survey managers to evaluate and analyse the system of standard quality indicators related to the production process. Thus SIDI is a complete tool for improving and standardising quality control.

16. It is worth noting that SIDI has been designed as a **centralised** information system that can be used for every ISTAT survey. This choice is motivated by the need to improve the quality of statistical information issued by ISTAT and to standardise both the survey procedures and quality control actions. Another advantage of a centralised system is the possibility of comparing the quality of different surveys.

17. One of the main purposes of the system should be to allow comparisons among different but homogeneous surveys. For instance, in order to analyse the impact of the editing phase, it can be useful to compare the standard quality indicators of all the surveys which use a particular generalised editing and imputation procedure to correct the data. The metadata managed into the system allow the users to select groups of surveys on the basis of several selection criteria. For example, surveys which observe a particular statistical unit (e.g. business surveys) or surveys which perform a certain operation (e.g.

automated editing) and/or perform a particular control action (e.g. preventive test of the automated editing procedure). Once a particular group of surveys has been chosen, the system shows the quality indicators for each survey of the group. The selection criteria depend on the set of standard quality indicators of interest. For example, in order to analyse the quality indicators of the data collection phase, the data collection technique is one of the most relevant selection criteria. On the other hand, the data collection technique does not affect the performance of the edit and imputation procedure. Examples of selection criteria for the editing phase are the type of observed variables (i.e. mainly quantitative or qualitative), the type of editing (i.e. manual, automated or interactive) and the type of preventive test if any (i.e. by simulating errors in a clean file and evaluating the performance of the editing procedure).

18. In order to allow comparisons among different surveys, both qualitative and quantitative information have been highly standardised. In order to equip the system with a rich set of enquiry functionalities it is necessary to have standard descriptions of the survey characteristics and the applied methodologies, as we have seen for the selection of homogeneous surveys. Moreover, it was necessary to define sets of standard quality indicators so as to enable the users to compare the quality of different surveys. In particular, the quality indicators have been standardised with respect to the evaluation function, which has been defined in a unique way, for each quality indicator managed into the system. However, there is the possibility to specify the standard quality indicators with respect to different classification variables. The classification variables can be common to all surveys such as the geographical area (whose minimum common degree of detail is the classification in five areas i.e. north-west, north-east, centre, south and islands) or can be differently specified for different groups. For example, the quality indicators for the editing phase can be differently specified with regard to the observed statistical unit. In particular, for business survey the quality indicators can be specified for the economical activity of the business, while for household surveys they can be specified for the characteristic of the municipality (i.e. metropolitan, in a metropolitan area, rural).

19. Another important goal is to allow the analyses of the quality of a specific survey over time. For instance, a survey manager is interested in monitoring the impact of the editing phase by analysing the proper set of standard quality indicators for different survey replications. For this purpose SIDI manages time series of each set of standard quality indicators together with the associated metadata. In this way it is possible to relate changes in the time series of a quality indicator (or a set of indicators) to improvements in the production process. For examples, it is possible to analyse the trend in the imputation rates knowing whether the editing procedure has been modified.

20. In the next section, we illustrate the quality indicators, which have been defined to evaluate the editing activity. Given that SIDI is a centralised information system, it manages quality indicators that can be compared among surveys. Therefore they might not be sufficient for an in-depth investigation of a specific phase. In fact, our idea is that SIDI should refer to **local** information systems where detailed information on a given phase or survey operation is managed and stored. For example the quality indicators detailed at the interviewer level (e.g. non-response rate per interviewer) are currently stored in the already mentioned interviewer database. In particular, some generalised edit and imputation procedures (such as SCIA) automatically evaluate rich sets of indicators, which must be analysed each time the procedure is applied. The same indicators are also very useful in the phase of tuning of the procedure.

IV. Metadata and quality indicators related to data editing

21. As we have seen, for the purpose of monitoring and documenting the quality of the survey production processes, both qualitative and quantitative information is required. With regard to metadata managed in SIDI, we have conceptually distinguished the operations (e.g. face-to-face interview) from the quality control actions that are regarded as a particular class of operations performed to improve the quality of survey data (e.g. interviewers' training). Even if the editing activity is aimed to correct non sampling errors, in the information system it has been considered as an operation since it is a current practice for all ISTAT surveys. This does not mean that the data editing is standardised, on the contrary,

there is a great variability among different surveys regarding editing (e.g. manual, automated or interactive) and the type of editing procedure (ad hoc or generalised) and so on. The metadata managed in SIDI allow the survey manager to describe different aspects of editing such as the current editing operations performed for each survey replication; the planning and the redesign of the editing. The preliminary test of the procedure and all its features have been considered as quality control actions related to the editing activity.

22. The standard quality indicators managed in SIDI are based on the comparison between raw and clean data regardless of the editing procedure, which has been applied. By raw data we mean the outcome file of data entry, while by clean data we mean the file of final data after the editing activity. In general, such an activity is differently performed among ISTAT surveys. For example, some surveys may apply more than one automated procedures or may include a direct contact to the edited units (e.g. large businesses) to correct the errors. In particular, the set of standard indicators allows the survey managers to analyse the impact of the editing activity both on the statistical units and on the observed variables. By comparing raw and clean data, it is possible to calculate some basic measures for each ISTAT survey. These basic measures are used as bases to calculate the set of standard indicators.

Examples of basic measures are:

U: number of statistical units

V: number of variables that can be imputed

I: total number of imputations

Note that in the V measure we exclude those variables, if any, which are not subjected to imputation in the editing and imputation procedure. For example, some identification variables such as the municipality code or the interviewer code may not be included in the edits because they have been manually revised and corrected before data entry. Examples of standard quality indicators obtained from the above-mentioned quantities are defined as follows:

the rate of total imputation $I_T = I / (U \cdot V)$

the mean number of imputations per statistical unit $I_U = I / U$

the mean number of imputations per variable $I_V = I / V$

23. Moreover, the set of standard quality indicators includes the main quartiles of the following distributions:

$f(x)$ that is the distribution of the statistical units per number x of imputations

$g(x)$ that is the distribution of the variables per number x of imputations.

24. As already mentioned, the quality indicators can be further detailed with respect to three classification variables. The SIDI system is also provided with a rich set of graphical and tabular representations associated with the quality indicators.

25. In the presentation, we will show examples of both metadata and quality indicators related to the editing activity. In particular, we will show the quality indicators calculated for different replications of the Labour Force Survey.

V. An example of data quality analysis by using indicators computed from different phases of the survey production process

26. As we have already seen, the analysis of the editing and imputation activity can provide the survey manager with a lot of information about non-sampling errors. The analysis can also be greatly enhanced if this information is combined with that deriving from other relevant sources, as such provided by the data collection activity.

27. In this section, we describe a study (Righi, 1996) of the effects of the field activity on the data quality of the survey “Italian citizen degree of satisfaction on the public services”, with particular attention to the **interviewer effect**. For this purpose, we considered both the information from the data collection operations and that from the edit and imputation procedure.
28. To collect information, the interviewers contacted the sampled households and each component was asked to answer a questionnaire on his/her satisfaction about the services supplied by public offices as for example postal offices or health institutions. Respondents had to fill in the questionnaires by themselves in the presence of the interviewer so that they could ask for help both for understanding the meaning of the questions and for answering the questions in the right order. The interviewer recorded in a proper field of the questionnaire whether he/she helped the respondent to fill in the questionnaire. It was then possible to exclude from the analysis all the cases in which the interviewers didn't help the respondents to accomplish their task.
29. To conduct the field activity the interviewers were divided into groups and each group was assigned the responsibility for covering a municipality. For this reason it was supposed that, in addition to the individual effect of a single interviewer on data quality, also the municipality could influence the level of non-sampling errors.
30. To measure the level of non-sampling errors affecting the data, 15 indicators were chosen regarding both the result of contacting and gaining the family's cooperation to the survey and the success of respondents in filling in the questionnaires properly. With respect to data collection the following cases were considered: non-contacts because of frame errors, non-contact of the family because not at home, family refusals, single person not interviewed because not at home or because of a refusal.
31. The quality of data completed in the questionnaire by the respondent was measured by monitoring the editing and imputation phase. In particular, for some relevant questions we enumerated the item non-responses and the errors in following the correct flow of the questionnaire. Quality measures were taken on every household and the results were summed up for the households surveyed by the same interviewer. At the end of this step, for each one of the interviewers, the 15 quality indicators were obtained.
32. To take into account the effect of interviewer characteristics on data quality, the following variables were considered: age of the interviewer, gender, education, type of employment and previous experience as an interviewer. In addition, the effect of the municipalities on the interviewer's performance was considered by recording the two following variables: geographical area and number of inhabitants in the municipality where the interviewers operated.
33. The fifteen indicators were combined into a smaller set of factors by using the principal components technique. The analysis returned six factors explaining about 60% of the whole variability measured for the fifteen original indicators. To interpret the six factors with respect to the survey production process a *varimax* procedure was applied to the data. While only a factor seems to be relevant with respect to the frame, three other factors were found related to the interviewer ability in contacting and convincing households to participate in the survey. Finally the two last factors were interpreted as related to the completeness and correctness in filling in the questionnaire.
34. Furthermore, association of the relevant factors with the interviewer characteristics was studied. To study the association, each factor was categorised according to the quartiles of the distribution over the interviewers. Then a X^2 was computed with respect to every combination of the factors and the interviewer characteristics. The findings of this second stage of analysis showed that:

- the factor related to the frame errors is influenced only by the municipalities characteristics;

- some of the interviewer characteristics are relevant for the factors related to the ability in contacting the households, in particular gender (males perform better than females) and previous experience seem to be related to the ability in contacting households;
- younger interviewers (18-35 years old) perform better than older ones (36 years and more) in helping the respondents during the self-administration of the questionnaire. This fact could be related to a greater motivation in doing their job, as suggested also by a worse performance of the permanent employees than of the contract employees. Past experience as an interviewer was also very relevant;
- Geographical area and number of inhabitants of the municipality resulted in a strong relation with all of the factors;
- Education was associated to the ability in contacting the households but interviewers with higher level of education obtain either high or low response rates. This might also be related to personal motivation.

35. The results described in this section can be considered neither general nor definitive. In fact the factors identified by the analysis don't account for much of the whole variability showed by the data. Moreover, some of the association measured between the factors and the interviewer characteristics are not strong enough and sometimes not well supported by other research (e.g. male interviewers perform better than females in gaining household participation to the survey). Nevertheless, this study shows how a connection between different pieces of evidence about quality of data is established and how a summary of results could be attempted. In this context the information provided by the editing and imputation source is essential not only to evaluate the process itself but also to analyse other important steps such as the collection phase.

References

G. Brancato, G. D'Angiolini, M. Signore (1998), "Building up the quality profile of ISTAT surveys", Proceedings of the Conference Statistics for Economic and Social Development, Aguascalientes, September 1-4.

L. Granquist and J. G. Kovar (1997), "Editing on survey data: how much is enough?", Survey measurement and process Quality, Lyberg et al. Eds., J. Wiley, New York.

P. Righi (1996), "Metodi di analisi fattoriale applicati agli indicatori di qualità dell'indagine Servizi resi dalle pubbliche amministrazioni e grado di soddisfazione dei cittadini", ISTAT internal document, (in Italian).