Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys

ISTAT		CBS		SFSO	
	Orietta Luzi (Coordinator) Marco Di Zio Ugo Guarnera Antonia Manzari		Ton De Waal Jeroen Pannekoek Jeffrey Hoogland Caren Tempelman		Beat Hulliger Daniel Kilchmann

Foreword

Overview

Eurostat and the European National Statistical Institutes (NSIs) need to provide detailed high-quality data on all relevant aspects (economic, social, demographic, environmental) of modern societies. Moreover, the quality declaration of the European Statistical System (ESS) states that continuously improving a programme of harmonized European statistics is one of the major objectives of the ESS. Given the differences among the existing national statistical systems, the development of new approaches aiming at standardizing the statistical production processes for improving the quality of statistics and for a successful harmonization is essential. In particular, one of the two main purposes of the "European Statistics Code of Practice", adopted by the Statistical Programme Committee on 24 February 2005, is: "To promote the application of best international statistical principles, methods and practices by all producers of European Statistics in order to enhance their quality".

In this framework, the European Community (represented by the Commission of the European Communities) is supporting the development of a set of handbooks on current best methods in order to facilitate the implementation of principle 8 of the Code of Practice - "Appropriate statistical procedures, implemented from data collection to data validation, must underpin quality statistics". This work should capitalize on what has already been done in Member Countries, consolidate or complement existing information and present it in a harmonized way.

This action is directly related to the recommendations of the Leadership Expert Group (LEG) on Quality (LEG on Quality, 2001). The LEG on Quality stated that Recommended Practices (RPs) can be considered a highly effective harmonization method which can, at the same time, provide to NSIs the necessary flexibility for using the "best" methods in their respective national context. A Recommended Practice Manual (RPM) is a handbook that describes a collection of proven good methods for performing different statistical operations and their respective attributes. The purpose is to help survey managers to choose among the RPs those that are most suitable for use in their surveys in order to ensure quality. The reason for providing a set of good practices is that it is difficult to define the best methods or standards for statistical methodology at European level.

In the area "Data validation methods and procedures during the statistical data production process", editing and imputation represents a relevant data processing stage mainly due to its potential strong impact on final data as well as on costs and timeliness: a better quality of processes and produced statistical information can be obtained more efficiently by adopting the most appropriate practices for data editing and imputation. Therefore, developing Recommended Practices for editing and imputation is considered an important task.

The RPM presented in this handbook focuses on cross-sectional business surveys. Developing a RPM for editing and imputation in this specific statistical area is considered a priority for a number of reasons:

- since important economic decisions are taken on the basis of information collected by NSIs, if data with errors or with missing values are not appropriately dealt with, then the quality of those decisions can be jeopardized. For this reason, the best practices and methodologies for the detection of errors and the imputation of missing and erroneous information for the different types of data in the Community ought to be used in the ESS;

- particularly in business surveys, there is a large heterogeneity in practices and methods used for editing and imputation not only at NSIs level, but also at European level. For this reason, comparisons of data from different statistical institutions in the ESS may be difficult;

- particularly in business surveys, editing and imputation is recognized as one of the most time and resources-consuming survey processes. Standardizing the editing and imputation process is expected to generate less expensive statistics and to shorten the time from data collection to publication of results while maintaining the quality of results;

- surveys conducted by NSIs, and in particular business surveys, are generally characterized by a lack of documentation concerning editing and imputation strategies, methods and practices. In this area, there is a strong need for performing systematic studies and for developing common practical and methodological frameworks.

The present RPM was developed in the framework of the European Project "Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys (EDIMBUS)", carried out in 2006/07 with partial financial support of Eurostat. The project was coordinated by the Italian National Statistical Institute (ISTAT) and involved the participation, as partners, of the Centraal Bureau voor de Statistiek Netherlands (CBS), and the Swiss Federal Statistical Office (SFSO).

Aim of the Recommended Practices Manual EDIMBUS

Developing and disseminating the RPM in the area of business statistics is expected to contribute to the development of a common practical and methodological framework for editing and imputation applications.

The RPM represents a guide for survey managers for the design, the implementation, the testing and the documentation of editing and imputation activities in cross-sectional business surveys at both European and NSIs levels.

At the European level, the handbook will facilitate the harmonization of statistical data production, thus contributing to the improvement of the comparability of the data produced in the ESS.

At single NSIs level, the RPM aims to reduce the heterogeneity of methods and practices for editing and imputation and to improve and standardize the procedures adopted for data editing and imputation. At individual survey level, the handbook can represent a support for survey managers in the adoption of the appropriate methodology in each specific context. Furthermore, the handbook is meant to disseminate a systematic approach for testing new editing and imputation processes or editing and imputation processes already in use, but that have not been appropriately tested before. Researchers and survey managers can find in the handbook a guide for planning, testing, evaluating and documenting an editing and imputation process. The RPM can also be used as a checklist for documenting and evaluating the effects on the final results of the performed editing and imputation activities. In general, the RPM is expected to contribute to the improvement of the quality and cost-effectiveness of the editing and imputation process, and the timeliness of the publication of the results.

It has to be underlined that the reader of the RPM is assumed to have some basic theoretical and practical knowledge on editing and imputation in cross-sectional business surveys. In effect, the good practices and recommendations given in the manual will not all apply in the same way to all survey contexts: their relevance and actual applicability in each specific context are to be carefully evaluated by subject matter experts and designers of the editing and imputation process, taking into account the survey objectives, organization and constraints.

Though the focus of this RPM is on cross-sectional business surveys, many of its elements can be extended to longitudinal business surveys and possibly to some cross-sectional and longitudinal household surveys.

Organization of the volume

The structure of the handbook follows a general prototype process of editing and imputation in cross-sectional business surveys.

Chapter 1 starts with the description of the general framework, with the embedding of the E&I process in the whole survey production process. Quality management aspects are also discussed. Chapter 2 is dedicated to the discussion of how to design and control an editing and imputation process in the area of business statistics. It illustrates the basic concepts, the overall criteria and the key elements to be taken into account when performing these activities. Furthermore, it also describes methods and approaches for testing and tuning editing and imputation methods: these activities allow survey managers to evaluate and monitor their own editing and imputation processes for their continuous improvement. Chapter 3 contains the description of the main types of errors which usually corrupt economic survey data. For each error type, the error detection methods which can be used to identify it are described. The methods which can be used to treat the errors once they have been identified are described in Chapter 4. Chapters 3 and 4 follow a similar structure. First of all, an overview of each method is provided. Then, for each method, the advantages and limitations relating to its use, as well as the most appropriate application context(s) are pointed out. Finally, recommendations are provided on the use of the method. Recommendations are provided in the form of a checklist on how to efficiently use the specific method. In Chapter 5 the problems relating to data analysis and estimation in presence of imputation, missing data and outliers are discussed. The documentation of editing and imputation processes is discussed in Chapter 6. Documentation has a crucial importance in the area of editing and imputation, since it guarantees that users and specialists are informed about the characteristics of the editing and imputation process, its costs and its impact on the data. Furthermore, documentation may support survey managers and researchers in understanding the error characteristics and the error sources for future improvements of the overall survey process. Chapter 7 contains a synthesis and general recommendations for the editing and imputation process.

Four Annexes deal with additional aspects or illustrate more in detail some particular issues.

Appendix A provides notational details.

Appendix B contains a description of the main results obtained by the state-of-the-art survey on current methods and practices carried out during the EDIMBUS project.

In Appendix C methodological details on the methods for error detection and error treatment illustrated in Chapters 3 and 4 are provided.

Appendix D contains a mathematical description of the indicators for testing, tuning and documenting editing and imputation processes.

Finally, a Glossary of terms is provided in Appendix E.

Acknowledgments

This manual is the result of the integration of the theoretical knowledge and practical experiences of the project's members. It also incorporates knowledge collected during the project activities. In particular, a state-of-the-art survey has been performed in European and overseas countries, in order to gather information about current practices and strategies for editing and imputation in different national contexts. Furthermore, a validation process on the preliminary version of the manual performed by external referees from different European and non-European NSIs has allowed the integration of the manual contents with different views, and has contributed to improve it.

The Authors wish to thank the following NSIs for contributing to the state-of-the art survey: Statistics Finland, Statistics Sweden, INSEE France, Office for National Statistics (ONS) UK, Statistics Austria, Statistical Service of Cyprus (CYSTAT), Czech Statistical Office, INS Belgium, Statistics Estonia, Federal Statistical Office Germany, Hungarian Central Statistical Office, Statistics Lithuania, Statistics Norway, Central Statistical Office of Poland, Instituto Nacional de Estatística (INE) Portugal, Statistical Office of the Slovak Republic, Statistical Office of the Republic of Slovenia (SORS), U.S. Census Bureau, Statistics Canada.

The Authors also wish to thank all the referees from partner and non-partner NSIs for revising the draft of the manuscript and for their helpful comments: for Statistics Sweden, Leopold Granquist and Gunnar Ardvison; for ONS, Alaa Al-Hamad; for INE Spain, Pedro Revilla, José Manuel Feito; for DESTATIS, Elmar Wein; for Statistics New Zealand, Vera Costa; for CBS Netherlands, Jelke Bethlehem and Petra Brocke-Snel; for SFSO Switzerland, Philippe Eichenberger, Bertrand Loison and Rick Trap; for ISTAT, Maria Gloria Narilli, Ersilia Di Pietro, Salvatore Filiberti, and Roberto Sanzo.

The Authors are especially in debt with John Kovar, Eric Rancourt, Jean-François Beaumont, Hélène Bérard, and Chris Mohl from Statistics Canada for carefully reading and insightful commenting the draft handbook, and for their highly constructive and detailed suggestions.

The Authors also like to acknowledge gratefully financial support from Eurostat.

Furthermore, the Authors thank their own Institutions (ISTAT, CBS, SFSO) for the generous support they gave to the project.

The views expressed in this handbook are those of the Authors (ISTAT, CBS, SFSO) and do not necessarily reflect the policy of Eurostat.

Contents

Chapter 1

General Framework of Editing and Imputation in Cross Sectional Business Surveys

1.1 Introduction

Editing and Imputation (henceforth E&I) is a set of activities aiming at detecting erroneous and missing data and treating these data accordingly. E&I is an important sub-process in the statistical data production process since it is time/resource consuming and it has a non negligible effect on survey data and their statistical properties. E&I activities can be performed at different stages in the survey process: data capture, data entry and post-data collection processing. In this handbook, we exclude the so-called *input editing* activities, i.e. activities for error detection and treatment performed before data are captured or keyed, and editing activities performed at the data capturing stage (e.g. through Computer Assisted Interviews). Hence, in this handbook, we focus on the E&I activities done at the post data capture stage, when the data are available in electronic format. We include also activities performed once all the data are available and preliminary estimates can be obtained to verify that not important errors are left in the data. The E&I process ends with the release of the final data ready for analysis and publication. However, E&I has to be viewed as an integral part of the overall survey process, and links between E&I activities performed at the different stages and other survey elements are to be properly managed in order to better exploit the potential of E&I (see Section 1.3).
In this RPM, editing stands for the identification of missing, invalid, inconsistent or anomalous entries. During the editing process, values are flagged as acceptable or suspicious . Imputation is the process used to resolve problems of missing, invalid or inconsistent responses identified during editing. During imputation, missing or inconsistent data items are substituted by estimated values. Imputation is part of a more general activity called data treatment which involves interactive treatment as well.
Traditionally, E&I has been viewed mainly as a tool for cleaning up the data. During the last years, however, more importance has been attributed to the role of E&I as a tool for gathering knowledge and information on errors and error sources, for improving the overall survey process, and for providing feedback to the other survey phases, in order to reduce the amount of errors arising at the earlier stages of the data collection and processing. To this aim, increasing importance has been put on the monitoring and the documentation of E&I activities and results, e.g. by means of indicators and meta data.
Nowadays, the international community has recognized that E&I should have the following main objectives (Granquist, 1995):

1. identify error sources in order to provide information for future improvements of the survey process;

2. provide information about the quality of the incoming/outgoing data;

3. identify and treat the most significant errors;

4. when needed, provide complete and consistent (coherent) individual data.

The notion of error is central to E&I. In this manual, an error occurs when the observed value is different from the actual "true" value. Errors are characterized by different aspects: for E&I, the most important ones are their source, their nature, their impact on data and how they appear in data. The source of error (e.g. definitions, questionnaire, interviewers, and so on) is particularly significant for future repetition/improvement of the survey in order to avoid the problems encountered during the current survey. The nature of error can be roughly divided in systematic and random. The distinction is mainly due to the different effects they have on data. In general, systematic errors have a deterministic effect once they occur, while random errors are characterized by a probability distribution with its own variance and expectation. This distinction is also important since the techniques used to deal with systematic and random errors are significantly different. Error impact is not an absolute concept, but depends on the target parameter and/or on the aggregation level: in fact, an error may be influential with respect to a certain estimate, but it may have a negligible impact on some other. This characteristic should be considered when balancing the trade-off between data accuracy and costs of E&I. Finally, the way the errors appear in data (outlier, missing, and so on) will influence the choice of methods to be used to detect them in the E&I process. All the above mentioned aspects overlap, so in order to deal with the variety of problems concerning error detection and treatment, all of them are to be considered. For this reason, in this handbook we have not chosen any specific perspective, but all of them are considered and described in Chapters 3 and 4. This choice is also motivated by the fact that each of these aspects generally requires the application of a specific methodology for error detections and/or error treatment. The decision on how to treat some type of errors is of course left to the reader that is assisted in his decision by the discussion given in Chapter 2.

Different types of methods are currently used, both for the detection and for the treatment of the different types of errors. Methods are characterized by different aspects as well. They differ by the kind of information that is used, how it is used, the type of error that can be detected and treated and the amount of resources needed. In order to improve the effectiveness of E&I processes, a combination of methods is usually applied. From an operational point of view, the most important distinction is between automatic and interactive methods (methods which contain human intervention) for error detection and treatment. Under optimal conditions, interactive methods may be considered enough accurate to deal with errors, but they can be very costly and, in some sub-optimal situations, they may introduce variability and bias and even new errors. Interactive treatment includes the possibility of calling back the respondent and clarifying errors. This means, in fact, that new original information is gathered. However, follow-up may originate additional respondent burden and does not guarantee better data quality. Automatic methods are cheap to apply, easy to reproduce and document, but may be less accurate at local level compared to optimal interactive treatment; however this could not necessarily represent a problem when the goal of the survey is about estimation at a global level.

E&I methods can also be classified into micro and macro approaches. In micro approaches, the detection and treatment of errors is done at the record or questionnaire level, in isolation from other responses in the current survey. Typically, micro E&I procedures involve applying edit rules. An **edit rule**, or edit, is a restriction to the values of one or more data items that identifies missing, invalid or inconsistent values, or that points to data records that are potentially in error. Edits are often classified as **fatal edits** (or *hard edits*) and **query edits** (or *soft edits*) depending on whether they identify errors with certainty or not.

Unlike micro approaches, macro approaches use all (or a large part) of the data for the current cycle to identify and treat possible errors. The distinction between micro and macro is particularly important not only for theoretical reasons but also because of operational aspects. Due to timeliness, it is often preferred to start the E&I activities as soon as the data are available in electronic format in order to not delay the survey process. For this reason, micro E&I techniques, which do not depend on the flow of responses, are generally used at the early stages of the E&I procedures. On the other hand, macro approaches should be applied as soon as most part of data are available.

Graphical representations like, among others, boxplots, scatter plots, histograms are also widely used for data E&I. The use of these tools in this context is generally referred to as **graphical editing**. Graphical editing can be used in different phases of the E&I process. It is used as an exploratory statistical tool, for instance to understand relationships between variables. It can also be applied to build edit rules, for instance as a tool for determining the acceptance bounds for a variable. It can be used as a direct tool for detecting errors, for instance in outlier detection. It can be adopted to monitor the effects of an E&I procedure, for instance by supporting the analysis of the change in the estimates due to error detection/treatment activities. To emphasize the importance of graphical editing, the peculiarities of this kind of methods are discussed within each single error detection and treatment subsection of this manual. In general, an experienced statistician, who does a careful analysis of the data by means of exploratory techniques in combination with other E&I methods, usually will obtain better results than those which can be obtained by any fully automatic method.

From this overview, it can be easily seen that E&I is a complex process, which must be carefully designed, tested, monitored, revised and documented. Therefore, besides E&I approaches and methods that can be used to detect and treat errors, this handbook provides elements and recommendations on all these aspects in the specific area of cross sectional business surveys. In the following Section 1.2, the main elements characterizing business surveys and E&I strategies in this area are described.

1.2 Editing and Imputation in Business Surveys

Business surveys are generally periodic surveys collecting mostly quantitative economic data. Business surveys can be classified in two broad classes: those producing either Short-Term Statistics and those focusing on Structural Statistics.

Short-term statistics collect information (variables) necessary to provide a uniform basis for the analysis of the short-term evolution of supply and demand, production factors and prices (Council Regulation (EC) No 1165/98 of 19 May 1998). All variables are to be produced more frequently than annually.

Structural statistics aim at analyzing, for example, the structure and evolution of the activities of businesses, the regional, national, Community and international development of businesses and markets, the business conduct of the small and medium-sized enterprises, the specific characteristics of enterprises related to particular groupings of activities (Council Regulation (EC) No 58/97 of 20 December 1997).

In this section, we provide a short overview of the problems relating to error detection and treatment for business survey data.

There are some main elements that characterize business survey data (Granquist, 1995). Firstly, responses to items of interest often present highly skewed distributions, in other words, a small number of units substantially contribute to the total estimate. Furthermore, information on the surveyed businesses is often available from a previous survey or can be drawn from administrative sources. Moreover, strong correlations usually exist between current and historical data of the same item and among many items. Finally, respondents generally obtain their data directly from their accounting systems. Particularly in business surveys, the respondents must have an in-depth understanding of the concepts and definitions underlying the questions. Differences between survey definitions and definitions used in accounting systems, for example the use of financial versus calendar year, may produce response errors which are difficult to identify. Ensuring coherence between definitions and concepts is essential. The survey vehicle (survey design, questionnaire, mode of data collection, respondent burden) has to be adapted to this objective in order to prevent response errors that the E&I process cannot solve.

Business surveys incur substantial E&I costs (producer costs, respondent costs, losses in timeliness), especially due to intensive follow-up and interactive editing. E&I activities commonly performed in business surveys generally do not produce a substantial increase in data quality that could justify the resources and time dedicated to them. Business survey data are often over-edited, though certain kinds of errors may be undetected by editing. Often much of the impact of E&I on the data depends on only a small percentage of the total changes. Furthermore, most interactive actions leave the original values unchanged. For a discussion on these topics see Granquist and Kovar (1997) and Granquist (1995).

1.3 Quality management in Editing and Imputation and connections to other parts of the survey process

In this section, the problem of error detection and error treatment activities performed at the E&I stage are described in a wider *quality management* perspective. The interactions between this stage and other survey aspects are underlined, the impact of E&I on quality dimensions, and the need for managing information flows among the E&I activities performed at the different stages of the survey process, are shortly discussed.

As stated in Section 1.1, the goal of E&I is not limited to the identification and treatment of errors on collected and captured data: an important role of E&I is related to its capability to provide information on the data quality and on the survey process, by means of appropriate indicators and documentation. E&I can contribute to gather intelligence on definitions, to improve the survey vehicle, to evaluate the quality of the data, to identify non-sampling error sources, to highlight problem areas, to optimize resources allocation. Hence, much importance has to be given to learning from E&I results, providing the basis for future improvements of the survey process (Granquist and Kovar, 1997; Statistics Canada, 2003; Granquist et al., 2006). In this *quality management* perspective, it is straightforward to see that the aim of E&I is also to prevent errors.

Furthermore, optimizing E&I on one hand contributes to the reduction of its costs and to the improvement of its effectiveness, on the other hand has an impact on all quality dimensions. E&I affects timeliness, as excessive traditional E&I leads to delays in publication. It has an impact on accuracy, as, on one hand, E&I contributes to eliminate some response and processing errors and to reduce nonresponse bias, but on the other hand it may introduce new errors and uncertainty (see Chapter 5). Accuracy is positively affected in case of E&I performed during the data collection, while the respondent is still available (e.g. in computer-aided interviews or web-questionnaires). Furthermore, follow-up activities focusing on identifying problems faced by respondents in the data help to sharpen definitions and questions, change the questionnaire design and the data collection instrument, so that respondents burden and nonresponse rates are reduced in future surveys. Coherence is positively affected by E&I, particularly when administrative data are used to check/treat data, and in general, in all cases when auxiliary data and all data sources available at the Statistical Agency are used for E&I. Finally, only a clear documentation leads to the necessary transparency of the survey process, forming the basis of an efficient quality enhancement of the different survey stages. Under this view on E&I, it is important to design the allocation and co-ordination of E&I activities at all survey stages, particularly in order to reduce the risk that data are edited and imputed more than once. The design of an overall E&I strategy is embedded in the design of the whole survey process. Therefore, links among the different phases where E&I activities are performed, may have to be managed at the survey design stage. Furthermore, re-engineering the E&I activities performed at the post-data collection stage implies that the whole survey vehicle should be re-thought, including the questionnaire, the mode of data collection and data entry, the use of generalized software, and the allocation of resources at the different survey stages. Concluding, the quality management perspective of E&I leads to the following evidences: high effort has to be put into error prevention, contributing in improving the quality of incoming data and in getting accurate responses from all respondents; more efficient E&I strategies are to be designed to save resources and redirect them into activities with a higher pay-off in terms of data quality, in particular data analysis and response errors (Granquist and Kovar, 1997). It is clear that this integrated view also implies a multidisciplinarity approach on E&I, since survey designers, questionnaire designers, E&I specialists and subject matter experts have to co-operate.

Chapter 2

Designing and Tuning Editing and Imputation

2.1 Introduction

In this chapter, the elements for the efficient design of an E&I process, as well as the problem of how an E&I process may be tested, tuned and monitored are discussed. The design of E&I strategies is dealt with in Section 2.2. In Section 2.2.1, the general criteria to consider when structuring an E&I strategy are illustrated. These criteria mainly rely on the characteristics and relative importance of units and errors. Section 2.2.2 discusses some basic notions for the description of the data and the process flow generated by an E&I strategy. Section 2.2.3 provides some examples about how some survey characteristics may influence the design of an E&I process.

Once the process is designed, it needs to be thoroughly tested and modified accordingly. Approaches for testing a process and appropriate quality indicators are discussed in Section 2.3. Finally, a tested process can be implemented in the actual survey process. The E&I process then needs to be continuously monitored and tuned, as also discussed in Section 2.3.

2.2 Designing the strategy

2.2.1 General criteria for the design of an Editing and Imputation process

The design of the E&I process should be part of the design of the whole survey process. The survey manager usually elaborates an E&I strategy, consulting specialists of each phase of the E&I process and involving the managers of related survey sub-processes. The design does not require actual data because it only establishes the general structure of the process. In this section, the criteria to be considered for the conceptual design of an E&I process are illustrated.

Given the wide range of business surveys, an effective E&I process can be obtained by differently combining approaches and "best practices". However, some general rules should be followed when building an E&I strategy:

1. firstly, identify and eliminate errors that are evident and easy to treat with sufficient reliability;
2. secondly, select and treat with great care influential errors, carefully inspect influential observations; automatically treat the remaining non influential errors;
3. check the final output to see if there are influential errors undetected in the previous phases or introduced by the procedure itself.

The previous order is based on an evaluation of the trade off between accuracy and resources spent for data treatment. The first two criteria can be met by means of either micro or macro E&I approaches

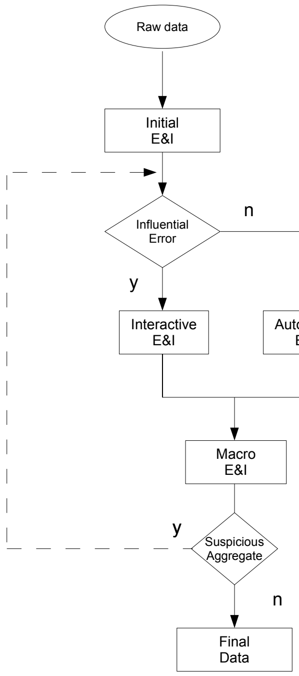

Figure 2.1: General flow of an E&I prototype process

(see Section 1.1): conceptually, in micro E&I, the units flow through the procedures separately whereas in macro E&I, sub-samples or the entire sample data flow through the procedures. On the other hand, the third point is generally accomplished by using macro E&I approaches. However, it has to be remarked that in many situations, a large amount of data is available near the beginning of the E&I process, so that macro approaches can be used early.
Based on this premise, a general E&I process is the one shown in Figure 2.1.

In the first phase (*Initial E&I*), the aim is to treat those errors that can be dealt with high reliability. Typically, systematic errors (see Section 3.3) are dealt with in this phase. These errors are generally automatically imputed by using deductive and/or rule-based imputation (see Sections 4.2.2 and 4.2.3).

The second phase is focused on identifying (potential) influential errors, i.e. values that are suspected to be in error and have a large impact on target survey estimates. The selection of influential errors is usually done through selective editing (see Section 3.4). In selective editing, a **score** is calculated for each record expressing the relevance of the potential error(s) in the record. This score is used to prioritize observations for accurate analyses, like expert interactive treatment, and/or follow-up (see Section 4.3). Limiting the interactive treatment to these most relevant units/errors is directly related to a cost-limiting criteria.

Since it frequently happens that a high number of observations are affected by less important errors, the time spent for the treatment becomes an important element to be considered. For this reason, non influential errors are usually automatically edited (see Sections 3.2 and 3.6) and treated (see Section 4.2) using appropriate software that process data on a record by record basis. Although automatic procedures are used for relatively unimportant errors, choosing the most suitable error detection and/or

imputation methods is still important. If non appropriate methods are used, especially for large amount of random errors and/or missing values, additional bias may be introduced, thus compromising the accuracy of final estimates.
In the E&I process flow, the interactively and the automatically treated observations will enter the last phase as a unique data set.

The last phase (*macro E&I*) typically involves the use of macro approaches, that take advantage of all the available collected information. In macro E&I, the suspicious values are usually identified by using outlier detection techniques (see Section 3.5). Outliers do not need to be errors but, if they are, they can have a large impact on survey estimates and it is therefore worthwhile to allocate resources to verifying them. At the end of the process, the validity of preliminary estimates is checked (*Suspicious Aggregates*) through macroediting techniques (see Section 3.4) using historical data or other sources of information. This process is also referred to as *data confrontation*. In the case of suspicious estimates, the underlying (most influential) units can again be identified and checked interactively, usually in a selective manner. Outliers and influential units identified at this stage may correspond to errors not properly identified or treated at the earlier stages of the E&I process, or introduced by some E&I activities.
It is straightforward to see that, given the variety of business surveys, the flow depicted in Figure 2.1 does not exhaustively describe all the possible strategies that can be developed in practice. This scheme is meant to give a broad picture of the process as a whole, and to serve as an introduction to the sections of this manual where the different methods and approaches are treated in detail (Chapters 3 and 4). For a discussion of the elements which may determine departures from this scheme, see Section 2.2.3.

2.2.2 Notions of the Editing and Imputation process flow

In this section, the basic notions underlying the design of the flow of an E&I process are discussed. Based on the concepts introduced in Section 2.2.1, in this section the E&I process is defined as a process with a parameterization. The implementation of the E&I process therefore requires a set of parameters.
The following main four principles of the E&I process lead to the notions used in this manual.

1. The data quality at the beginning and at the end of the E&I process must be assessed.

2. The E&I process has to be designed and executed in a way that allows for control of the process.

3. The data quality at the end of the process should satisfy the needs of the users.

4. The process should be as simple, cheap and fast as possible.

Principle 1 describes the performance of the process and helps to identify the error sources for future improvements, whereas principle 2 leads to the need of designing, testing, monitoring, documenting and archiving the process. Principle 3 states that the process should be effective. In order to know whether the process is effective the users' needs should be known. In practice, often, highest quality is targeted simply because the needs of the users are unknown. However, if such needs are indeed known, principle 3 may lead to final data containing inconsistencies or even missing data if the users' needs are met without treating them. Hence, quality of individual data may not be completely optimized in order to achieve perfect or even high quality. Principle 4 ensures that the process is also efficient focusing on the costs and the comprehensibility of the process itself. Therefore, good methods may not be implemented because of their complexity or their consumption of resources for development, implementation, testing, processing and monitoring.
In light of the above mentioned principles, the first step of the E&I process is the design of the process. Once the process has been designed, it has to be implemented with a parameterization, tested, tuned

and become productive. The design and production must be considered under their respective quality management aspects. The quality management of the design includes the indicators mentioned in Section 2.3.1, recommended practices for the design and an established procedure for obtaining approval of the design. As regards the quality management of the production, it must include a process design, an established procedure for approval of changes in the design, reallocation of resources and repetition of phases or procedures. Furthermore, the whole process should be monitored using the indicators described in Section 2.3.2. Both quality management actions include documentation, see Chapter 6.

For an efficient management the E&I process design a general process view must be adopted. Therefore, the E&I process will be described phase-by-phase, as shown in Figure 2.1. In the present context, a **phase** is defined by a compact set of **procedures** which are defined as implemented methods and techniques. A phase should be reproducible and repeatable using a phase specific set of parameters; the beginning or end of a phase must correspond to a milestone of the E&I process. A phase model allows to archive the data at the end of each phase and thus evaluate the impact of the phases. The three phases of the E&I process resulting from Figure 2.1 are: 1. *Initial E&I*; 2. *Identification of influential errors, interactive and automatic treatment*; 3. *Macro E&I*.
A more detailed scheme of the E&I process in Figure 2.1 is obtained when the four basic procedures of each phase are made explicit.

1. **Detection** of erroneous data (Section 3).

2. **Decision** about the treatment (Section 2.3.2).

3. **Treatment** (Section 4).

4. **Control** of the treatment (Section 2.3.2).

These four basic procedures are very natural but are often performed without explicit differentiation. In fact, usually when erroneous data are detected and located, a decision about their treatment will be taken. Such decision may be based on indicators, e.g. the impact of the erroneous data on final estimates or the number of units failing the edit rule. It may lead to doing nothing, transfer the treatment to another phase or treat the error with a given method (see Chapter 4). For example, in phase 3 a **drill-down** from macro to individual (micro) E&I is often performed to locate the errors and decide about their treatment. This happens, for example, when an error (an outlier) as described in Section 3.5 is detected. Often the most influential outliers are interactively treated (as in selective editing) while the rest of the outliers are treated automatically. By doing so, the decision about treatment may involve using similar procedures already performed during an earlier phase. Such a drill-down may be formalized with a **loop-back** in the process flow that leads from the decision taken in phase 3 that the data are to be individually edited, to the detection or treatment of phase 2. In other words, a phase or a procedure of a preceding phase is performed again with adapted parameters.
The detection procedures of phase 3 may reveal residual errors. These errors may be identified only in this phase, because some detection methods could only be applied in phase 3, e.g. comparison with the last survey estimates or with external data or multivariate analysis. These errors may also have been introduced during the E&I process. In such cases, the decision will often involve treating them, adapting the procedures of preceding phases, like adapted interactive treatment, adapted automatic imputation, etc.
The aim of managing the E&I process is, of course, to prevent such loops because they are costly and time consuming and generate unwarranted respondent burden. However, they should be considered in the design of the E&I process.
The treatment, third basic procedure of a phase, consists in applying implemented methods for treatment with their respective set of parameters.

The quality of the phase is ensured by the fourth basic procedure, that is control of the phase output. Such control includes, of course, analyzing the phase outcome in relation to the purpose of the phase, possibly taking into account preliminary estimates, external data and macro E&I rules. Furthermore, the control procedure contains a decision element regarding the acceptance of the output of the phase. This decision is based on afore-mentioned analysis. Again indicators related to the aims of the phase may be used to standardize and facilitate the decision. If the outcome of the phase is accepted, the succeeding phase will be started or the final data will be released after the control of phase 3. If the outcome of the phase is not satisfactory, it might be necessary to **repeat** the phase, i.e. a repetition of procedures with adapted parameters inside the phase, or a loop-back to a preceding phase.
At least for the first applications of the process, it is necessary to foresee loop-backs and repetitions of phases to be prepared and and to prevent wrong assumptions about the release date of final estimates.

2.2.3 Key Elements for the design of an Editing and Imputation process for business surveys

A typical E&I process for economic data usually involves all the phases of the process flow depicted in Figure 2.1. However, the actual flow can change depending on the characteristics of each specific survey, the users' needs, the available resources and the available auxiliary information. These elements are summarized in the following points.

1. **Survey characteristics**: type of survey (Short-Term Statistics, Structural Statistics, Economic Censuses), survey size (number of units, number of variables).

2. **Survey objectives**: target parameters (totals, means, ratios, covariances, etc.), level of detail of released data (micro data, aggregates).

3. **Available auxiliary information**: historical micro/aggregate data, administrative data, data from other surveys.

4. **Available resources**: human, time, financial resources, available equipment (software, infrastructures, etc.).

5. **Applied methods** and their integration: the method applied in one phase may have an effect on the choice of the methods to be applied in other phases of the E&I process.

When designing E&I as a process flow, all these elements will determine its general architecture as well as its phases. In the following some examples of how the previous elements may influence the E&I flow of Figure 2.1 are provided.

1. Survey characteristics

   (a) In Short-Term surveys, timeliness is a strong requirement since a short time period from data capturing to data publication is available, and often, preliminary estimates are to be provided. In such cases, it can be convenient to concentrate the resources mainly on the identification and the accurate (interactive) treatment of influential errors.

   (b) The amount of data to be checked (units and variables) is another element conditioning the design of an E&I procedure. Large scale surveys, like economic censuses and structural statistics, require the allocation of a high amount of resources for the identification, the analysis and treatment of both influential and non influential errors. Since data from structural statistics are generally used to build up information systems to be used as benchmark in some related short-term surveys, these data are to be treated in depth to achieve completeness and consistency in all variables. Although more resources are generally available for E&I activities, these surveys make an intensive use of automatic

E&I. n particular, since in these surveys many variables are usually collected in highly complex questionnaires, several automatic E&I procedures are often used in an integrated way, dealing each with subsets of variables.

Furthermore, in the case of hierarchical data (e.g. local unit of enterprises), the resulting E&I process will become more complex since it will be necessary to check the consistency of information at both local units and enterprise level. This may imply the integration of different E&I processes, in a predetermined hierarchy, in order to guarantee the overall data consistency.

2. Survey objectives

   (a) Due to the survey objective, the data flow may deviate from the one in Figure 2.1. As discussed above, when complete and consistent micro data are to be released, the automatic E&I procedures are used to deal with non influential inconsistencies and item nonresponses. In some situations (e.g. in some short-term statistics), when no individual data are to be released, it may happen that the automatic E&I procedures are not applied at all, or they imply few basic verifications and treatments.

   (b) Survey objectives have an impact on the E&I flow when preliminary estimates are required at a certain level of aggregation (e.g. regional level), and final estimates are to be provided at a different (e.g. more detailed) level of aggregation (e.g. provincial level). In this case, the E&I process may be divided in two different micro and macro editing phases with their respective parameters where, for instance, selective editing and outlier detection are performed considering the two different levels of aggregation.

3. Auxiliary information

   (a) When appropriate auxiliary information can be used to identify suspicious or unacceptable data and/or to predict acceptable values during imputation, more efficient E&I procedures can be built, and more accurate results can be obtained. This holds for both automatic and interactive methods. The type of available auxiliary information influences the choice of the methods that can be used. For example, if reliable register information is available at individual level, then this information can be used in micro editing approaches. If the register information is relatively good only at aggregate level, it can be used only in the macroediting phase.

   (b) The use of external auxiliary information (like registers or data from other surveys) in the E&I process may imply the need for checking the consistency and/or the completeness of the external data. This could require an additional phase with respect to the flow of Figure 2.1. In certain contexts, this phase may be not only resources and time consuming, but also complex, as it may itself imply an additional E&I process like the one in Figure 2.1.

4. Available resources

   (a) Resources are a crucial element to take into account when designing an E&I process, since they have a strong effect on both the general structure of the process and on the methods used in its phases. For instance, since follow-up of influential data is a costly activity, the amount of re-contacts which is possible to perform has to be planned taking into account the time, the human and budget resources which are available for the E&I process. In less advantageous situations, it can happen that no resources at all are available for re-contacts, hence the clerical review of influential data may be chosen as an alternative. This has an obvious impact on the quality of the data compared to the case where all suspicious influential data are re-contacted. The training of clerical reviewers is important and costly. If they are not well prepared, a consequence may be the so-called "creative editing" and the need of performing an additional phase of E&I on the interactively revised data (see Section 4.3). In some situations, neither re-contacts nor interactive treatment are possible due to limited resources or to the impossibility to re-contact units: in these cases, data are directly treated through automatic E&I procedures.

5. Applied methods

   (a) In the detection of influential values/errors and outliers for different target survey variables, if univariate methods are used, this phase has to be repeated until all variables have been treated. The number of repetitions can be reduced by using multivariate methods involving sub-groups of target variables.

   (b) The use of approaches specific for the detection of random errors generally requires that data are free of systematic errors, this implies that systematic errors must be detected in some previous E&I phase.

Although not exhaustive, all situations listed above are examples of how the key elements 1. to 5. (and combinations of them) may have an impact on the design of the process, determining an actual survey-tailored process flow which may differ to some extent from the scheme of Figure 2.1. Therefore, when designing an E&I process the flow of Figure 2.1 (and the reasoning behind it) should be strongly taken into account, while carefully considering the key elements shortly discussed in this section.

Generalized software

An important element to consider when designing an E&I process, is the availability of generalized tools. Some leading statistical agencies have developed generalized software implementing traditional methods and algorithms (e.g. for the analysis of edits, for outlier detection, for the identification of non influential random errors, for imputation) in order to support survey managers in performing some costly or complex data treatments on their own data. The use of these tools may allow to save costs and time for performing some specific process phases, since the implementation of the needed algorithms is already done, and the required parameters can be usually updated more easily than in ad-hoc procedures. It is worthwhile noting that developing some of the algorithms available in these software packages can be very costly.

2.2.4 Recommendations

1. The E&I process should be designed as a part of the whole survey process. The overall management of the E&I process and the interfaces with the other survey sub-processes should be considered. For each phase, the specific aim, the required quality, the expected inputs and outputs, the starting point, the possible loop-backs, the ending point and the parameters should be described. For each phase the resources and time needed to implement, test, execute and document it should be planned.

2. The E&I process should minimize the changes to the data. In other words, data consistency should be obtained by changing as few observed data as possible.

3. Edit rules should be designed in collaboration with subject matter specialists and should be based on the analysis of previous surveys. Consistency and non redundancy of edits should be verified. Edits should be designed cautiously in order to avoid over-editing.

4. The appropriate flags, the documentation including indicators and archiving should be part of the design.

5. Systematic errors should be detected and treated first (see Section 3.3).

6. Resources should concentrate on influential errors (including nonrespondents). Selective editing, outlier detection and detection of influential observations are means to this purpose.

2.3 Testing, tuning and monitoring the strategy

2.3.1 Testing Editing and Imputation methods

Once an E&I process has been designed for a given survey, it needs to be tested in order to assess its suitability for the observed data and to find the best set of techniques and/or parameters able to optimize its performance. The E&I process contains an integrated set of techniques, such as for example the detection of systematic errors, the detection of influential errors and the imputation of missing data. Consequently, the testing of an E&I process should be split into the testing of the different techniques used. The testing aims at evaluating the performance of the techniques in terms of efficacy (ability to achieve the objectives). Based on the results from testing, some of the design decisions and/or parameters could be revisited to optimize the performance. This section focuses on the evaluation of the accuracy of the E&I techniques. It must be pointed out that the different purposes of error detection, imputation, and interactive treatment require different criteria; therefore, we define separate evaluation criteria for the different activities of an E&I process. However, operational characteristics (resources required, reproducibility, flexibility to changes, ease of use, etc.) are also to be taken into account when evaluating competitive E&I methods.

Framework for evaluation studies

In order to evaluate the accuracy of the different techniques in an E&I process, one needs contaminated data (i.e. containing errors and/or missing items) as well as data which can be considered free of errors according to our best possible process. The latter are usually referred to as the "true" data. The tested techniques are applied to the contaminated data and the results obtained (treated data) are compared with the "true data" in order to compute a set of accuracy indicators used to select the best techniques. Investigations based only on the comparison between the treated data and the original raw data, without an error-free counterpart, can only result in a statement on the impact of the E&I process on the data and publication quantities. This kind of analysis does not assess the accuracy of the techniques used and is generally performed for monitoring and tuning purposes, see Section 2.3.2.

A comparative evaluation study can be done in different ways. Hereafter, two of the most used approaches are described. In the first approach, the results obtained with the current E&I process are considered the "true" dataset and inconsistencies and missing values are artificially introduced (by simulation) in it, using some sort of error or missing data mechanism. These mechanisms should be developed to approximate reality as much as possible, which can be quite difficult especially for generating inconsistencies, see Di Zio et al. (2005b). It should be noted that generating missing/inconsistent data and comparing results is a process that should be repeated a large enough number of times in order to take into account all the different stochastic mechanisms affecting the results.

The second approach to the comparative evaluation study is to apply the different techniques to the original raw data and to compare the results with the data relying on the gold standard E&I process, consisting e.g. in the use of external information sources, call-backs and subject matter specialists knowledge. The E&I procedures can then be assessed in their ability to approximate these gold standard values. The assumption for this approach is that there exists a mechanism, represented by the gold standard E&I process, leading to the "true" dataset.

Testing of error detection methods

The error detection techniques can be assessed in their ability to detect the errors. Evaluation criteria in this case could concern the following issues, see also EUREDIT Project (2004b).

1. The ability of the detection procedure to find the maximum number of errors that are present in the dataset. The number of correct errors found divided by the total number of errors in the dataset can be used for this purpose, see the indicator (D.1) in Appendix D.2.1.

2. The ability of the detection method to find the most influential errors. For example, by calculating the number of influential errors found divided by the total number of influential errors present, see the indicator (D.2) in Appendix D.2.1. As these errors have a substantial impact on the final survey estimates, it is important that they are located through the error detection mechanism.

3. Although the method must be capable of finding all errors present in the data (item 1), this should not be achieved at the cost of flagging items erroneously to be in error. The total number of incorrect errors found divided by the total number of flagged items may reveal an undesired behavior of the error detection method, see indicator (D.3) in Appendix D.2.1.

A more technical description of these indicators can be found in Appendix D.2.1. The method or combination of methods that performs best with respect to the quality requirements of a particular survey should be implemented in the final E&I process.

Testing of imputation methods

The accuracy of the imputation methods is measured by means of the indicators that summarize the differences between the imputed values and the "true" values. The differences can be referred to estimates and/or distributions or can focus on the individual data items. The following criteria are very important, see also EUREDIT Project (2004b).

1. The ability of the imputation procedure to preserve the individual item values. This can, for example, be assessed by calculating the (weighted) average deviation of the imputed data from the true data for continuous variables, see the indicator (D.6) in Appendix D.2.2.

2. The ability of the imputation method to preserve population estimates. The average deviation of the mean and dispersion estimates based on the imputed data and the true data can be calculated for this purpose, e.g. using the indicator (D.12) in Appendix D.2.3.

3. The ability of the imputation procedure to preserve the distribution of the data. This can be assessed by comparing the Kolmogorov-Smirnov distance between the imputed and the true data for continuous variables (e.g. using the indicator D.13 in Appendix D.2.3) and a χ^2-Test for discrete and categorical variables.

For a more technical description of these indicators and for other indicators, see Appendix D.2. The relevance of the different criteria generally depends on the survey objectives. For instance, when individual data have to be released or used in prediction models, the preservation of the individual values is a desirable feature. The imputation method or combination of methods that performs best with respect to the quality requirements of a particular survey should be implemented in the final E&I process.

2.3.2 Application of the process, tuning and monitoring

Once the error detection and imputation methods have been tested and implemented, they need to be monitored continuously during their execution in order to identify problems that may arise as soon as they appear and tuned accordingly. There is no dataset containing the "truth" at hand during production, which means that indicators for monitoring can only be based on the observed and/or treated data.

Diagnostics that may be informative in this case are, for example, the number of failed edit records, the rate of edit failures, the amount of editing done, the number of missing values, the impact of the E&I process on publication totals. See Appendix D.3 for a technical description of these indicators and additional diagnostics.

It is important to realize that, once the E&I process is implemented in the actual survey process, only slight changes should be made to monitoring and tuning in order to avoid structural breaks in the time series. The monitoring of the phases may be included in a general assessment framework for longitudinal evaluation.

2.3.3 Recommendations

1. The E&I process should be thoroughly tested before it is applied. Ideally a test data set in a "clean" and in a raw version should be available for testing. Error detection and treatment should be tested separately and together.

2. The E&I process and its phases need to be continuously monitored. Information on errors and error sources can be used for the continuous improvement of the E&I process as well as for improving the other sub-processes of the survey process.

3. The subset of indicators used during the testing should be selected on the basis of the survey characteristics and the evaluation objectives. Indicators on costs and timeliness should also be taken into account during testing and monitoring (see Appendix D.1).

Chapter 3

Detection of Errors

3.1 Introduction

The error detection phase consists of identifying values that are not acceptable with respect to some pre-defined logical, mathematical or statistical criteria. Since different error types may contaminate the observed data in a sample of units, usually the error detection process consists of a set of integrated error detection methods dealing each with a specific type of error. The detection of the different error types allows assessing the quality of the incoming data and helps to identify error sources for future improvements of the survey process (see Chapter 1). The result of error detection are flags pointing to missing, erroneous or suspicious values.

This Chapter describes different error types with the methods that can be used to detect them. Since, as discussed in Section 1.1, errors can be classified according to different criteria (random or systematic, influential or non influential, etc.), the method applied for a specific error depends on the criterion mainly being taken into account. The structure of this chapter reflects this approach to error and method classification.

Error detection is often based on the use of edit rules. For instance, edits can be used to detect systematic errors (see Section 3.3), or to set up automatic procedures for the random errors' localization. Therefore, a central role in editing is played by the edit specification and analysis. In particular, query edits that do not identify errors with certainty, must be designed with care in order to avoid performing many unnecessary checks. In edit specification, different elements have to be taken into account. First of all, edits should be based on subject matter knowledge, that is on an understanding of the social and economic conditions that are likely to influence respondents and the implications they have for the relationships between data items. Furthermore, the definition of acceptance regions through edits should be supported by specific statistical methods. In particular, analysis of the joint distributions may greatly facilitate the specification of appropriate edits by discovering relationships between variables and useful ratios (Whitridge and Kovar, 1990). Graphical methods can also be useful.

Interactive treatment, which could also be classified as an error detection approach, is not discussed in this Chapter: in effect, in the interactive treatment, values identified as erroneous are often also changed. For this reason, in this handbook, interactive treatment is discussed in Section 4.3 as a method for error treatment since, often, suspect units/values to be manually reviewed are identified by using automatic detection methods.

The chapter is structured as follows: Section 3.2 discusses missing values, Section 3.3 examines systematic errors. Section 3.4 is dedicated to influential errors, outliers are discussed in Section 3.5, finally Section 3.6 deals with random errors.

All the different error types are described together with the corresponding detection method. For each method, the application context, as well as advantages and limitations are illustrated.

3.2 Missing values

3.2.1 Definition of missing values

Missing values stem from questions the respondent did not answer. Nonresponse can be due to several reasons; the respondent may not know the answer, may not be willing to respond or may have simply missed a question. Sometimes missing values are set due to inconsistencies by the editing procedure, but setting a flag instead of replacing these values allows a more detailed control of the E&I process. Therefore replacing inconsistent or unusable data with missing values should be avoided. If the respondent provides some but not all answers, the missing values are also called **item nonresponses**. Item nonresponse is usually treated by imputation (see Section 4.2). The case where the respondent did not answer any question is called **unit nonresponse**. Unit nonresponse is usually treated by weighting the responding cases accordingly. Unit nonresponse is beyond the scope of this manual; however, in some applications even unit nonresponse is treated with imputation.

A special case of nonresponse arises due to filter questions that lead to unanswered questions in inappropriate parts of the questionnaire. These values are sometimes called **structurally missing** to distinguish them from the values that the respondents should have given. For example, if an enterprise gives the answer "no" to the question "Do you have expenditures for research?", then the enterprise will not have to fill in the part of the questionnaire on research-related expenditures. Therefore, the corresponding data items will be structurally missing. This is only a problem if the "no" answer was incorrect.

Ignoring missing values can seriously affect the quality of data, potentially leading to biased conclusions and loss of precision. Bias arises when respondents and nonrespondents have different characteristics with respect to the investigated phenomena (Little and Rubin, 2002). The loss of precision is due to the reduction of the sample size. Furthermore, estimates from different data sources (including time series) lack comparability without treating the missing values.

Terms related to the mechanism that guides the response behavior are: **missing completely at random** (MCAR), **missing at random** (MAR) and **not missing at random** (NMAR). When missing values are MCAR, estimates based only on completely observed data are unbiased if the corresponding complete-data estimators are unbiased. However, MCAR often is an unrealistic assumption. MAR missing values may be modeled with observed covariates, but the assumption that MAR holds cannot be tested with the observed data. The NMAR nonresponse mechanism is difficult to model, because under this assumption, the probability that a value is missing depends on the value itself. See Appendix C.2 for a formal description of the nonresponse mechanism.

3.2.2 Detection of missing values

The detection of missing values usually is simple though the patterns of missing values may be complex. One problem is the distinction between structurally missing values and other missing values. A similar problem is that of the distinction between missing values and zeroes for quantitative variables. Such a problem can determine severe bias in the estimates especially in presence of semi-continuous variables.[1].

Since missing values lead to edit failures, they are often detected in practice by using the same edit checks used for detecting other types of errors (see Section 3.6). These edits can then be used in a deterministic or probabilistic framework. Nevertheless, when such an approach is not applicable because of the lack of edits expressing strong relationships between the involved items, sometimes

[1] Semi-continuous variables have a distribution where singular values, e.g. 0, occur with positive probability and for the rest their distribution is continuous. Variables of this type often occur in business surveys where distributions are frequently characterized by a mixture of zeros and continuously distributed positive numbers. Investments is a good example.

a probabilistic approach can be followed. It consists in distinguishing structurally missing values (or zeroes) from genuine nonresponses through the prediction of a suitable binary variable. In these cases, the corresponding model could be estimated by using auxiliary information, for instance historical data, provided that a sub-set of clean data is available to be used for estimation of the model parameters. Missing values detection results in a response indicator matrix R. The elements r_{ij} take the value 1 if the variable j contains an answer for the respondent i, and 0 otherwise. Structurally missing values may be given a particular value in the response indicator matrix, or a further indicator variable may preferably be used. The response patterns are defined by the rows of R.

Context (type of data) of use
Checks for missingness can be defined for all types of variables (categorical and numerical). Detection of missing values is needed when missing values and genuine zeroes (or structurally missing values) are not coded in a different way.

Advantages and limitations
The advantage of using edit checks for identifying missing values is that this approach does not imply the definition of additional edits with respect to the ones developed for other purposes. The main limitation is that, in some situations, it cannot be applied since logical or mathematical relations between variables do not exist. The probabilistic approach requires a modeling effort which may be of low quality if the needed auxiliary information is not available.

3.2.3 Recommendations

1. Replacing missing values by zero is neither an acceptable imputation procedure nor an acceptable alternative to flagging missing values.

2. Appropriate indicators on missing values should be calculated (see Chapter 6 and Appendix D). At the least, the number of observed values and the number of missing values and structurally missing values should be recorded for each observation.

3. The indicators should be analyzed to gain information on nonresponse mechanisms (see Chapter 2.3).

4. Variables and cases with many missing values should be studied and a decision on the variables and observations to be imputed should be taken in view of the amount of missingness. It may be necessary to discard certain variables or certain cases from the analysis or from the imputation because too many values are missing.

3.3 Systematic errors

3.3.1 Definition of systematic errors

A **systematic error** is an error that is reported consistently over time by responding units. It is a phenomenon caused either by the consistent misunderstanding of a question during the collection of data, or by consistent misinterpretation of certain answers in the course of coding. Systematic errors do not lead necessarily to consistency errors but always seriously compromise statistical results (United Nations, 2000), because they lead to bias in an estimate.
A well-known type of systematic error is the so called unity measure error. This error occurs when respondents report the value of a variable in a wrong unity measure. For example, let us suppose total turnover must be reported in thousands of Euros, but the entire amount is erroneously declared in Euros. This kind of error has strong biasing effects on target estimates and distributions. Unity

measure errors can remain undetected by consistency edits because if, for instance, total turnover is involved in a balance edit, it can happen that this edit is satisfied because all the components of turnover are reported with the same wrong unity measure.

Systematic errors could also be caused by the collection vehicle itself either because of a bug or because the embedded edits are not (or no longer) in line with the economic reality. It could be related to inadequate codes. Finally, it can be a diagnostic of interviewers who may need guidance.

Other examples of systematic errors are the following: errors associated with the frequent misunderstanding of the skip rules related to filter questions in the questionnaire; errors due to the respondent's inability to provide information based on the required classification/definition; sign errors that occur when respondents systematically omit a minus sign for variables, such as profit, that can be negative; systematic missing values such as the omission of the total value for a number of component items.

3.3.2 Detection of systematic errors

Systematic errors, especially unity measure errors, can result in outlying values in a specific direction which may be detected by outlier detection methods, see Section 3.5.

Methods introduced in this section aim at identifying the presence of systematic errors and error-generating mechanism. In fact, the decision of which variable is in error is generally guided by the knowledge of the error generating mechanism.

The main limitation of the techniques for detecting systematic errors in general is that one has to have an idea of what kind of systematic errors to expect and what their underlying error generating mechanism can be. For instance, regarding a specific systematic error mechanism like the unity measure error, it is possible to model the problem and to use a sophisticated technique based on finite mixture models like the one described in Di Zio et al. (2005a) and Di Zio et al. (2007). Without such, often specialized, subject-matter knowledge, it is usually hard to adequately detect the presence of systematic errors. If we have no idea of the kind of errors to expect, the following approaches may indicate the presence of systematic errors. Since systematic errors can result in outlying values in a specific direction, analysis of the nature of outliers detected through methods illustrated in Section 3.5 may be useful. The following techniques based on edits can be also effective.

Analysis of fatal edits

High failure rates of fatal edits may indicate the presence of systematic errors in one or more of the involved variables. The analysis of frequently failed edits may allow one to identify the variable(s) in error as well as the source of the error. Many errors within a single record indicate a possible problem with the respondent's understanding while the same error across many respondents indicates a possible questionnaire problem.

Analysis of ratio edits

There are two variables involved in a ratio edit (see the Glossary in Appendix E). Since a ratio edit generally identifies non-fatal errors, the study of the failure rate must be jointly performed with the graphical analysis of the two variables in order to detect potential systematic errors in these variables. For instance, considering the thousand error on turnover, when the total amount of employees is available and it is not affected by the same error, the edit based on the ratioturnover/employee would be useful to identify the cluster of observations affected by this error (resulting in a much higher value of the ratio than for the non-erroneous units). This example shows that ratio edits are particularly useful when variables potentially in error are related either to observed variables not affected by the same systematic error, or to variables from external sources or to historical data.

Once analysis of edits shows that data are affected by a specific systematic error, deterministic checking rules need to be formulated in order to indentify the units where this error occurs.

Context (type of data) of use
The analysis of fatal edits for detecting systematic errors can be applied to both categorical and numerical variables. The analysis of ratio edits can be applied to numerical variables only.

Advantages and limitations
A major advantage of systematic error detection is the fact that these errors can be imputed with a high level of confidence in the final results. Systematic error detection methods are cost-effective in terms of time and resources. As mentioned before, the main limitation of these methods is that one has to have an idea of what kind of systematic errors can occur. Without such knowledge, it is usually hard to adequately detect systematic errors.

3.3.3 Recommendations

1. Systematic errors should be detected and treated before dealing with random errors, in particular when the Fellegi-Holt method is used (see Section 3.6), and before selective editing (see Section 3.4).

2. The analysis of indicators on edits helps to find systematic errors mechanisms. If systematic error mechanisms are found by examining edits, then appropriate deterministic checking rules to detect errors due to the systematic error mechanism should be added.

3. If systematic error mechanisms are found, then improvements to the survey process (Questionnaire, interviewer training, coding, processing) should be made to prevent similar errors.

3.4 Influential errors

3.4.1 Definition of influential errors

Influential errors are errors in values of variables that have a significant influence on publication target statistics for those variables. Strictly related to the concept of influential error is that of **influential observation**. An influential observation is an observation that has a large impact on a particular result of a survey, i.e. a statistic. It may be correct or not and, in this latter case, it can generate an influential error. In business surveys influential observations are quite common. An important reason is that a small number of businesses can be much larger than others, in terms of number of employees, or other survey variables. Another reason is that some businesses may be given large sample weights so that, even though these companies are not large, their contribution to the target estimates is significant. The concept of influential errors is closely connected to an editing approach known as **selective editing**. In this approach, potentially influential errors are identified for each publication cell and editing efforts are mostly (or even exclusively) spent on these errors, usually by interactive treatment. Thus, selective editing is a way to make efficient use of the available editing resources to target predefined levels of accuracy for the most important estimates (Granquist, 1995; Granquist and Kovar, 1997; Hidiroglou and Berthelot, 1986; Latouche and Berthelot, 1992).

3.4.2 Detection of influential errors

The common approach to detect influential errors is **Selective editing**. This approach is illustrated in next section.

Selective editing

Selective editing can be done as soon as raw data records become available. In this approach raw data are split into a critical stream, with potential influential errors, and a non-critical stream. Selective editing is part of the micro editing phase, where each record is assessed individually. Since of course the true value is unknown, in order to identify potentially influential errors, selective editing requires a "prediction" of the true value, or **anticipated value**. The anticipated value can be based on a model, on validated values of variables from an earlier period, or on data from an external source. The anticipated value should be close to the expected true value of a variable. However, it is not necessary that this estimate is completely accurate; it is only used to make a comparison across units to the extent to which the values are atypical in order to prioritize the editing efforts. The estimate is not meant to be used as an imputation. The number of records with potential influential errors depends on the number of variables in the record: The larger the number of variables, the higher the probability of an influential error. Investigation of the raw and validated data of an earlier period may give insight as to the amount of influential errors that may be expected.

Score functions are often used to detect influential errors in records (Farwell and Raine, 2000; Hedlin, 2003; Hidiroglou and Berthelot, 1986). For the construction of a score function we need to distinguish two important components: **influence** and **risk**. The influence component quantifies the relative influence of a record on a publication estimate. The risk component quantifies either the extent to which a record violates edits or the extent to which it deviates from anticipated values.

A score function should make a correct selection of records that are sent to the critical stream and are usually treated interactively. This means that records that are likely to contain errors with a significant effect on the target estimate should be selected for interactive treatment. For instance, let us assume that the target estimate, based on a sample of n units, is the total for a variable y and that P% of the records are to be selected for interactive treatment. The selection should be such that the bias in the estimated publication total is minimized. Since the correct values are unknown, this bias cannot be evaluated and an approximation is used. This approximation is called relative pseudo bias and can be expressed in terms of the weighted differences between the raw values y_i and the anticipated values $\widetilde{y}_i$ in the units i not selected for interactive treatment:

$$\Delta(\widetilde{y}, y, \widetilde{M}) = |\frac{1}{\hat{Y}} \sum_{i \in \widetilde{M}} w_i(\widetilde{y}_i - y_i)|,$$

where $\widetilde{M}$ is the set of records not selected for interactive treatment and $\hat{Y}$ is a reference estimate of the total for the variable y, for instance, $\hat{Y} = \sum_i w_i \widetilde{y}_i$.

A **local score function** is a score function for one variable. It is often defined as a scaled difference between raw and anticipated values of a variable y in a unit i, also taking into account the influence of unit i. In this case it can be viewed as the contribution of the expected error in unit i to the relative pseudo-bias for the variable y. An example of a local score function is $SF_i = w_i|y_i - \tilde{y}_i|$, which can be thought of as the product of an influence component $w_i\tilde{y}_i$ and a risk component $|y_i - \tilde{y}_i|/\tilde{y}_i$.

In order to make the decision to edit a record interactively one needs a measure for the whole record rather than for specific variables within a record. To that end a **global score function** is defined by combining the local scores into a record level measure. A **cut off value** is set to decide when a record should be treated. That is, a given record is suspicious if the value of a score function exceeds a certain value.

To configure this method for a survey it is necessary to test it on the basis of a raw and a validated data file; for instance, from a previous period. Budget constraints can set a limit on the percentage of records that can be edited interactively. Alternatively, if the maximal deviation in publication estimates due to selective editing that is acceptable can be specified, then the necessary amount of interactive editing can be approximated using the estimated pseudo bias.

An alternative way of detecting influential errors uses edit rules. This approach assumes that influential errors will violate edit checks. It consists of the following three steps:

1. Select records that fail edits.
2. For each of these records, estimate the amount of changes for variables involved in failed edits needed to make the records satisfy edit constraints.
3. Use the estimated amounts of changes to build a score function to prioritize records to be manually reviewed.

Continuing the previous example, let us assume that the quantity to be estimated is the total Y, and that a record i fails a balance edit involving y. The amount of change on y can be estimated as the absolute difference between the reported total and the sum of its individual components. A simple score function can be obtained by multiplying this estimate by w_i and standardizing with respect to $\hat{Y}$.

Context (type of data) of use

Selective editing can be applied to continuous numerical variables.

Advantages and limitations

An advantage of selective editing is the gain in time and efficiency, because fewer records are edited interactively within acceptable levels of accuracy. However, especially in the case of many important variables it can be difficult to detect all records with influential errors in the micro editing phase. Further influential errors that are not detected by score functions may be detected in the macro editing phase.

Poor accuracy of the anticipated value with respect to the actual true value may produce bias on the score function. Units for which the score function cannot be calculated have to belong to the critical stream to prevent biasing the estimates.

Macroediting

In general, macroediting methods allow one to identify suspicious data in individual records on the basis of the analysis of preliminary survey estimates. These techniques are typically applied at the end of the E&I process to assess the validity of preliminary estimates and identify possible residual influential errors in data. Initial estimates from the treated survey are compared with results from an earlier period or a different source with comparable variables. The estimates considered are often means, totals and ratios, but other statistics like correlations, quantiles, variances may be considered, too. Relations between estimates across different sub-populations (economic activity, size class, regions) are examined to detect anomalies. Drilling down to sub-populations and individual units is usually necessary for the examination of anomalies and influential data. The contribution of individual observations to initial estimates is examined, e.g. by the sensitivity function (see Section 3.5.1). Starting from the most influential observation, individual data are manually inspected. The manual inspection stops when macro edits are satisfied or when estimates are considered of acceptable quality.

A traditional method that can be used to identify suspicious data which are influential on preliminary estimates is the *aggregate method* (Granquist, 1995). In the aggregate method checks are carried out first on aggregates, and then on the individual records of the suspicious aggregates. For a given variable, both aggregate and individual level checks usually compare current and past (weighted) data. Aggregates corresponding to values of checks exceeding pre-defined acceptance bounds are selected as suspicious. Checks at individual level are then carried out on all units belonging to a suspicious aggregate. Starting from the most extreme values in the ordered list of checks values, or focusing on units corresponding to check values out of pre-defined acceptance bounds, units are manually reviewed until there is no noticeable effect on estimates.

Macroediting methods rely on classical statistical techniques, in particular exploratory data analysis with graphical techniques. Scatterplots, scatterplot matrices and graphical displays of high dimensional point-clouds are used mainly to detect special groups and outliers. For practical applications of graphical editing supporting macroediting refer for example to De Waal et al. (2000), Weir et al. (1997), Houston and Bruce (1993).

Macroediting requires statistical skills, subject matter knowledge and good information on all previous stages of the survey and the E&I process. Often experts for the comparison values have to be consulted too.

As is the case for other statistical analysis it can become difficult to keep track of decisions and actions during macroediting. For effective macro-editing it is essential to fix objectives and strategies before starting it.

Context (type of data) of use

Macroediting can be applied to continuous numerical variables.

Advantages and limitations

Macroediting is crucial to end up with acceptable aggregates. However, there is a danger of over-adjusting the data to the past when doing macroediting. The inherent sampling and non-sampling variance of the aggregates examined should always be considered.

Macroediting is more efficient than micro editing. However, the time and cost of macroediting is difficult to predict. Drilling down to individual units may be necessary and parts of the micro editing phase may have to be repeated or added.

3.4.3 Recommendations

Selective editing

1. Selective editing is an appropriate method to focus attention on critical observations without generating a detrimental impact on data quality (Granquist and Kovar, 1997).

2. Priorities should be set according to types or severity of errors or according to the importance of the variable or the reporting unit. Priorities should be reflected in the score functions of selective editing. Score functions should include a risk and an influence component.

3. The thresholds of score functions should be chosen carefully during the tuning and testing of the E&I process. The thresholds and parameters of score functions should be revised whenever the survey process is changed (Questionnaire, data entry, removal of systematic errors).

4. The quality of the anticipated value should be assessed at least for a sub-sample of the survey. E&I flags have to be taken into account if data of earlier periods are used as anticipated values.

5. Units, for which the score function cannot be calculated have to belong to the critical stream to prevent biasing the estimates.

Macroediting

1. Macroediting should be performed before releasing data for final estimation.

2. Important publication aggregates, publications cells and publication sub-populations should be considered in macroediting.

3. The quality of reference data and problems like inflation and structural differences (e.g. definitions) should be taken into account.

4. Macroediting should be properly documented (Estimates, comparisons, graphs, anomalies, outliers etc.).

3.5 Outliers

In this section methods for the detection of outliers and influential observations are described. These methods include several techniques which are applicable once a sufficient amount of data is available to do statistical analysis (macro approaches). Specific macro editing approaches, including graphical macroediting, usually applied at the end of the E&I process to assess the validity of preliminary estimates are also described. The treatment of errors detected through outlier detection and other macroediting approaches is the same as for other detection methods (see Chapter 4). In particular, outlier detection may lead to interactive treatment and call-backs for individual units (as in selective editing), but also to robust estimation and imputation.

3.5.1 Definition of outliers

An **outlier** is an observation which is not fitted well by a model. The model can be a parametric distribution; e.g. we may declare observations in the tails of a normal distribution as outliers or the model can be a more loosely defined concept like "close to the center of the data". In the latter case an outlier is an observation which is not close to the center of the data. These concepts must be operationalized in practice. The model an outlier is compared with refers to the underlying population, not to the sample, and often there are different models appropriate for different sub-populations. For example a model may be appropriate only for the businesses with a particular economic activity.

An outlier may be defined with respect to only one variable (univariate outlier) or with respect to a set of variables (multivariate). Multivariate outliers are much more difficult to detect than univariate outliers. However, there may be genuine multivariate outliers which no univariate or bivariate method is able to detect.

The definition of an outlier is strictly related to the concept of an influential observation. As stated in Section 3.4, an influential observation is an observation that has a large impact on a particular result of a survey, i.e. on a statistic. Depending on the statistic under consideration an influential observation may or may not show up as an outlier. The statistic may be an estimator, like the Horvitz-Thompson estimator, or a p-value of a t-test for comparing the means of two domains of study. The influence of an observation on the statistic may be measured by the sensitivity curve evaluated at the observation. Thus the influence is the difference between a statistic $\hat{\theta}$ evaluated at the sample including the observation under consideration, say i, and the same statistic but without observation i, denoted $\hat{\theta}_{(i)}$. In other words the sensitivity of $\hat{\theta}$ to the observation i is $SC(y_i, \hat{\theta}) = c\ (\hat{\theta} - \hat{\theta}_{(i)})$, where c is a suitable standardization constant, e.g. $c = \sum_{i \in S} \hat{\theta}_{(i)}/n$.

Often influential observations show up as outliers. Since the large number of possible statistics makes it impossible to check all possibly influential observations individually usually one concentrates on detecting outliers at the macro editing phase.

Outliers in a sampling context may be classified as representative and non-representative outliers (Chambers, 1986). The latter are either incorrect observations whose true values would not show up as outlying, or are unique, but correct values, in the sense that one should not extrapolate them to other observations in the population. Representative outliers are correct observations which may have similar units in the population. The distinction is of importance for the treatment of the outlier but less for the detection.

Detection and treatment of outliers can be separated or they may be combined into robust procedures, in particular robust estimators. The separation of detection and treatment has the advantage of allowing to identify those outliers which correspond to incorrect observations through interactive treatment (if enough resources are available), in order to treat them as appropriate, while robust procedures allow a better tuning for efficiency.

The score functions in selective editing (see Section 3.4) are used to indicate the potential usefulness to revise an observation thoroughly (Lawrence and McKenzie, 2000). Score functions are similar to sensitivity curves of totals or ratios but instead of using $\hat{\theta}_{(i)}$ score functions use an anticipated value $\tilde{y}_i$, which should be a good predictor for y_i. The use of linear statistics, e.g. weighted means, and anticipated values opens the possibility to treat each observation separately, i.e. regardless of the data distribution.

3.5.2 Detection of outliers

The result of outlier detection often can be expressed as a robustness weight u_i per observation i with $0 \leq u_i \leq 1$. A value below 1 indicates an outlier. The robustness weight u_i may be just dichotomous, i.e. take values 0 or 1 only. In that case it is nothing but a flag. If the robustness weight is continuous it also indicates a degree of outlyingness. Robustness weights with intermediate values between 0 and 1 are often more efficient than dichotomous weights. The robustness weight may be specific to a variable or to a set of variables. Thus in the end there may be several robustness weights per observation. Robustness weights may be used directly in a robust estimator (see Section 5.3).

Univariate methods

Many methods consider the distance from a robust location estimator, often the median, and declare observations as outliers if their distance to the robust location is beyond some threshold. In order to obtain a scale invariant distance measure a robust scale must be estimated. An example of such a robust scale is the median absolute deviation but other robust scales, like the interquartile range, may be used, too. A simple robustness weight based on a (weighted) median and a (weighted) median absolute deviation is described in Appendix C.3.2.

Winsorized means and trimmed means based on the weighted empirical distribution function may be used as robust estimators of location. Implicitly they also define a robustness weight and thus can be used for outlier detection independently from the estimated value (see Appendix C.3.3).

Outliers in periodic data (Hidiroglou-Berthelot)

For periodic data Hidiroglou and Berthelot (1986) proposed to use a score for the detection of outliers which is based on the ratio of two consecutive measures on the same unit (see Appendix C.3.4). The method takes into account a distance of the ratio to the median ratio as well as the magnitude of the values themselves. The method needs three parameters to be specified and can be adapted to many situations. It has shown good performance and applicability in practice.

Methods based on regression models or tree-models

Instead of directly detecting outliers in a variable y_j we also may look for outliers in the residuals when a model is fitted to y_j, e.g. a regression model. The model must be estimated robustly to be able to detect outliers.

Also tree based methods like Classification and Regression Trees (Breiman et al., 1984) can be used for outlier detection if the models used in the splitting criterion are robust like in Chambers et al. (2004). The advantage of using models is that sub-populations and covariates may be taken into account. Generally the better the model fits the bulk of the data the simpler is outlier detection. For example it may be much easier to find outliers in turnover when a (robust) regression model with the explanatory variable number of employees is fitted first.

Multivariate methods

Many multivariate outlier detection methods use the (squared) Mahalanobis distance of an observation y_i (here a vector) from a center m when the covariance matrix is C: $d^2 = (y - m)^\top C^{-1}(y - m)$. Observations with a Mahalanobis distance above some threshold get a robustness weight u_i smaller than 1. For the choice of the threshold it is often useful to plot the Mahalanobis distance versus the quantiles of a Fisher distribution $F_{p,n-p}$, where p is the number of variables and n is the number of observations. The Mahalanobis distance is suited to a situation where the bulk of the data is distributed elliptically, e.g. the multivariate normal distribution.

To be useful for outlier detection the center m and the covariance matrix C must be robust estimators. Several methods exist but all have their disadvantages. Some are computationally very intensive, others are not robust enough when the number of dimensions grows. Most of them cannot deal with missing values and many cannot deal with sampling weights.

For data without missing values a minimum volume ellipsoid or minimum covariance determinant algorithm (Rousseeuw and van Driessen, 1999) or methods based on projections (Franklin and Brodeur, 1997) may be used.

An iterative procedure which can cope with a moderate amount of missing values and with sampling weights is the BACON-EEM algorithm (EUREDIT Project, 2004a). A simpler non-iterative procedure which can cope with a considerable number of missing values is the Transformed Rank Correlations algorithm (Béguin and Hulliger, 2004).

Context (type of data) of use

Outlier detection can be applied to numerical variables.

Advantages and limitations

Outlier detection allows the survey managers and methodologists to get acquainted with the data and the production process. It helps to form an understanding of the quality of the data, and to end up with acceptable aggregates. However, there is a danger of over-adjusting the data to the past when doing outlier detection. The inherent sampling and non-sampling variance of the statistics examined should always be considered.

Many outlier detection methods (and imputation methods or robust estimation methods) depend on tuning constants. These tuning constants allow one to choose the degree of robustness. It should depend on the amount and type of contamination or outliers in the data, which is unknown. Therefore tuning constants are often difficult to choose. For the skewed distributions often encountered in business surveys the treatment of outliers usually introduces a bias in estimates. The choice of the tuning constant has direct consequences for the bias. This must be taken into account already in the detection and even more in the treatment of errors.

Graphical techniques are very useful and may lead rapidly to the detection of outliers and influential observations. The well known rules for good graphs must be followed. It is often useful to start with a first exploratory analysis of the data, then apply a given outlier detection method, whose tuning can profit from the exploratory analysis, and in the end do statistical data analysis again during result validation.

Graphical techniques also have their limitations. For example, it is very difficult to detect anomalies and outliers in more than two dimensions by graphical techniques. The shear number of possible graphs may be overwhelming.

3.5.3 Recommendations

1. Outliers and influential observations should be detected. Such observations may be errors or correct but sometimes their correctness is unclear. Errors should be treated and also correct or unclear influential or outlying observations may have to be treated to prevent potential bias and high variability.

2. The influence on important results should be controlled even after selective editing and outlier detection.

3. Outlier detection methods should be robust against outliers.Therefore methods based on means or weighted means and standard deviations, which are not robust, should be avoided.

4. Simple univariate methods, graphical displays or more complex multivariate techniques should be used for outlier detection depending on possible models among variables. Different models may have to be applied in different sub-populations.

5. When choosing tuning constants for detection or treatment several tuning constants should be tested and the corresponding impact on estimates and their variances observed. Often the treatment of a few outliers is acceptable.

3.6 Random errors

3.6.1 Definition of random errors

Random errors are errors that are not caused by a systematic reason, but by accident. They primarily arise due to in-attention by respondents, interviewers and other processing staff during the various phases of the survey cycle. An example of a random error is an observed value where a respondent by mistake typed in a digit too many. Random errors are often defined as non systematic errors. In the statistical context the expectation of a random error is typically zero. In our context, however, the expectation of a random error may also differ from zero. This is, for instance, the case in the above-mentioned example.

Random errors can result in outlying values and can be detected by methods for detecting outliers (see Section 3.5). Random errors can also be influential and can then be detected by the methods described in Section 3.4. Here we will concentrate on the detection of random errors that are not influential and do not result in outlying values but lead to inconsistent records because some edit rules are violated.

3.6.2 Detection of random errors

Once the edit rules are defined and implemented, it is straightforward to check whether the values in a record are inconsistent in the sense that some edit rules are violated. It is, however, not so obvious how to decide which variables in an inconsistent record are in error. This activity is usually called **error localization** There are two different approaches for error localization. The first approach is to use deterministic checking rules. The second approach is to use a general guiding principle for the localization of the erroneous fields in an inconsistent record.

Deterministic checking rules state which variables are considered erroneous when the edit rules are violated in a certain record, for instance because that variable is less reliable. An example of a simple deterministic checking rule is: if component variables do not sum up to the corresponding total variable, the total variable is considered to be erroneous. However, this is not the correct way to deal with random errors as the same variable is considered to be in error for all records. Often the deterministic checking rules are coupled with deterministic imputations (see Section 4.2).

There are several general guiding principles for the localization of the erroneous fields in an inconsistent record. The best-known and most-used of these general guiding principles is the **Fellegi-Holt paradigm**, developed by Fellegi and Holt (1976). This paradigm is, in fact, only one of three principles for automatic edit and imputation proposed by Fellegi and Holt. These three principles are:

1. The data in each record should be made to satisfy all edits by changing the fewest possible items of data (fields);

2. As far as possible the frequency structure of the data file should be maintained;

3. Imputation rules should be derived from the corresponding edit rules without explicit specification.

In the context of localizing random errors the first one of these principles is referred to as the "Fellegi-Holt paradigm" With regards to error localization it is the most important principle of the three. The other two principles relate to automatic imputation after random errors have been detected.

In due course the original version of the Fellegi-Holt paradigm has been generalized to: the data of a record should be made to satisfy all edits by changing the fewest possible (weighted) number of fields. Here each variable in a record is given a weight, the so-called reliability weight of this variable. A reliability weight is a measure of "confidence" in the value of this variable. Variables that are generally correctly observed are given a high reliability weight; variables that are often incorrectly observed are given a low reliability weight. A reliability weight of a variable corresponds to the error probability of this variable, i.e. the probability that its observed value is erroneous. The higher the reliability weight of a variable, the lower its error probability.

Usually, software based on the Fellegi-Holt paradigm also has a module for imputing missing and erroneous data. Fellegi and Holt basically note in their second principle, which was originally formulated in the context of categorical data, that imputation should result in the preservation of the distribution of the true data.

The Fellegi-Holt paradigm can often be used successfully in the case of random errors. Although error localization and imputation are related and should be applied in combination - basically this is the third principle of Fellegi and Holt - the Fellegi-Holt methodology only ensures the construction of internally consistent records; not the construction of a data set that possesses certain (distributional) properties. That is, the the Fellegi-Holt paradigm does not ensures that edited and imputed values equal the true values or that the means and (co)variances of the edited data equal the means and (co)variances of the true data. The only result that can be obtaoned with certainty is that the edited data satisfy the edits and as few as possible data values are changed to achieve consistency. However, provided that the set of edits used is internally consistent and sufficiently powerful, application of the Fellegi-Holt paradigm generally results in data of higher statistical accuracy than other approaches like deterministic checking rules, with respect to location and scale estimates. For a more detailed description of the application of the Fellegi-Holt paradigm we refer to Appendix C.1.

Context (type of data) of use

The above-mentioned techniques for detecting random errors can be applied to categorical and numerical variables. They require that the relations between variables can be expressed through fatal edits.

Advantages and limitations

Deterministic checking rules

The advantages of using deterministic checking rules are the transparency and simplicity of the correction process.

A limitation of the deterministic checking rule approach is that many detailed checking rules have to be specified. Developing such rules can be quite time consuming and may require many resources. Maintaining and checking the validity of a high number of detailed checking rules can be a practical problem. In some cases it may be impossible to develop deterministic checking rules that are powerful enough to identify errors in a reliable manner. Another disadvantage is the fact that a systematic bias can be introduced, as a random error is detected in a systematic manner. If this bias is small, this will not pose any problems. The problem is that usually it is not possible to estimate this bias.

Fellegi-Holt method

Advantages of the Fellegi-Holt paradigm are that it aims to preserve as much of the originally observed data as possible, that it leads to consistent data satisfying all edits, and that it is suitable for application to many variables simultaneously. Other strong points in comparison to the deterministic checking rule approach are that less detailed rules are required and that bias due to the deterministic treatment of random errors will not be introduced.

A limitation of adopting the Fellegi-Holt paradigm is that a system based on the Fellegi-Holt paradigm considers all edits specified for a certain record as fatal ones. Other drawbacks are that the Fellegi-Holt paradigm can only be successfully adopted if sufficiently powerful edits can be specified and that the method is not easy to implement. Most developed software packages for the Fellegi-Holt paradigm are not generally available, or may either be considered too expensive or unsuited to fit into the overall IT-architecture of the statistical office.

3.6.3 Recommendations

1. If appropriate software is available, random errors should be detected and treated by applying the Fellegi-Holt paradigm.

2. If appropriate software for applying the Fellegi-Holt paradigm is not available, random errors may be detected by means of deterministic checking rules. However, the error localization becomes more arbitrary with deterministic checking rules.

Chapter 4

Treatment of Errors

4.1 Introduction

During the editing phase, the potentially erroneous values in the data are determined. In a second step, the items that are believed to be in error are replaced with more plausible values. Although error detection and error treatment are conceptually distinct, in practice it is often hard to clearly separate the former step from the latter. For instance, the localization of erroneous values might be based on estimating values first and then determining the deviation between the observed and estimated values. The observed values that differ most from their estimated counterparts are then considered erroneous, or in any case as suspicious. In this approach the detection of erroneous values and the estimation of better values are highly intertwined. During interactive treatment (see Section 4.3) the detection of erroneous values and the "estimation" of better values are often not separated activities as well. This "estimation" often simply consists in filling in correct answers obtained by re-contacting the respondents.

In order to replace values in error or to fill in missing values, basically two approaches are available. The first approach is based on interactive treatment, where either respondents are re-contacted or subject-matter knowledge is used. In Section 4.3 we discuss interactive treatment in more detail. The second approach is based on statistical estimation techniques.

In principle one could use a statistical estimation technique to estimate the values of the target variables directly, without estimating values for the individual erroneous and missing values. In practice, however, this approach often has to be based on overly simplified models, otherwise the approach quickly becomes very complex or computationally demanding. Therefore, instead of directly applying a statistical estimation technique, one often replaces the erroneous or missing values by estimated ones at the record/variable level. This process is called **imputation** and is described in Section 4.2. After imputation of the missing and erroneous values a complete data set is provided, from which estimates can be obtained by using relatively simple and easy to apply standard estimation techniques. Imputation thus simplifies the estimation procedure. For instance, in a traditional sample survey where sample units are weighted to represent the corresponding population units, one first imputes the missing values and then uses the weights to calculate the population parameters by means of a standard weighting scheme.

4.2 Imputation

Missing data resulting from nonresponse (generally item nonresponse) or values flagged as erroneous during editing may be treated by imputation. Imputation consists of replacing these values by plausible ones to meet the demands on data quality for analysis and dissemination.

Imputation methods can be partitioned in four broad classes: rule-based imputation, deductive imputation, model based imputation and donor-based imputation. In some imputation methods the imputed values are determined uniquely (**d**eterministic imputation methods). In some others a random residual is added to the imputed values (**s**tochastic imputation methods) in order to take the variance of the nonrespondents partially into account.

The main problems caused by missing values is that estimates may be biased and the variance may be underestimated (see Section 3.2). Bias can be reduced by imputation if the nonresponse is characterized by a MAR mechanism, (Rubin, 1987; Schafer, 2000). Otherwise bias will persist and usually its impact on estimation cannot be estimated. In many cases MAR mechanism is assumed within subgroups of data (**i**mputation cells) and separate treatments are performed in the different sub-groups. Further, imputation introduces an extra component of variability that must be considered in estimation. Analyses performed on imputed values treated as if they were observed can be misleading when estimates of the variance do not include the error due to imputation (variance underestimation). Some techniques can be used to deal with variance underestimation, like analytic techniques, resampling methods, multiple imputation (see Section 5.3).

The choice of the most appropriate imputation method strongly depends on the survey objectives. For example, a method that leads to unbiased estimates of the mean may produce a strong bias in estimates of quantiles.

4.2.1 Imputation cells

Description

Imputation cells partition data into homogeneous sub-groups such that the subsequent imputation reduces the item nonresponse bias (Haziza, 2002).

Imputation cells are defined by using auxiliary variables like stratification variables or variables used as publication domains, so that the variables to be imputed are as homogeneous as possible within the cells. The choice of imputation cells corresponds to an implicit model for the variables to be imputed. The variables used to form the imputation cells must be known for all observations in the sample.

Context (type of data) of use

Any type of data can be used to form imputation cells. Imputation cells can be used for any imputation method.

Advantages and limitations

The bias due to item nonresponse can be reduced considerably if the nonresponse mechanism is MAR or MCAR within cells. Furthermore, a strong association between variables to be imputed and auxiliary variables used for the definition of the imputation cells, results in a more accurate imputation. On the other hand, introduction of many imputation cells can cause that some cells contain only few observations, resulting in a high variability of the estimates.

4.2.2 Rule-based imputation

Description

With this method, through subject-matter expert knowledge, the values to be imputed are determined by rules based on the values of the other fields and/or the erroneous values to be replaced. Rule-based imputation is generally based on "IF THEN" rules and is often not separated from the error localization procedure. Example: "If *number of employees*=0 and *worked hours*>0 then *worked hours*=0".

Context (type of data) of use

Rule-based imputation is appropriate when, in the presence of systematic errors whose nature is known, the imputation action is quite obvious. (for instance in the case of systematic unity measure error, see Section 3.3). Rule-based imputation can be used for both categorical and numerical variables.

Advantages and limitations
Rule-based imputation is usually simple to implement and it allows one to recover the true value when the error source is easy to identify. Nevertheless, in general, it should be used very carefully. In fact, such imputation rules may lead to severe bias in the estimates if the error source cannot be identified with certainty in all cases. Further it is generally difficult to set up a set of rules that ensures consistency of the imputed data with respect to a large set of edits.

4.2.3 Deductive imputation

Description
Deductive imputation is performed when, given specific values of other fields, and based on a logical or mathematical reasoning, a unique set of values exists causing the imputed record to satisfy all the edits (e.g. when items must sum up to a total and only one item in the sum has to be imputed, then its value is uniquely determined by the values of the other items).

Context (type of data) of use
Deductive imputation can be used in any context. It is particularly useful when for several observations the constraints lead to a unique set of values, values that allow the record to pass the edits. This typically occurs in the presence of balance edits.

Advantages and limitations
Deductive imputation is often viewed as a reliable method because the result is deterministically defined at the unit level and based on logic reasoning. It leads to the true values with certainty if errors in the data have been perfectly localized. Consistency constraints concerning variables not included in the reasoning are not ensured and may compromise the uniqueness of the solution. Deductive imputation is the simplest and cheapest method of imputation.

4.2.4 Model based imputation

Description
In model based imputation the predictions of missing values are derived from explicit models. An imputation model predicts a missing value using a function of some auxiliary variables. Under the MAR assumption the model can be correctly estimated using the observed data. The auxiliary variables may be obtained from the current survey or from other sources. Typical auxiliary variables are the variables of the sampling frame (e.g. size class, branch of economic activity), historical information (e.g. value of the missing variable in a previous period) and administrative data. The most common types of model based imputation are regression imputation (see Appendix C.4.1), ratio imputation (see Appendix C.4.2) and mean imputation (see Appendix C.4.3) which can all be described as regression imputation methods.
For categorical variables predictions usually result from logistic or log-linear models.

Context (type of data) of use
Model based imputation can be used for any type of variables. The auxiliary variables can be either continuous or categorical. Categorical variables are generally transformed to dummy variables if they are used as predictors in regression models. Imputation is usually done within imputation cells

Advantages and limitations
Model based imputation is easy to apply since standard software exists and is usually fast enough for the treatment of large data sets. The properties of the used models are well known. For instance, if

the relation between the variable to be imputed and the auxiliary variables is the same for respondents and non respondents, the mean of the imputed data is an approximately unbiased estimator for the true mean. If imputation cells are used, this holds within cells. Generally valid imputation are obtained only under MAR condition.

Model based imputation provides good predictions for the imputed values if the model is carefully specified. Specifying models can be time consuming since it must be done for each target variable separately. Extensions of linear models to the multivariate case, where a number of variables with missing values are imputed simultaneously makes it unnecessary to specify different models for different patterns of missing values and can preserve covariances between variables. The size of the data, on which the multivariate models are based on, is however more limiting in comparison with univariate models. This can compromise the validity of the multivariate model. Consistency and non-negativity constraints of the data after imputation are not ensured if they are not explicitly taken into account. Non-negativity constraints can be handled by an appropriate transformation of variables. To achieve consistency, the imputed values can be adjusted by a separate algorithm (Pannekoek and De Waal, 2005). A further problem with model based imputation is that the model usually refers to the population, not to the sample and, therefore, the sample design should be appropriately reflected in the estimation procedure for the parameters. In particular, a weighted estimation procedure should be used, with weights equal to the inverse of the inclusion probabilities to ensure consistent estimates of the model parameters (Skinner et al., 1989).

4.2.5 Donor-based imputation

Description

In donor-based imputation the missing or inconsistent values of units (called recipients) which fail any edit rules are replaced with observed values from another record (the donor). If the donor is chosen among the observations in the actual survey the method is called hot-deck, otherwise if the donor belongs to other sources (e.g. historical data) the method is called cold-deck. The set of donors (**donor pool**) generally consists of observations considered not erroneous. The way used to choose the donor differs among several types of donor imputation. For instance, in random donor imputation a donor is randomly chosen in the donor pool for each unit with missing values, while in **nearest neighbor imputation** (NNI) the donor is chosen in such a way that some measure of distance between the donor and the recipient is minimized (Kalton and Kasprzyk, 1982). If more than one donor has the same minimum distance for a given recipient then the donor should be randomly selected among them.

Context (type of data) of use

Every type of variable can be treated. Random donor imputation is usually performed inside imputation cells. A substantial number of donors in the donor pool is needed to ensure good performance.

Advantages and limitations

Donor-based imputation can handle variables that are difficult to treat by explicit modeling, for instance semicontinuous variables (Javaras and van Dyk, 2003). Since observed values are used for imputation, no synthetic values can be imputed. Under certain conditions donor-based imputation can preserve population distribution (Chen and Shao, 2000). Consistency of the imputed observations with respect to edit rules is generally not ensured. However, consistent data can be enforced by adjusting the imputed values by a separate algorithm, or restricting the donor pool to donors which result in consistent imputation. In the latter case, however, it can be almost impossible to find appropriate donors when missing values refer to variables involved in balance edits.

When all missing values of a given recipient are imputed from the same donor, the multivariate relationships are better preserved. However, if more variables are treated at the same time difficulties arise to find a sufficient number of donors. On the other hand, splitting up the set of variables results in splitting up the covariance matrix and may induce inconsistencies in the imputed data. In NNI, the choice of a distance function needs careful analysis of the data and of the results of the imputation: often conflicts between distance functions and number of available donors occur. Generalized software for donor-based imputation in business surveys exists.

4.2.6 Recommendations

1. Deductive imputation, where only one possible correct value exists, should be the first method to be considered.

2. Rule-based imputation should be used only when the error nature is well understood (i.e. in case of systematic errors) and should be avoided in all other cases.

3. The models used for imputation should be carefully validated for each variable separately and for groups of variables. Models, in particular in the form of imputation cells, should yield good prediction of missing values and should be stable. Small imputation cells and large models should be avoided. It may be necessary to consider the sampling weights or to include the variables of the sample design into the models.

4. For donor imputation the distance function should be chosen carefully during design, tuning and testing and it should be documented in detail. The size of the donor pool, the utilization of each donor, the donor per recipient and the corresponding distance should be monitored, documented and analysed.

5. The consistency of imputed observations should be checked and the impact of imputations on estimates and variances should be evaluated.

6. The imputed values should be flagged.

4.3 Interactive Treatment

In some cases, once errors and/or anomalous values have been detected, data are interactively treated instead of being processed by automatic procedures. In this procedure, generally supported by ad-hoc or generalized software, the reviewer starts by analysing the type of inconsistencies occurring in the erroneous record and tries to identify the source of error. Sometimes, the error source can be easily identified by simply checking data after they are coded and stored in electronic files. In some situations, on the contrary, the reviewer has to check the hard-copy questionnaires or scanned versions of these questionnaires. Checking the original data can help the reviewer to discover some (possibly systematic) errors in filling in the questionnaire, due to the misunderstanding of some questions by the respondent, or to imperfections in the data capture strategy. As an example, we can consider the situation where in a table of a questionnaire the answers are shifted by a row, so that there is no correspondence between reported values and variables they are referring to. This kind of error is difficult to detect by automatic procedures, while it can be easily identified through the inspection of the hard copy questionnaire or its electronic image. Furthermore, going back to the questionnaire allows for discovering possible discrepancies between original data and electronic data due to errors in the coding and/or data capture phases. On the other hand, often errors are not obvious, and it may be difficult to remove inconsistencies from data if there is no evidence of which items are responsible for those inconsistencies. Thus, it is extremely important that interactive treatment is performed by subject matter specialists, and that clerks without a deep matter knowledge limit themselves to correct only obvious cases. In fact, interactive treatment may increase the data quality if subject-matter knowledge is used to treat the data, since in this case, for each value that is considered erroneous, a tailor-made approach can be applied, where, for instance, detailed information (e.g. historical) about this particular respondent is used. This means that the specialist must have deep and confirmed knowledge of the treated unit and all variables of the questionnaire as well as deep knowledge of the E&I process. On the other hand, interactive treatment performed by not expert clerks may lead to highly subjective, procedures (creative editing). Unfortunately, using subject-matter knowledge for each erroneous value is highly time and resources consuming. and the resources spent on interactive treatment activities are often not justified by the resulting improvement in data quality, contributing to the so-called over-editing problem. This is the reason why generally, interactive treatment is restricted to some small sub-set of incorrect records that are previously selected for their high suspiciousness or their potential impact on the target estimates. When items in error cannot be localized with certainty, the ''true'' values often cannot be deduced by the erroneous ones. In these cases the respondent has to be re-contacted (followed-up) to correct the data. Calling back respondents can also be useful to validate or correct data that lie in the tail of the observed data distribution, like outliers, even though they do not fail any consistency rule. However, call back should be restricted to as few cases as possible, in order to avoid an excessive burden on respondents and too high costs. Furthermore, respondents' ability to report should not be overestimated. In fact, if the structure of the questions does not fit their understanding, no amount of badgering will get the "correct" answers out of them. To this regard, it is worth noting that follow-up should be also focused on the problems the respondents have in providing the correct answers instead of merely on trying to get more plausible values of questionable data.

Advantages and limitations

Interactive treatment of incorrect/suspicious data allows, more than automatic procedures, to put subject-matter knowledge into editing and imputation. Furthermore, the activity of interactively checking erroneous data may have as an output a better understanding of possible systematic sources of error. Compared to automatic deterministic E&I techniques based on IF-THEN rules, interactive editing can manage each individual situation in a different way, allowing appropriate treatment when the nature of the error is evident, or flagging records to be adjusted through re-contact of the respondent.

An important limit of interactive treatment is the risk of creative editing, that is that reviewers' adjustments are subjective. When data changes are not justified by evidence and merely reflect the opinion of the reviewer, interactive treatment can lead to serious bias in data. Interactive treatment often has a high inter and intra-personal variability which is not reflected in subsequent variance estimates. Repeatability of the interactive editing cannot be guaranteed. Consistency and completeness could be hardly ensured by interactive editing especially when many checking rules have been specified. Another serious limitation of interactive treatment is that it is an expensive process in terms of producers costs, respondents costs, and losses in timeliness (Granquist, 1995). Due to its high costs, interactive treatment is generally limited to a small sub-sample of data corresponding to influential units and/or to potentially influential errors in order to maximize the gain in data quality. However, for surveys with many publication variables interactive treatment cannot always be restricted to a small subset of records. When many records have to be checked, interactive treatment is often not performed by experts but by clerks which are trained on specific data characteristics. In these circumstances the quality of the editing is generally poorer than it would be if questionable values were inspected by specialists, or even treated automatically.

4.3.1 Recommendations

1. Interactive treatment should be restricted to problems which cannot be solved by automatic treatment. Interactive treatment should be limited to the most relevant errors (e.g. influential errors and outliers).

2. Interactive treatment should be performed by specialized and trained reviewers. Interactive treatment, and in particular manual imputation, must follow strict guidelines, which are designed, tested, tuned and monitored closely.

3. Information from previous surveys or administrative data and access to the original questionnaire should be used when available.

4. Number and length of call-backs to respondents should be kept minimal in order to reduce the burden on respondents and to avoid a negative impact on response rates. When doing call-backs, information on problems encountered by respondents should be collected to improve the survey.

5. Interactive treatment and, in particular, changes to the data should be fully documented and flags should be set for each action.

Chapter 5

Subsequent Analysis and Estimation

5.1 Introduction

The prototype E&I process as sketched in Figure 2.1 ends with the release of the final micro data, ready for the estimation of primary results such as totals or ratios and sometimes also for further multivariate analysis in the course of scientific research. In these final data, errors and missing values may have been detected and treated by the methods of Chapters 3 and 4. However, this does not mean that at the stage of estimation and analysis we can ignore the existence of errors and missing values in the original data, see Lundström and Särndal (2001) for an overview of problems with estimation on imputed data. There are two general reasons why, even after the E&I process is finished, we are still confronted with the consequences of these problems.

The first reason is that the data now contain imputed values that cannot be treated as truly observed data. As mentioned in Section 4.2, the imputations may lead to biased parameter estimates. Consequently, if an important part of a statistic is based on imputed values then the results should be cautiously interpreted because of this danger of bias. Even if globally the amount and impact of imputation is small, it may happen that this is not true for certain sub-populations, therefore the amount and impact of error treatment and imputation should be checked when analyzing sub-populations. The bias depends on the combination of imputation method and target parameter. For instance, for the estimation of means and totals, deterministic regression imputation will have the same bias as its stochastic counterpart, because the added residuals have expectation zero. However, for the estimation of measures related to the spread of a distribution such as quantiles, methods that impute estimated expected values (e.g. mean, ratio and regression imputation without added residuals) will be biased downwards. Apart from the possible bias in estimators, imputation, and in particular imputation of estimated expected values, will also lead to underestimation of the variance of estimators if standard variance estimators are used. Alternative variance estimators that are developed for imputed data are discussed in Section 5.2.

The second reason is that some missing values and outliers are explicitly not treated by the E&I process because it is considered more appropriate to deal with them by special estimators that take missing values or outliers into account. Examples of such estimators are discussed in Section 5.3. It can also happen that for some purposes the missing values, errors or outliers are not sufficiently treated by the E&I process since the E&I process of a survey is targeted to the most important statistical results. Subsequent analysis of particular aspects of a survey, e.g. a regression analysis or analysis of sub-populations, may need further processing appropriate for that particular purpose.

5.2 Variance estimation in the presence of imputed data.

Imputed data usually exhibit less variability than true data. This has important consequences for the estimation of the variance of parameter estimates. The precision of these estimates is overstated if the imputed values are treated as actual observed values. As a result the confidence intervals of the estimates will be too short and consequently this may lead to erroneous conclusions. In order to obtain valid variance estimators for parameter estimates based on imputed data, the following three general approaches have been advocated.

Analytical methods
In this approach a formula is derived for the variance of an estimator that relies partly on imputed data. It consists of a component that estimates the variance in case of a complete sample and adds variance component(s) to reflect the uncertainty associated with the estimation (imputation) of the missing values. These methods rely on either the assumption of the correctness of the imputation model (model assisted approach) or assumptions on the response mechanism (two phase sampling approach), see Appendix C.2.1. In both instances these methods will result in formulas that depend both on the target parameter and the imputation method used. Typically they have been derived for simple parameters such as means or totals and for simple imputation methods e.g. mean or ratio imputation and using simple random sampling. For more complex survey designs or imputation algorithms these methods becomes analytically difficult.

Resampling methods
Resampling methods use a large number of samples drawn from the observed sample to simulate the distribution of a statistic. The jackknife technique is a resampling method that can be used to obtain a valid variance estimate based on imputed data (Rao and Shao, 1992). The jacknife variance estimate of an estimator is calculated as follows. Each sampled element i, $i = 1, \ldots, n$ is removed from the sample once resulting in n samples of size $n - 1$. Next each sample is imputed and the estimator of interest is calculated. The jackknife variance estimate is based on the n values for the estimator.
Another well-known variance estimation procedure that falls in this category and can be applied to the problem of variance estimation with imputed data is the bootstrap technique (Shao and Sitter, 1996). In this case a random sample with replacement of size n is drawn from the observed sample. This means that in the bootstrap sample some respondents may be present more than once and others may not be present at all. Next the bootstrap sample is imputed and the estimator of interest is calculated. The bootstrap variance estimate is based on the different values for the estimator for a large number bootstrap samples.
A major advantage of these methods is the fact that the variance of complicated estimators can be calculated relatively easy without the theoretical derivation of variance formulas as is the case with the analytical method. On the other hand the methods may be computationally intensive and they have to be adapted to the particular sample design of the survey.

Multiple imputation
The variance due to imputation can also be estimated by stochastically imputing several, say m, times and calculating the variance based on a combination of the within and the between variance of these m datasets. This is referred to as *multiple imputation*, which was developed by Rubin (1987). This method makes variance estimation in the presence of imputation relatively simple. An advantage of multiple imputation, that it shares with resampling methods, is that it is independent of the parameter to be estimated: the same approach can be use many different parameters .
The main consideration in choosing between multiple imputation and the resampling or analytical methods is whether the benefit of an immediate relatively simple variance estimation outweighs the

simplicity of imputing only once (single value imputation). Although theoretically appealing since it incorporates the idea that imputations have a certain variability, multiple imputation is not used often in (large) surveys or at statistical agencies because of the practical implications. Multiple imputation requires maintaining and storing multiple complete data sets, which is operationally difficult. Besides, data dissemination such as the tabulation of data is complicated by the existence of multiple data sets.

Moreover, in order to obtain valid inference with multiple imputation, the imputation method used needs to be proper (Rubin, 1987). In words, a proper method is a method that has enough variability between replicates to provide appropriate variance estimates. Some commonly used imputation methods, including random hot deck and random regression imputation, are improper because these draws do not represent the full uncertainty in estimating the data for purposes of variance estimation with multiple imputation. Another consequence of this is that multiple imputation can only be used for stochastic imputation methods, as there will be no variability with deterministic procedures. Note that this means that the popular nearest neighbor method cannot be used. Taking all this into account, statistical agencies currently prefer the use of single value imputation.

5.3 Estimation in the presence of missing values and outliers

Missing values

Analogously to the calibration for dealing with unit nonresponse, estimators may be calibrated for dealing with item nonresponse (Särndal and Lundström, 2005). In this case the calibration should be carried out separately for each variable (or pairs of variables). This leads to a multitude of different weights which are difficult to handle. Therefore this approach is not recommended in general.

The EM-algorithm (Dempster et al., 1977) is another alternative to deal with missing values directly. It is designed to handle incomplete multivariate data and can provide estimates of model parameters. An example is the use of the EM-algorithm for the estimation of the mean vector and covariance matrix of a multivariate normal distribution based in incomplete data. The mean and covariance may be the parameters of interest themselves but the target parameters can also be the coefficients of a multiple regression that can be derived from them. However, the estimates are based on an explicit stochastic model and i.i.d. assumption and straightforward application of such methods do not take the sampling design into account. This may lead to invalid inference especially if the sampling design differs from a simple random sample (e.g. unequal inclusion probabilities, clustering).

Robust estimation

Robust estimators detect outliers and modify either the final weight of an outlier or the value of one or more variables. Often downweighting and changing the value is equivalent.

Censored means (Searls, 1966) and winsorized means (Fuller, 1970) replace values beyond a pre-specified cut-off value by the cut-off value or replace values beyond a quantile (with pre-specified probability) of the empirical distribution function by the quantile (see Appendix C.3.3). M-estimators (Huber, 1964) can be adapted to the sampling situation (Gwet and van Rivest, 1992; Hulliger, 1995) and allow a flexible handling of outliers. In business surveys often only the large observations are treated. For the skew distributions often encountered in business surveys robust estimators have a bias which is more than compensated for by their low variance. However, the choice of the trimming, winsorization or tuning constant (for M-estimators) is sometimes difficult (see Appendix C.3).

A simple robust estimator for linear characteristics can be formed with a robustness weight $u_i \in [0, 1]$. Robustness weights can be derived from outlier detection methods, e.g. from the mad-rule (see Appendix C.3.2), from robust estimators like M-estimators (Hulliger, 1999) or winsorized means (see Appendix C.3.3), or even from multivariate outlier detection (see Section 3.5). If a robustness weight u_i is available, a robust estimator of location can be formed by replacing the weighted mean estimator $\sum_S w_i y_i / \sum_S w_i$, which is an approximately unbiased estimator of the population mean, by a robustified version $\sum_S u_i w_i y_i / \sum_S u_i w_i$.

It is often advisable to use the robustness weights of the outlier detection just to form an initial guess for a robust estimator and recalculate the robustness weights for estimation. The reason is that while the distinction between representative and non-representative outliers (see Section 3.5.1 and Chambers, 1986) is not important for detection, for imputation or estimation representative outliers ideally should be down-weighted less than non-representative outliers to avoid a strong bias. Unfortunately, often we cannot distinguish between the two types of outliers and must account for possible representative outliers just by choosing the tuning constant of the robust estimator somewhat larger than a corresponding tuning constant for outlier detection. As a rule of thumb the tuning constant should be chosen such that the mean robustness weight should be close to 1 (usually above 0.98).

There are robust procedures for many estimation problems with complete data. However, these methods have to be adapted to the sampling context and to missing values.

Model based robust estimators for sample surveys are introduced in Chambers (1986), robust calibration estimators in Duchesne (1999) and robustified Horvitz-Thompson estimators in Hulliger (1995). A particular downweighting method is to group the non-representative outliers in a poststratum where all observations get the final weight $w_i u_i = 1$. This is proposed in Hidiroglou and Srinath (1981). An overview of robust estimation methods in business surveys is given in Lee (1995). In practice often the simple methods of Hidiroglou and Srinath (1981), univariate winsorisation (see Appendix C.3.3) or the one-step robustified ratio estimators of Hulliger (1999) (see Appendix C.3.5) are used to cope with outliers.

5.4 Recommendations

1. Estimation and analysis should take into account the foregoing E&I procedures. Any automatic treatment by imputation, deletion or downweighting of outliers and any interactive treatment may have consequences for the estimation procedures, the estimates and the analysis. Bias of totals, means and proportions may have been reduced or even increased by E&I procedures and correlations may have been affected, too.

2. Problems that can be solved relatively easily at the estimation stage should not be solved by E&I procedures. In particular weighting for unit-nonresponse often is preferable to imputing complete units. On the other hand, problems that have been solved by E&I procedures should not be addressed at the estimation stage.

3. Robust procedures which can be expressed by robustness weights are often suitable for business surveys. The robustness weights should be joined to the final data set.

4. The classical variance estimators are often not appropriate because the additional variability introduced by the E&I procedures is not accounted for or because imputed data is treated as if it was properly observed. Therefore special care and special methods are needed to estimate correctly the variance of estimators based on E&I treated data or of robust estimators.

Chapter 6

Documenting Editing and Imputation

6.1 Introduction

This chapter aims at providing the reader with general recommendations which should be considered independently of the characteristics of the E&I process and the survey where it is integrated in.

The aim of documentation is to inform users, survey managers, respondents and also E&I specialists about the data quality, the performance of the E&I process, its design and the adopted strategy. The documentation of the E&I process can be divided in three types of documentation: methodological documentation, reporting, and archiving. These types will be discussed in more detail in the next sections.

Quality indicators can be calculated on any procedure used in E&I. Usually these indicators are simple to calculate but difficult to analyze, particularly if a large set of indicators is used.

A particular set of indicators are those on resources spent and duration. Possible resource indicators are those reported in Appendix D.1.

Which indicators are useful depends on the E&I process, on the information available, on the accounting system of the organization and on the particular purpose in mind.

Generally the quality indicators can be calculated overall or per variable/observation. Furthermore, weighted or unweighted versions of indicators can be calculated, when appropriate.

The indicators may be used as standard reporting devices which are routinely calculated. The very basic set for reporting to users and managers is discussed in Section 6.3. However, internal reporting usually needs a larger standard set. It is part of the design to decide on the set of indicators needed for the particular E&I process. If appropriate flags are set during the process, and data sets are archived properly, then an ex-post calculation of particular indicators is possible. Often a regularly calculated set of E&I indicators leads to the need of more in-depth analysis of the E&I process: this task usually needs the support of methodologists.

Usually the resources spent on documentation depend on a trade off between cost and usefulness. Detailed description by means of several statistical indicators may be used to enhance the strategy and the design of the process in repeated surveys (see also Section 2.2), and to ensure reproducibility and repeatability of the E&I process. Minimal documentation consisting in the information given in reporting to the users may be sufficient if the survey is not repeated or if the E&I process is simple.

Archiving is necessary for being able to respond to any questions concerning the E&I process in particular when secondary analysis is done or when the data is used as input for other statistical results. Moreover, data, programs and the corresponding meta-data have to be archived if an E&I process will be repeated or if new methods or whole E&I phases will be tested for a specific survey. Confidentiality aspects have to be taken into account in the design of archiving.

6.2 Methodological documentation

The methodological documentation of the E&I process describes the strategy and the flow of the process and the methods used for editing and imputation. The main users are experts in E&I, methodological staff and subject matter specialists.
The flow of the process can be described as explained in Section 2.2, so that each phase of the process and the milestones of strategic decisions are well defined. Indicators used to make a decision about the strategy should be discussed and evaluated in the methodological documentation.
The description of each phase should be accompanied by an analysis of the input data and the output data. This analysis covers at least descriptive statistics about the most important variables and statistical measurements of the impact due to the E&I phase, see Appendix D.3 and Appendix D.4.
The documentation of the procedures of the E&I process should have two aspects: i) the impact of the procedures on the actual data, and ii) the performance of the methods in for the type of data at hand. The impact of an E&I procedure is measured during the E&I process using indicators based on the input and the output of the procedure. Compared with this, the performance of an E&I method is based on the comparison of the "true" data to the data by the procedure under consideration (as discussed in Section 2.3 on testing E&I methods). The impact shows the effects of the procedure on the true data, whereas testing allows to evaluate the accuracy of the E&I methods in the specific survey context. It is usually not possible to evaluate the performance on real data during the E&I process because the true values are not available. Therefore, the impact of each applied E&I procedure and the performance of its underlying method should both be documented. This kind of documentation is particularly useful for the continuous improvement of the current system and for the development of new, but similar processes.

6.3 Reporting

Reporting on the E&I process should inform the user about the main aspects of the E&I process and data quality before and after the E&I process. This kind of documentation is the minimal documentation and is sometimes added to the general description of the survey or added to the main statistical publication. Some accuracy indicators used for reporting on the E&I process are listed in Working group on quality (2005). They are:

1. unweighted unit response rate (D.38);
2. weighted unit response rate (D.39);
3. unweighted item response rate (D.19);
4. weighted item response rate (D.20);
5. weighted item response ratio (D.21);
6. unweighted imputation rate (D.23);
7. weighted imputation ratio (D.31).

Unit response indicators can be included in the set of measures for reporting in case unit nonresponses are treated at the E&I stage. The interpretation of the related indicators is highly dependent on the metadata provided. Useful metadata are:

1. Specific definitions for various categories of units (respondents, in-scope units, etc.). Note also that, according to the Statistics Canada Quality Guidelines (Statistics Canada, 2003) for each unit the reasons for nonresponse must be recorded.

2. Precise description of the weighting method including auxiliary variables as appropriate.

3. Description of the data collection technique.

4. Information on whether substitution was adopted or not.

5. Precise description of the imputation process (including methods for re-weighting).

As for indicators relating to **item response**, relevant metadata are similar to those of of unit non-response plus a self-assessment on the response burden related to the contents of the questionnaire (i.e. presence of sensitive questions, length, etc.). Indicators for item response are usually computed for the key variables only.

Additionally to these indicators, it should be stated whether the imputation has been taken into account in the variance estimation or not (see Section 5.2). Furthermore, the impact of E&I on the main statistical results of the survey should be evaluated and documented (see Appendix D.3.3 and Appendix D.4.2).Finally, references to more detailed documentation should be included in reporting.

6.4 Archiving

The archiving of the E&I process consists of storing the most important information and data of this process to guarantee reproducibility[1] and repeatability[2] of the process and for further investigation of the data. Archiving is also needed for testing new methods and processes (see Section 2.3) and for developing new quality measurements. Moreover, archiving is necessary to ensure that a different E&I process can be built when new user needs emerge (e.g. for secondary analysis with different objectives than the main survey objectives).

Data input and output at each phase should be archived in particular to guarantee repeatability of the corresponding phase. Further, the programs used and their description, including guidelines for interactive treatment and the set of edit rules, as well as a description of the data flow should be stored to guarantee reproducibility.

Often the data flow, and in particular the flow of each survey unit, cannot be completely archived because of the huge amount of data and its complexity or because of confidentiality constraints. Therefore, the needs and the legal rules should be taken into account at the design stage of the E&I process.

6.5 Recommendations

1. The documentation should be established appropriately for different target readers and stakeholders: producers, users, survey managers, methodologists.

2. Resources and time for documenting E&I activities should be allocated at the design stage of the E&I process.

3. The minimal set of indicators used for reporting should always be published together with the data.

[1] A process is reproducible if it is possible to restart the implementation of the process from scratch and arrive at the same result.

[2] A process is repeatable if given the implementation it is possible to apply it with the same parameters and input and arrive at the same result.

Chapter 7

Summary and General Recommendations

7.1 Introduction

This handbook is meant to disseminate a systematic approach for designing and managing E&I processes. Although the focus of this RPM is on cross-sectional business surveys, there are many elements which can be applied in longitudinal business surveys and in other types of surveys, like household surveys.

In Chapters 1-6, the E&I process has been discussed with respect to the different phases of its life-cycle: design, testing, tuning and monitoring, alternative approaches for error detection and error treatment, impact on subsequent analysis and estimation, documentation. The handbook represents a complete guide for survey managers and editing designers, supporting them in selecting the most appropriate approaches for dealing with the different types of non-sampling errors under each particular survey context. It also provides a guide for the design, the management and the analysis of an E&I process.

As far as the design is concerned, E&I processes are viewed from two different but complementary perspectives. The first perspective (see Section 2.2.1) concentrates on the design of the process: the E&I process is the result of the integration of different methodologies, dealing each with a specific data problem, and structured based on the survey objectives and on the data characteristics. Under this perspective, the process flow is dominated by the need of balancing between costs, time, and quality requirements. The second perspective (see Section 2.2.2) focuses on the management and control of the E&I process: the E&I strategy is analyzed as a process subject to parametrization, where some basic conditions are to be met to guarantee the efficient management and control of each phase of the E&I process. In this perspective, once the process has been designed, it has to be implemented with a parameterization, tested and tuned and get productive.

As for testing, tuning and monitoring an E&I process (see Section 2.3), the discussion provided in the handbook aims at highlighting the relevance of these activities for the continuous improvement of process efficacy and data quality.

An issue which is assuming an increasing relevance at statistical agencies concerns the evaluation of the impact of an E&I process on statistical analysis and on estimation performed on edited and imputed data. The discussion provided in Chapter 5 should contribute to the development of a common framework at NSIs in the area of data analysis and estimation based on edited and imputed data.

As for the documentation of E&I activities and of effects of E&I on data (see Chapter 6), the aim of the RPM is to support the survey managers in satisfying the increasing demand (from users, data producers, respondents, and E&I specialists) of information about the performance of the E&I process, its design and the adopted strategy, its costs and timeliness, and the data quality.

The declared focus of the RPM (see Section 1.1) is on the E&I activities performed at the post data capture stage, when data are available in electronic format. However, in the handbook the importance of an integrated approach to the design and the management of E&I activities performed at the different stages of the survey process is underlined. Furthermore, given the impact of the E&I activities on other phases of the statistical survey process, the need for a multidisciplinary approach to the E&I problem is recommended.
Taking into account this summary, in next section general recommendations for an efficient design and management of and E&I strategy for cross-sectional business surveys are provided.

7.2 General recommendations

1. E&I is cost and time consuming. Given the resources and time constraints, the trade-off between data accuracy and resources spent should be optimized. Resources should be focused on the most relevant errors. Selective editing directs relevant errors to the best available treatment, which is often interactive and may include call-backs to respondents. Often the best treatment is also more expensive and should not be applied to less important errors, or to errors that can be resolved with less expensive automatic treatment. Appropriate measures, like quality and impact indicators, should be taken into account to allow for this optimization. More in general, facilities should be provided to evaluate the performance of the E&I process to collect information relating to its impact on survey data and on survey outcomes (e.g. production of diagnostics, quality indicators and impact measures).

2. An important role of E&I derives from its ability to provide information about non-sampling error characteristics and their possible sources. This information can be used on one hand to provide quality measures for the current survey, on the other hand to suggest future improvements of the survey process. Appropriate time and resources should be allocated for a systematic analysis of information produced by the E&I process, supporting a continuous learning cycle.

3. Audit trails should be established to allow for a detailed analysis of the data processing, and to identify possible problem areas.

4. The links and the information flows between the E&I activities carried out at the different stages of the survey process are to be considered, integrated and harmonized. An appropriate resource allocation has to be planned at the design stage also taking into account information coming from previous surveys and the knowledge of the main survey problems.

5. An efficient design and analysis of edits used for E&I should be guaranteed. In general, processes should be designed which allow for the systematic specification, management and update of edits. Tools for the analysis of internal consistency of edits should be incorporated in the E&I process.

6. The design and tuning of E&I strategies in general, and the interactive treatment in particular, should avoid over-editing. Over-editing occurs when the resources and the time spent for editing are not justified by the resulting improvements in data quality.

7. Existing generalized software for E&I should be used when possible. Ad-hoc software implementing specific algorithms for E&I should not be developed if these algorithms are available in already accessible tools or programs.

8. Documentation about costs and time for E&I should be provided on a sufficiently detailed basis to allow for the process optimization.

9. Information on the quality of the data (error types and the methods used to deal with them) should be provided to users to allow for the proper use of the data in subsequent analysis.

10. People working on E&I and related aspects should be properly trained from both a theoretical and an operational point of view. They should have a sound knowledge of the competitive methods they can adopt, and they should be aware about the links between the E&I activities and the other parts of the survey process. Furthermore, they should be conscious about the importance of testing, monitoring and documenting E&I processes. To these purposes, appropriate training courses should be given periodically for survey managers/editing specialists to disseminate knowledge about E&I methodologies and on E&I in general.

11. A multidisciplinary approach to the design of E&I strategies should be adopted: survey designers, questionnaire designers, E&I specialists, computer scientist, methodologists and survey managers should co-operate.

12. For an actual harmonization of E&I, NSIs should promote new strategies for planning or reengineering E&I procedures, developing formal procedures to be internally followed to obtain approval to E&I processes, identifying centralized/expert staff in charge of formal, highly specialized and comprehensive evaluations. The above mentioned training (see point 10) could be facilitated if appropriate specific guidelines are developed concerning how to design, implement, manage and document E&I processes for statistical survey data.

Appendix A

Notation

Symbol	Meaning	Remarks
i, h	Index for unit i or h	preferably i
n	Number of units	
j, k	Index for variable j or k	preferably j
p	Number of variables	
y_i	Vector of values for unit i	vector of length p
y_j	Vector of raw values for variable j	vector of length n
$\hat{y}_j$	Vector of imputed values for variable j	vector of length n
y_j^*	Vector of true values for variable j	vector of length n
y_{ij}	Raw value of variable j at unit i	
$\hat{y}_{ij}$	Imputed value of variable j at unit i	
$\tilde{y}_{ij}$	Anticipated value of variable j at unit i	for score functions for selective editing
y_{ij}^*	True value of variable j at unit i	for evaluation studies
x_{ij}	Auxiliary variables	
w_i	(Sampling) weight of unit i	Usually $w_i > 0$
r_{ij}	Response indicator	$r_{ij} = 1$ for response
$\hat{r}_{ij}$	Response indicator after E&I	$\hat{r}_{ij} = 1$ for response
e_{il}	Error indicator for edit rule l	$e_{il} = 1$ for failure
o_{lj}	Involvement indicator for variable j in rule l	$o_{lj} = 1$ for involvement
f_{ij}	Failure indicator	$f_{ij} = 1 - r_{ij} \prod_l (1 - e_{il} o_{lj})$
b_{ij}	Structurally missing values flag	$b_{ij} = 1$ for structurally missing values
$\hat{b}_{ij}$	Structurally missing values flag after E&I	$\hat{b}_{ij} = 1$ for structurally missing values
u_i	Robustness weight	$u_i \in [0, 1]$
S	Sample	most often the net sample
M	Set for intensive review in selective editing	usually manual review
θ	Characteristic of the population, i.e. a function of the population values	E.g. the population total
$\hat{\theta}$	Estimator of θ, i.e. a function of the sample values	E.g. a Horvitz-Thompson estimator
$E(\hat{\theta})$	Expectation of estimator $\hat{\theta}$	Bias is $E(\hat{\theta}) - \theta$
$V(\hat{\theta})$	Variance of estimator $\hat{\theta}$	
$\hat{V}(\hat{\theta})$	Variance estimator for estimator $\hat{\theta}$	
$SF(y_{ij}, \tilde{y}_{ij})$	Score function for value y_{ij}	
$SC(y_{ij}, \hat{\theta})$	Sensitivity curve at y_{ij}	
$ISC(y_i^*, \hat{y}, \hat{\theta})$	Imputation sensitivity curve at y_i^*	Also with y_i instead of y_i^*

Appendix B

The state-of-the-art survey

B.1 Introduction

In developing this handbook, information on currently used approaches and practices has been collected from various sources such as existing manuals, best practices guides, congress proceedings and the general literature concerning data editing and imputation. In addition, as part of the EDIMBUS-project, a state-of-the-art survey has been carried out among all the European countries (including the three partner institutions: Istat Italy, CBS Netherlands and SFSO Switzerland) as well as some leading statistical agencies outside of Europe (Statistics Canada, the U.S. Bureau of the Census, and the ABS). The questionnaire used for this survey included a preliminary version of the glossary of this manual in order to explain the used terminology to the respondents.

A total of 70 questionnaires have been filled in by 21 different NSIs, 31 of these responses were obtained at the three partner NSIs and the 39 other responses came from 18 different external NSIs. In Section B.2 a few main results of this survey are highlighted. The full questionnaire, together with a summary of the results for each question is included in Section B.3.

B.2 Some main results

One of the objectives of the survey was to understand which approaches, and to which extent, are currently used in business surveys to deal with the different types of errors. Table B.2 contains the frequencies of respondents using the various classes of approaches considered in the questionnaire. As can be seen, a very high percentage of NSIs (86%) use Manual editing/follow-up in their surveys. Macroediting is also widely used (71%), while Selective editing and Graphical editing are adopted in a lower percentage of cases: 36% and 26% respectively. Deterministic checking rules for error localization are widely used as well (74%), while the Fellegi-Holt principle is used in only 9 surveys (13%).

As for imputation, Model-based approaches are more used (59%) than Donor-based techniques (24%). Deductive imputation is widely used as well (56%). Note also that 89% of surveys declare to make use of other sources and/or historical data in their E&I procedures.

Question 9 asks the respondent to indicate for which type of errors these methods are used. Although some of these results are what one would expect (macroediting is often used for outliers) some other results are suspicious (use of the Fellegi-Holt paradigm for systematic errors). This subdivision in error types appears to be difficult for respondents, which is generally due to somewhat unclear or ambiguous definitions and concepts in the glossary. This has led to enhancements in the description of error types and methods and improvements in the glossary.

Question (15): *Do you perform preliminary test of the E&I process in your survey?*, was answered negatively by more than 50% of the respondents. Concerning the reasons, 54% answered that there are not enough resources, 34% said that no suitable data are available, and 46% claim there is a lack of time. This result seems to confirm that there is a lack of preliminary evaluation and tuning of E&I procedures before they are used in survey production processes.

It is known that some statistical agencies (such as Statistics Canada, Statistics Netherlands, Census Bureau), have developed high-level generalized software for the E&I of business survey data. In this respect, among the analyzed surveys, 28 ones (40%) declare to make use of some type of generalized software (question 13). However, in most cases this software consists of ad-hoc SAS programs. Actual generalized software specific for E&I tasks adopted at the responding statistical agencies are Banff (one Canadian and two Italian surveys), SLICE (Netherlands) SCIA (used in two Italian surveys to deal with qualitative information) and two different specialized tools developed by Eurostat (Poland and Italy).

It is well known in the literature that E&I activities may account for a considerable amount of survey resources (human, budget, time). The state-of-the-art-survey (question 17) confirms this fact, the results suggest that in the survey manager's perception the amount of resources spent for E&I is very high, 33 of the 64 respondents claim to spend more than 50% of their overall resources on data E&I. However, since it is not easy to define and measure the concept "amount of resources" and it is difficult to separate specific E&I activities as defined in this manual from other data processing activities, this result should be interpreted with some caution.

Table B.1: Frequency of use of E&I approaches

Method	Frequency	% of respondents
Interactive treatment / follow-up	60	86
Selective editing	25	36
Macroediting	50	71
Graphical editing	18	26
Deterministic checking rules	52	74
Minimum change error localization (Fellegi-Holt principle)	9	13
Deductive imputation	39	56
Use of other sources and/or historical data	62	89
Model-based imputation (e.g. regression, ratios, mean)	41	59
Hot-deck imputation (e.g. nearest neighbor, random donor)	17	24
Robust estimation / re-weighting	18	26

The problem of standardization has inspired two questions (18.1, 18.2): *Do you have manuals or guidelines, developed at your institute, describing recommended Edit and Imputation processes or methods in general?* and *Do you use manuals or guidelines, developed at other institutes, describing recommended Edit and Imputation processes or methods in general?* Of the 62 respondents to these questions 23 answered negatively on both questions. Of the affirmative responses many refer to Eurostat manuals, UN/ECE documents, or "non-standard" methodological documents developed at the statistical agencies. From these results it is evident that there is a definite need for developing and disseminating standard manuals on E&I at statistical agencies and/or at European level, which is the aim of the EDIMBUS project.

The purpose of the question (19): *In your institution, do you have an established procedure for obtaining approval for the E&I strategy of a survey?* was to investigate the problem of the standardization of at least the high-level organization and control of E&I processes at statistical agencies. Only 30% of the respondents declare to have such procedure, which shows that there is much room for improvement of standardization of processes within NSIs.

B.3 The survey

Part 1, Type of your survey

1. Full name of the survey

2. Type of survey

1. Census	10
2. Random sampling	37
3. Register-based	10
4. Other	19

3. Number of records and variables

1. Number of records (net)	18150 (median)
2. Number of variables	38 (median)

4. Survey period

1. Decennial	2
2. Annual	33
3. Quarterly	4
4. Monthly	24
5. Other	7

If available, please send us a general documentation of the methodology for E&I of the survey (in particular published methodology reports).

Part 2, Edit and Imputation process in your survey

5. Please describe in a few words the E&I process of your survey

6. Use of Computer Aided Interviews or data verification at the data entry/data capturing stage (e.g. manual entry, CATI/CAPI, electronic questionnaires, and so on)

1. Yes	43
2. No	26
Missing	1

7. In your Editing and Imputation process do you use information from other sources?

1. No	19		
2. Yes	51	1. At aggregate level (e.g. totals, means, distributions)	23
		2. At unit level	61

Edit and Imputation methods usually deal, sequentially, with different kinds of errors. Often, the following types of errors are distinguished: Systematic errors, Outliers, Influential errors, Random errors, Item non-response, Residual errors.

8. Which types of errors are identified with your Editing methods?

Systematic error	37
Outliers	69
Influential errors	50
Random errors	38
Item non-response	54
Residual errors	34
Other (please specify)	12

Often used methods for Editing and Imputation of various kinds of errors include the ones in the rows of the table below.

9. For each of these methods could you indicate if you use it and so, for which types of errors you use it? (multiple responses per method allowed).

	Used for error type						
Method	systematic error	outliers	influential errors	random errors	item non-response	residual errors	other
Interactive treatment / follow-up	30	49	37	20	39	24	11
Selective editing	4	20	17	4	14	4	3
Macroediting	8	43	26	6	5	24	4
Graphical editing	3	12	7	2	0	3	3
Deterministic checking rules	21	34	22	26	23	9	21
Minimum change error localization (Fellegi-Holt principle)	2	3	0	9	6	0	1
Deductive imputation	10	11	7	10	28	6	5
Use of other sources and/or historical data	11	40	31	16	36	12	12
Model-based imputation (e.g. regression, ratios, mean)	1	12	8	8	35	5	6
Hot-deck imputation (e.g. nearest neighbor, random donor)	2	4	4	10	12	2	4
Robust estimation / re-weighting	0	7	2	4	4	3	9
Other (please specify)	1	1	1	0	2	0	0

10. Do you think that your methods have particular advantages? (The question refers to the methods you mentioned in question 8. Multiple responses per method allowed).

	Advantages				
Method	quality of results	simplicity	low costs	timeliness	other
Interactive treatment / follow-up	53	28	7	7	1
Selective editing	15	8	18	16	0
Macroediting	35	19	20	24	0
Graphical editing	3	14	10	10	0
Deterministic checking rules	36	37	22	24	0
Minimum change error localization (Fellegi-Holt principle)	5	3	6	5	0
Deductive imputation	18	25	27	16	0
Use of other sources and/or historical data	37	22	35	21	2
Model-based imputation (e.g. regression, ratios, mean)	23	18	19	22	1
Hot-deck imputation (e.g. nearest neighbor, random donor)	13	7	8	6	0
Robust estimation / re-weighting	12	4	5	4	0
Other (please specify)	3	1	1	1	0

11. Do you think that your methods have particular disadvantages? (The question refers to the methods you mentioned in question 8. Multiple responses per method allowed).

	Disadvantages/limitations				
Method	quality of results	complexity	costs	timeliness	other
Interactive treatment / follow-up	5	4	43	36	6
Selective editing	9	2	3	4	0
Macroediting	8	10	3	12	1
Graphical editing	3	0	2	0	1
Deterministic checking rules	5	6	7	5	2
Minimum change error localization (Fellegi-Holt principle)	0	1	1	2	1
Deductive imputation	6	8	1	2	0
Use of other sources and/or historical data	11	15	7	14	2
Model-based imputation (e.g. regression, ratios, mean)	8	13	0	2	2
Hot-deck imputation (e.g. nearest neighbor, random donor)	4	5	1	4	0
Robust estimation / re-weighting	4	7	0	2	0
Other (please specify)	0	1	1	2	0

12. Do you have documentation of the Edit and Imputation methods used for your particular survey?

1. No	19
2. Yes	51

If Yes, it would be very helpful if you could send a copy to the return address of this questionnaire.

13. Do you use generalized software in the Editing and Imputation process of your survey? (Software that can be used for a range of applications rather than for a specific survey)

1. No	42
2. Yes	28

14. How do you manage the Edit and Imputation process in your survey?

14.1 Information used in setting up an E&I system:

	Yes	No
1. Strategy for the Edit and Imputation process	55	15
2. Indicators of the performance of the E&I process (e.g. remaining errors)	30	40
3. Preliminary tests of the E&I process	37	33
4. Experience of the previous surveys	65	5
5. Guidelines, recommended practices	50	20
6. Other, please specify:	1	68

14.2 In case you detect a problem during your E+I process. What management actions do you take?:

	Yes	No
1. Redirection of the data flow	23	46
2. Reallocation of available resources	22	47
3. Repetition of data processing with adapted input parameters	42	26
4. Change of the design (e.g. add new procedures)	46	23
5. Increase the resources of the E&I process	12	57
6. Other, please specify:	0	69
Missing	1	

15. Do you perform systematic preliminary tests of the editing and imputation process in your survey?

1. No	35	1. Lack of time	16
		2. Not enough resources	19
		3. No suitable data available	12
		4. Other	2
2. Yes	33		
Missing	2		

16. Do you document the results of your editing and imputation process?

1. No	15			
2. Yes	55	1. Computation of indicators	39	
		1. for the main variables		23
		2. for all the variables		16
		2. Methodological report	25	
		3. Technical report	35	
		4. Quality Report for Eurostat	23	
		5. Other:	6	

17. Could you range the used resources for the Editing and Imputation process in respect to the workload of the whole survey?

<10%	0
10% - 20%	5
20% - 30%	6
30% - 40%	9
40% - 50%	11
50% - 60%	7
60% - 70%	11
70% - 80%	6
80% - 90%	6
>90%	3
Missing	6

Part 3, General aspects of E&I processes in your Institution

18.1. Do you have manuals or guidelines, developed at your institute, describing recommended Edit and Imputation processes or methods in general?

1. No	31
2. Yes	31
Missing	8

If Yes, it would be very helpful if you could send a copy to the return address of this questionnaire.

18.2. Do you use manuals or guidelines, developed at other institutes, describing recommended Edit and Imputation processes or methods in general?

1. No	44
2. Yes	19
Missing	7

If Yes, could you please give the references:

19. In your institution, do you have an established procedure for obtaining approval for the E&I strategy of a survey?

1. No	42
2. Yes	18
Missing	10

19.1 If *Yes*, do you have to submit the following documents concerning the E&I strategy of a particular survey?

	Yes	No
1. Design of the E&I strategy of the survey	17	11
2. Changes of the design of the E&I strategy	15	13
3. Assessment of the E&I strategy	4	24
4. Documentation of the E&I strategy	16	12
5. Other documents, please specify:	2	26
Missing	42	

Appendix C

Methodological Details

C.1 Random errors

C.1.1 The Fellegi-Holt methodology

The (generalized) Fellegi-Holt paradigm for localizing errors is given by: the data of a record should be made to satisfy all edits by changing the values of the variables with the smallest possible sum of reliability weights. In mathematical terms, this can be formulated as follows. Let p be the number of variables. For each observation or record $(y_{i1}, ..., y_{ip})$, we wish to determine, or more precisely: wish to ensure the existence of, a synthetic record $(\hat{y}_{i1}, ..., \hat{y}_{ip})$ such that the synthetic record satisfies all edits, none of the $\hat{y}_{ij}$ $(j = 1, ..., p)$ is missing, and

$$\sum_{j=1}^{p} w_j \delta\left(y_{ij}, \hat{y}_{ij}\right) \tag{C.1}$$

is minimized, where $\delta\left(y_{ij}, \hat{y}_{ij}\right)$ equals 1 if y_{ij} is missing or differs from $\hat{y}_j$ and 0 otherwise, and w_j is the so-called reliability weight of the j-th variable $(j = 1, ..., p)$.

A reliability weight of a variable expresses how reliable one considers the values of this variable to be. A high reliability weight corresponds to a variable of which the values are considered trustworthy, a low reliability weight to a variable of which the values are considered not so trustworthy. Reliability weights are non-negative.

The variables of which the values in the synthetic record differ from the original values, together with the variables of which the values were originally missing, form an optimal solution to the error localization problem. In the original Fellegi-Holt paradigm all reliability weights were set to one in formula C.1. A solution to the error localization problem is basically just a list of all variables that need to be changed. There may be several optimal solutions to a specific instance of the error localization problem.

Formulated in the above way, error localization based on the (generalized) Fellegi-Holt paradigm becomes a mathematical programming problem. In this mathematical programming problem the edits define the restrictions, while the objective function is given by formula C.1.

The resulting mathematical programming problem can be solved by techniques known from the operations research literature. Some of these techniques have been adapted especially for the error localization in order to improve their efficiency. For a recent overview of algorithms for solving the error localization problem automatically based on the Fellegi-Holt paradigm we refer to De Waal and Coutinho (2005).

C.2 Missing Values

C.2.1 Nonresponse mechanism

Terms related to the mechanism which guides the response behavior are "missing completely at random" (MCAR), "missing at random" (MAR) and "not missing at random" (NMAR).

MCAR. The missing data are said to be missing completely at random if the fact that a certain item is missing does not depend on the missing nor the observed data. This means that MCAR is like simple random sampling. No missing data adjustment is therefore needed, but the assumption of a MCAR nonresponse mechanism is quite unrealistic.

MAR. In this case the nonresponse mechanism is random conditional on the observed covariates. Therefore, imputation methods using the relevant observed covariates as auxiliary information can reduce nonresponse bias. This means that MAR is like simple random sampling within the classes determined by the relevant covariates.

NMAR. If the nonresponse mechanism depends on unobserved data, such as variables outside the survey or the target variable itself, it is said to be NMAR. In this instance the nonresponse bias cannot be reduced by imputation as the respondents and the nonrespondents differ from each other, even after conditioning on the covariates. For example, if - in a survey on income - the households with relatively high and/or low income have more nonresponse, there may be no way to reduce the resulting bias.

Now a more formal description will be given. Let y_i denote the vector of responses of respondent i on k questions. Then denote the observed part of variable y_i by y_i^o and the missing part by y_i^m. The binary vector r_i is the response indicator, where $r_{ij} = 1$ means that respondent i answered question j; otherwise $r_{ij} = 0$. Let x_i be the vector of auxiliary information, let z_i be a vector with unobserved variables outside the survey and let ξ be a nuisance parameter vector. The above mentioned nonresponse mechanisms may then be formalized as follows:

$$MCAR: \quad P[r_i|y_i^o, y_i^m, x_i, z_i, \xi] = P[r_i|\xi] \tag{C.2}$$

$$MAR: \quad P[r_i|y_i^o, y_i^m, x_i, z_i, \xi] = P[r_i|y_i^o, x_i, \xi] \tag{C.3}$$

The nonresponse mechanism is NMAR if conditions C.2 and C.3 do not hold, that is if $P[r_i|y_i^o, y_i^m, x_i, z_i, \xi]$ cannot be simplified and the nonresponse mechanism remains dependent on the missing data y_i^m and/or z_i. See Rubin (1987) and Schafer (2000) for more details on the nonresponse mechanism.

Another important issue for nonresponse mechanisms is the point of ignorability. A nonresponse mechanism is ignorable if the missing data are MCAR or MAR and if the nuisance parameter ξ is distinct from the parameter θ that is to be estimated. If the nonresponse mechanism is NMAR, it is nonignorable. In this case there is a systematic difference between respondents and nonrespondents, even after conditioning on auxiliary information. Refer to Rubin (1987) for a formal description of ignorability and nonignorabilty. If the nonresponse mechanism is ignorable, inference on the parameter of interest, θ, can be based solely on the observed data.

C.3 Outliers

C.3.1 Weighted quantiles

The univariate weighted median or any univariate weighted quantile of a variable $y_j = y_{1j} \ldots, y_{nj}$ can be calculated from the weighted empirical distribution function for the sample S. The weighted quantile $q_\alpha(y_j, w)$ of the observations $y_{1j}, \ldots, y_{nj}$ for a fixed $0 \leq \alpha \leq 1$ and for the vector of weights w is calculated as follows:

1. Sort the observations in S according to variable y_j and denote the indices of the sorted observations by $i = 1, \ldots, n$.

2. Calculate the partial sums of weights $s_i = \sum_{h=1}^{i} w_h$.

3. Determine $l = \min\{i : s_i \geq \alpha s_n\}$ and $u = \min\{i : s_i > \alpha s_n\}$.

4. The α-quantile is then
$$q_\alpha(y_j, w) = \frac{w_l y_l + w_u y_u}{w_l + w_u}.$$

The unweighted quantile, i.e. $q_\alpha(y_j, w)$ for w a vector of 1s, is denoted $q_\alpha(y_j, 1)$.

C.3.2 MAD-rule

The weighted median $\text{med}(y_j, w)$ is the weighted quantile $q_\alpha(y_j, w)$ for $\alpha = 0.5$. Taking absolute residuals from the weighted median and calculating their weighted median again we obtain the (scaled) weighted Median Absolute Deviation (mad):

$$\text{mad}(y_j, w) = 1.4826 \ \text{med}(|y_j - \ \text{med}(y_j, w)|, w). \tag{C.4}$$

The "mad-rule" for outlier detection corresponds to the robustness weights

$$u_{ij} = \min\left(1, \frac{c \ \text{mad}(y_j, w)}{|y_{ij} - \ \text{med}(y_j, w)|}\right). \tag{C.5}$$

The tuning constant c has to be chosen by the statistician (see Section 3.5). The particular form of the robustness weight of the "mad-rule" corresponds to a Huber type one-step M-estimator.
Note that u_{ij} depends on the variable y_j and, in fact, we may calculate p robustness weights u_{ij}. These univariate robustness weights may be combined to a robustness weight, e.g. by $u_i = \min(u_{i1}, \ldots, u_{ip})$. However, such a robustness weight which is derived from univariate robustness weights does not protect weighted means against multivariate outliers.
Often the mad has a high variance and it is 0 if more than half of the (weighted) observations are concentrated at the median. Then instead of the mad we may prefer the inter-quartile range or even a larger quantile like the inter-decile-range.
The weighted inter-quartile range is $\text{iqr}(y_j, w) = 1.4826 \ |q_{0.75}(y_j, w) - q_{0.25}(y_j, w)|$ and similar for other quantiles.
The "iqr-rule" for outlier detection is

$$u_{ij} = \min\left(1, \frac{c \ \text{iqr}(y_j, w)}{|y_{ij} - \ \text{med}(y_j, w)|}\right). \tag{C.6}$$

C.3.3 Winsorized mean and trimmed mean

The winsorized mean is based on weighted quantiles (C.3.1) of a variable y_j: For a tuning constant α with $0.5 \leq \alpha \leq 1$ declare observations smaller than $q_{\alpha/2}(y_j, w)$ or larger than $q_{1-\alpha/2}$ as outliers. This is the implicit detection rule of the winsorized mean and the trimmed mean. The winsorized mean then imputes $q_{\alpha/2}(y_j, w)$ for the small outliers and $q_{1-\alpha/2}(y_j, w)$ for the large outliers. Note that for positive variables, i.e. if $y_{ij} > 0 \ \forall i$, the winsorized mean can be described by robustness weights. Simply set $u_{ij} = q_{\alpha/2}(y_j, w)/y_{ij}$ if $y_{ij} < q_{\alpha/2}(y_j, w)$ and $u_i = q_{1-\alpha/2}(y_j, w)/y_{ij}$ if $y_{ij} > q_{1-\alpha/2}(y_j, w)$. The trimmed mean is more extreme, setting u_{ij} to zero outside these quantiles.

C.3.4 Hidiroglou-Berthelot

The Hidiroglou-Berthelot method (Hidiroglou and Berthelot, 1986) is used to detect errors in repeated surveys.
Let $y_{ij,1}$ and $y_{ij,2}$ be measures of a variable y_j on the same unit i at two time points $t = 1$ and $t = 2$. The ratio of these values is called trend:

$$t_{ij} = \frac{y_{ij,2}}{y_{ij,1}}. \tag{C.7}$$

Using the median of these trends med$(t_j) = q_{0.5}(t_j, 1)$ (usually the unweighted median is used) a transformation of the trends ensures more symmetry of the tails of the trend distribution:

$$s_{ij} = \begin{cases} 1 - \text{med}(t_j)/t_{ij}, & 0 < t_{ij} \leq \text{med}(t_j); \\ t_{ij}/\text{med}(t_j) - 1, & t_{ij} \geq \text{med}(t_j). \end{cases} \tag{C.8}$$

The magnitude of the data is combined with these scores s_{ij} to yield the effects

$$E_{ij} = s_{ij}[\max(y_{ij,1}, y_{ij,2})]^{c_1}, \tag{C.9}$$

where $0 \leq c_1 \leq 1$ is a tuning parameter. A scale for the right tail of these effects is defined by

$$d_{j,right} = \max(|q_{0.5}(E_{ij}, 1) - q_{0.75}(E_{ij}, 1)|,\ c_2\,|q_{0.5}(E_{ij}, 1)|), \tag{C.10}$$

and similarly a scale of the left tail replacing $q_{0.75}(E_{ij}, 1)$ by $q_{0.25}(E_{ij}, 1)$. The second part of the maximum condition with tuning constant c_2 helps to avoid too small scales. Observations outside the interval

$$[q_{0.5}(E_{ij}, 1) - c_3 d_{j,left},\ q_{0.5}(E_{ij}, 1) + c_3 d_{j,right}] \tag{C.11}$$

are declared outliers. The constant c_3 allows to adjust the width of the acceptance interval.

C.3.5 One-step robustified ratio estimator

The one-step robustified ratio estimator of Hulliger (1999) with an auxiliary variable $x_i > 0$ is based on a preliminary estimate of the ratio by $\hat{\beta}_0 =$ med$(y_j, w)/$ med(x, w) and the standardized absolute residuals $a_{ij} = |y_{ij} - \hat{\beta}_0 x_i)|/\sqrt{x_i}$. Let the median of the absolute residuals $\hat{\sigma}_r =$ med(a_j, w). Then robustness weights are defined as

$$u_{ij} = \begin{cases} 1, & a_{ij} \leq c\,\hat{\sigma}; \\ c\,\hat{\sigma}/a_{ij}, & a_{ij} > c\,\hat{\sigma}. \end{cases} \tag{C.12}$$

Again, c is a tuning constant to be chosen. The one-step robustified ratio estimator is the ratio of sampling and robustness weighted totals, i.e.

$$\hat{\beta}_j = \frac{\sum_i w_i u_i y_{ij}}{\sum_i w_i u_i x_i}. \tag{C.13}$$

C.4 Imputation

C.4.1 Regression imputation

The imputed value is obtained by a multiple regression model of the form

$$\hat{y}_i = \hat{\beta}_0 + x_{1,i}\hat{\beta}_1 + \ldots + x_{k,i}\hat{\beta}_k \tag{C.14}$$

with $x_{1,i}$ to $x_{k,i}$ the values of auxiliary variables that are observed for respondent i and $\hat{\beta}_0$ to $\hat{\beta}_k$ (weighted) least squares estimates of the regression coefficients based on the sub-sample S_r of responding units. Sometimes it is convenient to model jointly all the variables of interest through a multivariate parametric distribution and derive the regression models corresponding to the different missing patterns from the conditional distributions of this model. Since the parameters of the regression models needed for imputation can be expressed in terms of the joint distribution parameters, only the latter have to be estimated. One of the most popular method for the estimation of the joint data distribution is the **Expectation Maximization** (EM) algorithm (Dempster et al., 1977). It allows the maximum likelihood estimation of the model parameters from incomplete data by iteratively applying, until convergence, two steps (E-step and M-step), that in some cases (e.g. in the normal case) can be easily implemented through simple computer programs (Schafer, 2000). If one is interested in preservation of distribution moments (higher than the mean), a residual can be randomly added. In this way the method becomes stochastic.

C.4.2 Ratio Imputation

This method is a very often used imputation model for business statistics. The imputed value, can be written as

$$\hat{y}_i = \hat{R}x_i \tag{C.15}$$

with x_i the value of an auxiliary variable and $\hat{R} = \frac{\sum_{i \in S_r} y_i}{\sum_{i \in S_r} x_i}$ the ratio of the mean of y to the mean of x for the responding units. The ratio imputation is a special case of regression imputation with only one auxiliary variable x and $\hat{\beta}_0 = 0$ if the variance of the model is assumed to be proportional to x.

C.4.3 Mean imputation

The respondent mean imputation imputes the mean of the respondents $\bar{y}_r$:

$$\hat{y}_i = \bar{y}_r \tag{C.16}$$

This method is a special case of a ratio imputation with $x_i = 1, \quad \forall i \in S_r$.

C.4.4 Random donor imputation

In random donor imputation donors are chosen completely at random (with replacement) from the donor pool, D, and the selected donor's value is assigned to the recipient, that is, $\hat{y}_i = y_h$ for some $h \in D$ such that $P(\hat{y}_i = y_h) = \frac{1}{n_D}$ with n_D the number of units in D. This method is the simplest form of donor-based imputation. It does not necessarily yields the same imputed value given the sample if the imputation process is repeated hence it is a stochastic method. It is generally performed within "imputation classes" constructed using categorical auxiliary variable(s) whose values are observed for all units (random donor imputation within classes).

C.4.5 Nearest neighbor imputation

In the nearest neighbor imputation (NNI) the missing and inconsistent values for a unit i come from the responding unit closest to i with respect to a distance measure based on the values of a vector of observed auxiliary (or matching) variables. Let $x_i = (x_{i1}, ..., x_{iJ})$ be the value of J auxiliary variables for a unit i for which y_i is missing. If values of these variables are used to define imputation cells, the distance measure

$$d(i,h) = \begin{cases} 0 & if \quad i,h \text{ in same cell} \\ 1 & if \quad i,h \text{ in different cells} \end{cases}$$

yields the random donor imputation within classes. A common choice is the l_p distance: $d(i,h) = (\sum_{j=1}^{J} |x_i - x_h|^p)^{1/p}$ that, for example, when $p = 1$ is the Manhattan distance, when $p = 2$ is the Euclidean distance and when $p = \infty$ is the Maximum deviation: $d(i,h) = \max_k |x_{ik} - x_{hk}|$. Usually the auxiliary variables are standardized so they are all on the same scale. Another commonly used distance, that does not need scaling variables, is the Mahalanobis distance: $d(i,h) = (x_i - x_h)^T S_{xx}^{-1}(x_i - x_h)$, where S_{xx} is an estimate of the covariance matrix of x_i. If there are multiple nearest neighbors of i then the donor is randomly selected from them.

A method that can be considered a variant of the NNI is the **Predictive Mean Matching** (PMM) (Little, 1988). With this method, a model (typically the normal model with continuous numerical variables) is used only to define a suitable distance function. In particular for each missing pattern, the predictive means of the missing values conditional on the observed ones are computed based on the assumed model, and each nonrespondents is matched to the respondent with the closest predictive mean.

Appendix D

Indicators

The indicators and measurements listed in this section are based on the indicators described in Della Rocca et al. (2003) and in EUREDIT Project (2004b). The notations of Appendix A are used and enlarged in this paragraph.
Note that most of the listed indicators are until now rarely described and only a small subset is usually used. Therefore it is difficult to assess their utility in general. The list is by far not exhaustive, in particular weighted versions and versions for subsets of observations (e.g. observations with structurally missing values), subsets of variables and subsets of edit rules (e.g. query edits) are not listed explicitly.

Notation

We assume that there are n eligible observations in the sample under consideration and p variables may have missing or inconsistent values. Note that we may restrict the sample to a domain of interest for most of the indicators described here. This may be necessary to find error sources.

- r_{ij} is the actual response indicator for unit i and variable j: it takes the value 1 for response and 0 for nonresponse.

- $\hat{r}_{ij}$ is the response indicator for unit i and variable j after E&I: it takes the value 1 for response and 0 for nonresponse.

- e_{il} is the indicator variable related to the edit rule l and takes the value 1 if observation i fails the edit (detection of erroneous or suspicious values) and 0 if the edit is passed by observation i. Strictly speaking e_{il} is not defined if an observation has missing values in the variables involved in the edit rule. Often $e_{il} = 1$ is also set in these cases and we use this convention here.

- Often the detection of the erroneous variable values is performed using a binary vector of length p with all components corresponding to the involved variables in the edit equal 1 and all the others equal 0. Therefore, $o_{lj} = 1$ if variable j is involved in the edit rule l and 0 otherwise.

- The failure indicator f_{ij} is the flag which is 1 if variable j of unit i is detected to be erroneous and 0 otherwise.

- Structurally missing values are flagged with b_{ij}=0. In all other cases (observed values and genuine nonresponses), b_{ij}=1.

- Structurally missing values after E&I are flagged with $\hat{b}_{ij}$=0. $\hat{b}_{ij}$=1 when the value is either observed or imputed.

- w_i is the sampling weight of observation i.

- $I(\cdot)$ is an indicator function taking the value 1 if its argument is true and 0 otherwise. For example the expression $I(y_{ij} \neq y^*_{ij})$ yields the value 1 if the raw value differs from the true value. If the raw value is not observed we follow the convention that the expression yields 1.

- c_{SF} denotes the cut-off value of the score function SF.

The index j always varies from 1 to p, the index i varies from 1 to n, if not specified differently.
The following indicators usually can be calculated with and without weights. For example the unweighted item response rate for variable j is $\sum_i \hat{r}_{ij}/n$ (indicator D.19). Its weighted version is $\sum_i w_i\hat{r}_{ij}/\sum_i w_i$ (indicator D.20).
Instead of w_i we may also weight with the so called response fraction w_ix_i for an auxiliary variable x_i, e.g. for x_i the number of employees. This may describe the impact on totals better than weighting with the sampling weight w_i alone.
An additional issue concerns the use of the above defined flags in the definition of indicators depending on the context. In particular, in the testing phase, indicators are defined using the actual (true) values of flags, e.g. r_{ij} and b_{ij}. On the contrary, at the tuning and monitoring stage, where true values are not available, the only flags we can use for computing the indicators are those determined by the E&I process, i.e. $\hat{r}_{ij}$ and $\hat{b}_{ij}$. On the other hand, although true values of flags r_{ij} and b_{ij} are unknown, their product con be observed also on the raw data: in effect, $r_{ij}b_{ij} = 1$ only if the observed value of variable j for unit i is different from missing. For example, the modification rate (indicator D.28) for a variable j is defined as

$$\frac{\sum_i I(\hat{y}_{ij} \neq y_{ij})(r_{ij}b_{ij})\hat{b}_{ij}}{n}$$

The unit i contributes to this indicator if and only if the raw value for j is not missing ($r_{ij}b_{ij} = 1$) and it does not result as structurally missing after E&I ($\hat{b}_{ij} = 1$).

D.1 Resource indicators

These indicators depend heavily on the particular E&I process and the type of resources involved in it. Whether these indicators can be calculated and how detailed they can be calculated depends on the accounting system of the organization. For example if the E&I process is contracted out, the information may be quite different than from an in-house process. Usually the design and development phase is different from the actual application phase and different cost and resource indicators may be appropriate. Possible resource indicators follows here:

- **Duration overall:** Elapsed time between the start of the E&I process after data capturing and the release of all the data for analysis.

- **Duration breakdown:** Elapsed time for each phase of the E&I process.

- **Person-hours overall:** Overall person-hours for E&I.

- **Person-hours breakdown:** Breakdown of person-hours into skill levels, phases (interactive treatment), types of treatment etc.

- **Number of persons involved:** Usually a breakdown according to skills, functions, and phases is needed.

- **Number of call-backs:** The number of call-backs may have to be supplemented by more detailed information on the duration of call-backs.

- **Cost overall and breakdown:** Includes person-hours with tariffs plus equipment and computing costs etc. A breakdown into phases or procedures may be necessary.

D.2 Testing Editing and Imputation methods

In a testing phase we have the raw data y_{ij}, the true data y^*_{ij}, anticipated values $\tilde{y}_{ij}$ and the imputed data $\hat{y}_{ij}$ (see also the notation in Appendix A).

D.2.1 Efficiency of error detection

The following indicators refer to the error detection capability taking into account all the edits and all the data.

$$\text{Correct error detection rate} = \frac{\sum_{ij} f_{ij} I(y_{ij} \neq y^*_{ij})}{\sum_{ij} I(y_{ij} \neq y^*_{ij})} \tag{D.1}$$

$$\text{Correct influential errors detection rate} = \frac{\sum_{ij} I(SF(y_{ij}, \tilde{y}_{ij}) > c_{SF}) I(SF(y_{ij}, y^*_{ij}) > c_{SF})}{\sum_{ij} I(SF(y_{ij}, y^*_{ij}) > c_{SF})} \tag{D.2}$$

$$\text{Incorrect error detection rate} = \frac{\sum_{ij} f_{ij} I(y_{ij} = y^*_{ij})}{\sum_{ij} I(y_{ij} = y^*_{ij})} \tag{D.3}$$

The following indicator, called *Hit rate*, refers to the error detection capability of a single **query edit** l. It measures the proportion of the edit failures that point to true errors (i.e. that are due to true errors in at least one variable involved in the edit).

$$HR(l) = \frac{\sum_i e_{il} \left[1 - \prod_j (1 - I(y_{ij} \neq y^*_{ij}) o_{lj}) \right]}{\sum_i e_{il}} \tag{D.4}$$

D.2.2 Efficiency of error treatment

For each variable j the following indicators can be computed.

$$\text{Imputation error rate} = \frac{\sum_i I(\hat{y}_{ij} \neq y_{ij}) I(\hat{y}_{ij} \neq y^*_{ij})}{\sum_i I(\hat{y}_{ij} \neq y_{ij})} \tag{D.5}$$

$$\text{Weighted average imputation error} = \frac{\sum_i w_i (\hat{y}_{ij} - y^*_{ij})}{\sum_i w_i} \tag{D.6}$$

$$\text{Weighted relative average imputation error} = \frac{\sum_i w_i (\hat{y}_{ij} - y^*_{ij})}{\sum_i w_i y^*_{ij}} \tag{D.7}$$

$$\text{Weighted } \alpha\text{-relative imputation error} = \frac{\left(\frac{\sum_{i=1}^n w_i |\hat{y}_{ij} - y^*_{ij}|^\alpha}{\sum_{i=1}^n w_i} \right)^{1/\alpha}}{\sum_i w_i y^*_{ij} / \sum_i w_i} \tag{D.8}$$

$$\text{Weighted imputation error ratio} = \frac{\sum_i w_i I(\hat{y}_{ij} \neq y_{ij}) I(\hat{y}_{ij} \neq y^*_{ij}) \hat{y}_{ij}}{\sum_i w_i I(\hat{y}_{ij} \neq y_{ij}) \hat{y}_{ij}} \tag{D.9}$$

Indicators (D.6)-(D.9)can be computed on numerical variables only. All the previous indicators can be computed on the subset of observations containing at least one error, or on the subset of observation which have been changed by the E&I procedure, depending on the objectives of the analysis.

D.2.3 Effects on estimates

Indicators for measuring the impact on estimated statistics by a distance between the estimates computed on true and edited and imputed data are listed below.

Individual effect on estimates

The **imputation sensitivity** $ISC(y_i^*, \hat{y}, \hat{\theta})$ for the estimator $\hat{\theta}$ at the true value y_i^* is the difference between a statistic $\hat{\theta}$ evaluated after imputation and the same statistic but with the true value y_i^* in place of the imputed value $\hat{y}_i$. The imputation sensitivity of $\hat{\theta}$ to the observation i with respect to the true value is

$$ISC(y_i^*, \hat{y}, \hat{\theta}) \quad = \quad c\left(\hat{\theta}(\hat{y}_1, \ldots, \hat{y}_i, \ldots, \hat{y}_n) - \hat{\theta}(\hat{y}_1, \ldots, \hat{y}_{i-1}, y_i^*, \hat{y}_{i+1}, \ldots, \hat{y}_n)\right) \tag{D.10}$$

where c is a suitable standardization constant, often $c = 1$. Note that $ISC(y_i^*, \hat{y}, \hat{\theta})$ is different from the sensitivity of the estimator $\hat{\theta}$ to the observation i, $SC(y_i, \hat{\theta})$ (see Section 3.5).

Overall effect on estimates

The **estimation error due to E&I** for the estimate $\hat{\theta}$ is the difference of $\hat{\theta}$ after imputation at the sample and the same statistic with raw or true values. The first is noted $TIMP(y, \hat{y}, \hat{\theta})$ the second is noted $ERRIMP(y^*, \hat{y}, \hat{\theta})$;

$$ERRIMP(y^*, \hat{y}, \hat{\theta}) \quad = \quad \hat{\theta}(\hat{y}_1, \ldots, \hat{y}_i, \ldots, \hat{y}_n) - \hat{\theta}(y_1^*, \ldots, y_i^*, \ldots, y_n^*). \tag{D.11}$$

The **relative estimation error due to E&I** is

$$RERRIMP(y^*, \hat{y}, \hat{\theta}) \quad = \quad \frac{|\hat{\theta}(\hat{y}) - \hat{\theta}(y^*)|}{\hat{\theta}(y^*)} \times 100 \quad = \quad \frac{|ERRIMP(y^*, \hat{y}, \hat{\theta})|}{\hat{\theta}(y^*)} \times 100. \tag{D.12}$$

The analysis based on the survey data are usually multivariate. Therefore, measurements like correlations or robustified measurements of the covariance matrix may be used to evaluate the impact of imputation on the multivariate structure of the data.
Usually the target quantities $\theta(y_j^*)$ are totals, means and standard deviations.

Overall effect on distributions

The **Kolmogorov-Smirnov** index (KS) can be used to assess the difference between the marginal distributions of a variable y_j in the raw and edited and imputed data. The KS distance is defined as:

$$KS(F_{y_j^*}, F_{\hat{y}_j}) \quad = \quad max_t|F_{y_j^*}(t) - F_{\hat{y}_j}(t)| \tag{D.13}$$

where

$$F_{y_j^*}(t) \quad = \quad \frac{\sum_{i=1}^n w_i I(y_{ij}^* \leq t)}{\sum_{i=1}^n w_i}$$

and w_i are the sampling weights. The definition of $F_{\hat{y}_j}(t)$ is analogous.
Adapted versions of the KS were discussed in EUREDIT Project (2004b).

D.3 Tuning and monitoring Editing and Imputation methods

D.3.1 Impact of error detection

Number of observations:

$$\text{with at least one missing value} \quad = \quad \sum_i (1 - \prod_j \hat{r}_{ij}) \tag{D.14}$$

$$\text{failing at least one edit rule} \quad = \quad \sum_i (1 - \prod_l (1 - e_{il})) \tag{D.15}$$

The corresponding rates can be obtained by dividing the previous indicators by n.

$$\text{Failure rate for edit } l = \frac{\sum_i e_{il}}{n} \tag{D.16}$$

For each observation i*:*

$$\text{Missingness proportion} = \frac{\sum_j (1 - \hat{r}_{ij})}{p} \tag{D.17}$$

$$\text{Inconsistency proportion} = \frac{\sum_j f_{ij}\hat{r}_{ij}}{p} \tag{D.18}$$

For each variable j*:*

$$\text{Unweighted item response rate} = \frac{\sum_i \hat{r}_{ij}}{n} \tag{D.19}$$

$$\text{Weighted item response rate} = \frac{\sum_i w_i\hat{r}_{ij}}{\sum_i w_i} \tag{D.20}$$

$$\text{Weighted item response ratio} = \frac{\sum_i w_i\hat{r}_{ij}\hat{y}_{ij}}{\sum_i w_i\hat{y}_{ij}} \tag{D.21}$$

$$\text{Rate of inconsistent data} = \frac{\sum_i f_{ij}\hat{r}_{ij}}{n} \tag{D.22}$$

Indicators for variables should always be calculated with and without weights.

D.3.2 Impact of error treatment

For each observation i*:*

$$\text{Imputation proportion} = \frac{\sum_j I(\hat{y}_{ij} \neq y_{ij})}{p} \tag{D.23}$$

$$\text{Modification proportion} = \frac{\sum_j I(\hat{y}_{ij} \neq y_{ij})(r_{ij}b_{ij})\hat{b}_{ij}}{p} \tag{D.24}$$

$$\text{Net imputation proportion} = \frac{\sum_j I(\hat{y}_{ij} \neq y_{ij})(1 - r_{ij}b_{ij})\hat{b}_{ij}}{p} \tag{D.25}$$

$$\text{Cancellation proportion} = \frac{\sum_j I(\hat{y}_{ij} \neq y_{ij})(r_{ij}b_{ij})(1 - \hat{b}_{ij})}{p} \tag{D.26}$$

For each variable j*:*

$$\text{Unweighted imputation rate} = \frac{\sum_i I(\hat{y}_{ij} \neq y_{ij})}{n} \tag{D.27}$$

$$\text{Modification rate} = \frac{\sum_i I(\hat{y}_{ij} \neq y_{ij})(r_{ij}b_{ij})\hat{b}_{ij}}{n} \tag{D.28}$$

$$\text{Net imputation rate} = \frac{\sum_i I(\hat{y}_{ij} \neq y_{ij})(1 - r_{ij}b_{ij})\hat{b}_{ij}}{n} \tag{D.29}$$

$$\text{Cancellation rate} = \frac{\sum_i I(\hat{y}_{ij} \neq y_{ij})(r_{ij}b_{ij})(1 - \hat{b}_{ij})}{n} \tag{D.30}$$

$$\text{Weighted imputation ratio} = \frac{\sum_i w_iI(\hat{y}_{ij} \neq y_{ij})\hat{y}_{ij}}{\sum_i w_i\hat{y}_{ij}} \tag{D.31}$$

Indicators for variables should always be calculated with and without weights.

D.3.3 Impact of error detection and treatment

The indicators (D.6)-(D.8) and (D.11)-(D.13) in Section D.2 can be adapted to evaluate the impact of E&I during production by replacing the true values y_j^* and y_{ij}^* by the raw values y_j and y_{ij}. The following impact indicators are then obtained:

$$\text{Weighted average imputation impact} \quad = \quad \frac{\sum_i w_i(\hat{y}_{ij} - y_{ij})}{\sum_i w_i} \tag{D.32}$$

$$\text{Weighted relative average imputation impact} \quad = \quad \frac{\sum_i w_i(\hat{y}_{ij} - y_{ij})}{\sum_i w_i y_{ij}} \tag{D.33}$$

$$\text{Weighted } \alpha\text{-relative imputation impact} \quad = \quad \frac{\left(\frac{\sum_{i=1}^n w_i|\hat{y}_{ij} - y_{ij}|^\alpha}{\sum_{i=1}^n w_i}\right)^{1/\alpha}}{\sum_i w_i y_{ij} / \sum_i w_i} \tag{D.34}$$

As far as impact on estimates and distributions, the following indicators can be calculated:
Total impact of imputation

$$TIMP(y, \hat{y}, \hat{\theta}) \quad = \quad \hat{\theta}(\hat{y}_1, \ldots, \hat{y}_i, \ldots, \hat{y}_n) - \hat{\theta}(y_1, \ldots, y_i, \ldots, y_n). \tag{D.35}$$

Relative change

$$RC(y, \hat{y}, \hat{\theta}) \quad = \quad \frac{|\hat{\theta}(\hat{y}) - \hat{\theta}(y)|}{\hat{\theta}(y)} \times 100 \quad = \quad \frac{|TIMP(y, \hat{y}, \hat{\theta})|}{\hat{\theta}(y)} \times 100. \tag{D.36}$$

Impact on distribution

$$KS(F_{y_j}, F_{\hat{y}_j}) \quad = \quad max_t|F_{y_j}(t) - F_{\hat{y}_j}(t)|. \tag{D.37}$$

For tuning and monitoring purposes, also the **Hit rate** defined as in (D.4) can be used: in this context, it provides indications on the proportion of edit failures that originate a change in at least one variable involved in the failed query edit.

D.3.4 Analysis of the indicators

It is usually not possible to asses the editing performance during production because the 'truth' is not known and therefore the "true" errors are not known.
The indicators for observations are used together to detect the observations with the most erroneous and/or missing values. The units may be weighted for this comparison. Variables with the most problems are detected in a similar way. The indicators for the variables are often only calculated for the key variables.
It is hardly possible to fix acceptable values for the indicators beforehand without any experience, i.e. without values form preceding surveys or similar surveys. Therefore, in repeated surveys the level of the indicators may be defined in advance based on indicators of preceding surveys.
Indicators can also be computed for the different phases of the E&I process to evaluate the impact of each of them on survey results at micro or aggregate level.

D.4 Documenting Editing and Imputation

D.4.1 Eurostat Indicators

Eurostat has proposed a set of *"Standard quality indicators that can be used, from the point of view of the producers, for summarizing the quality of the statistics as reported according to the Standard*

Quality Report (Working group on quality, 2003) from various statistical domains. The objective is to have a limited set of indicators that can be used to measure and follow over time the quality of the data produced in the European Statistical System (ESS), i.e. data that most of the time are collected by the members of the ESS and then published by Eurostat" (Working group on quality, 2005).
Concerning non-sampling errors and the specific quality dimension of *Accuracy*, the following indicators are considered as indirect measures of the data quality for E&I:

1. **Item response rate**:

   Unweighted item response rate. (D.19)

   Weighted item response rate. (D.20)

   Weighted item response ratio. (D.21)

2. **Imputation rate and ratio**: to calculate these indicators it is necessary to retain a flag in the data set when a change due to imputation occurs or to store the original data set. They refers to a single variable.

   Unweighted imputation rate. (D.27)

   Weighted imputation ratio. (D.31)

3. **Unit response rate**:

$$\text{Unweighted unit response rate} = \frac{\sum_i \prod_j (1 - r_{ij})}{n + n'} \qquad \text{(D.38)}$$

$$\text{Weighted unit response rate} = \frac{\sum_i w_i \prod_j (1 - r_{ij})}{\sum_i^n w_i + \sum_i^{n'} w_i} \qquad \text{(D.39)}$$

where n' is the number of **eligible unknown units**, i.e. units of the sample for which eligibility is not known because it has not been possible to ascertain their eligibility during data collection.

D.4.2 Other Indicators

A subset of the indicators listed in Appendix D.3 can also be used to document the quality of the data and the effects of E&I at micro and/or aggregate level.
In particular, starting from indicators (D.19)-(D.22) and (D.23)-(D.31), that refer to single variables, useful overall measures for documentation purposes can be computed taking into account all the variables subject to E&I. For example, an **overall unweighted imputation rate** can be calculated as follows:

$$\frac{1}{p} \sum_j \left(\frac{\sum_i I(\hat{y}_{ij} \neq y_{ij})}{n} \right) \qquad \text{(D.40)}$$

This indicator corresponds to the average over the p variables subject to E&I of the **overall unweighted imputation rate** indicator (D.23) for single variables.

Appendix E

Glossary

The glossary covers the main definitions used in the handbook and is aimed to help the reader's understanding but does not claim to be exhaustive or to have an official status. The definitions used in this glossary are partly based on the EUROSTAT Glossary of Quality Terms (Eurostat, 2003) and on the UNITED NATIONS Glossary of Terms of Statistical Data Editing (United Nations, 2000). Some of these definitions had to be adapted to the notations and definitions used in this handbook.

Accuracy
Accuracy in the general statistical sense denotes the closeness of computations or estimates to the exact or true values.

Anticipated value
Anticipated values are used in score functions and are predictions for the values which are expected in the actual survey. The prediction is usually based on the data of previous surveys or auxiliary information.

Audit trail
A method for keeping track of changes to values of data items and the reason for each of these changes.

Balance edit
An edit which checks that a total equals the sum of its parts. Example: Labour Cost=Wages+..-..

Bias of an estimator
The bias of an estimator is the difference between the expectation of the estimator and the true value of the parameter being estimated.

Checking rule
See Edit.

Coding
Coding is a technical procedure for converting information into numbers or other symbols, which can be more easily counted and tabulated.

Cold deck
A donor questionnaire is found from a different survey as the questionnaire with the missing item. This donor questionnaire is used to supply the missing or inconsistent values. An example would be using prior year's data. (see also hot deck).

Computer Aided Interview - CAI

Use of computers during interviewing. Any contradictory data can be flagged by edit routines and the resultant data can be immediately adjusted by information from the respondent. Furthermore, data capturing (key-entry) is performed at interview time.

Consistency error

Occurrence of contradictory values with respect to a predefined relationship between these values of at least two variables.

Cut-off value

In the context of score functions: a discriminating value leading to a division of the data flow in a critical and a non-critical stream with respect to the applied score function.

Creative editing

An action whereby manual reviewers invent imputations to avoid reviewing another error message from subsequent machine editing. The error detection and localization may be influenced in the same way by creative editing.

Deductive imputation

Deductive imputation is performed when, given specific values of other fields, and based on a logical or mathematical reasoning, a unique set of values exists causing the imputed record to satisfy all the edits (e.g. when items must sum up to a total and only one item in the sum has to be imputed, then its value is uniquely determined by the values of the other items).

Deterministic checking rules

A deterministic checking rule determines whether data items are incorrect with a probability of 1. Example: "If `number of Employees`>0 and `Wages`=0 then set the inconsistency flag for `number of Employees`=1" means that the number of employees is classified as erroneous. As in this example deterministic checking rules may be questionable.

Deterministic imputation

A deterministic imputation method determines one unique value for the imputation of a missing or inconsistent data item. This means that when the imputation process is repeated, the same values will be imputed.

Donor

A donor is a record whose values (all or only a subset) are copied to the corresponding fields in a record (recipient) where these values are missing or inconsistent.

Donor pool

The donor pool is the set of records which may be used as donors in donor imputation. The donor pool usually consists of records considered not erroneous, i.e. edit-passing records.

Drill-down

Drill-down is an action where a possible error of a high-level aggregate detected during macroediting is investigated by focusing on lower-levels of aggregates and following the error down to individual observations (micro editing).

Edit

A logical condition or a restriction to the value of a data item or a data group) which must be met if the data is to be considered correct. Also known as **edit rule** or **checking rule**.

Editing
Detection of missing, invalid or inconsistent values.

Editing and Imputation process
The E&I process is defined as a process with a parameterization aiming at detecting erroneous and missing data and treating these data accordingly.

Electronic questionnaire
An electronic questionnaire is a questionnaire taking the form of a computer application that can be filled in by the respondent or the interviewer directly on the computer or through the Internet.

Error detection
An activity aimed at detecting erroneous data, using predefined criteria for completeness and consistency, which may be defined by the means of several edit rules.

Error localization
The (automatic) identification of the fields to be imputed in an edit-failing unit. Often an optimization algorithm is used to determine the minimal set of fields to impute such that the imputed unit will not fail edits.

Expectation maximization algorithm (EM)
The EM algorithm allows the maximum likelihood estimation of the model parameters from incomplete data by iteratively applying, until convergence, the expectation of a likelihood function (E-step) and its maximization (M-step).

Explicit edit
An edit explicitly written and defined.

Fatal edit
Identifies data errors with certainty. Examples are an economic activity that does not exists in a list of acceptable economic activities and balance edits. Also known as hard edit.

Fatal errors
Errors identified by fatal edits.

Flag
Auxiliary variable resulting from error detection or from treatment and indicating if an action has been performed (detection of an error, treatment of an error). Usually a binary variable with the value 1 if the action has been performed and 0 otherwise.

Global score function
A global score function is the combination of all defined local score functions, i.e. score functions defined for individual variables.

Graphical editing
Using graphs to identify anomalies in data. While such graphical methods can employ paper, the more sophisticated use powerful interactive methods that interconnect groups of graphs automatically and retrieve detailed records for manual review and editing.

Hard edit
See Fatal edit.

Hit rate
The "success" rate of an edit, the proportion of edit failures which point to true errors (see indicator D.4).

Hot-deck imputation
A donor record is found from the same survey as the record with the missing item(s). This donor record is used to supply values for the missing or inconsistent data item(s).

Implied edit
An unstated edit derived by logical operations from explicit edits.

Imputation
Imputation is the treatment of data used to treat problems of missing, invalid or inconsistent values identified during editing. This is done by substituting estimated values for the values flagged during editing and error localization.

Imputation rate
(Unweighted) imputation rate as defined by EUROSTAT (see indicator D.23).

Imputation ratio
(Weighted) imputation ratio as defined by EUROSTAT (see indicator D.31) .

Imputation variance
A component of the total variance of the survey estimate introduced by the imputation procedure.

Influential error
In selective editing an observation with a score above the cut-off value.

Inconsistency
Cf. consistency error.

Influential observation
An observation that has a large impact on a particular result (statistic) of a survey.

Inlier
An inlier as a data value that lies in the interior of a statistical distribution and is in error. Because it is often not possible to find a model that separates inliers from good data (in which case the inliers would become outliers w.r.t. that model) inliers are often difficult to detect and correct. An example of an inlier might be a value repeatedly, over time, recorded without change.

Input editing
Editing that is performed as data is input, e.g. during an interview. The editing may be part of the data entry process.

Interactive editing
Computer aided manual editing after the data capturing process. Interactive editing is usually assisted

by automatic editing.

Interactive treatment

Computer aided manual treatment of values flagged as erroneous during editing usually directly performed on the computer and assisted by implemented edit rules. Often interactive treatment includes also interactive editing.

Invalid value

A value of a data item that is not an element of the set of permissible values or codes assigned to that data item.

Item nonresponse

Item nonresponse occurs when a respondent provides some, but not all, of the requested information, or if some of the reported information is not usable.

Item response rate

Unweighted item response rate (see indicator D.19) and weighted item response rate (see indicator D.20) defined by EUROSTAT.

Local score function

A local score function is a score function for one variable.

Loop-back

A loop-back is a formalized drill-down in the process flow leading from phase 3, where the decision that the data has to be individually edited was taken, to treatment (maybe detection) of phase 2. In other words, the data flow is lead to a phase or a procedure of a preceding phase which will be performed again with adapted parameters.

Macro editing approaches

In macro editing approaches sub-samples or the entire sample are checked together, i.e. an important part of the data is edited simultaneously as opposed to micro editing approaches where the data is individually edited. The checks are typically based on statistical models which are often found by graphical methods. The purpose of macro editing approaches is still to treat individual data and therefore a drill-down is usually necessary.

Matching fields

In donor imputation, the fields in an edit-failing unit that do not fail any edits are matched against possible donors (see donor). Often a distance is calculated between the matching fields to chose among possible donors.

Mean square error

The expected value of the square of the difference between an estimator and the true value of a parameter. If the estimator is unbiased, then the mean square error is simply the variance of the estimator. For a biased estimator the mean squared error is equal to the sum of the variance and the square of the bias.

Micro editing

Finding errors by inspection of individual observations without referring to actual estimates based on all or the main part of the observations. Editing done individually at the record, or questionnaire level.

Missing values

Missing values stem from questions the respondent did not answer. This can happen for several reasons; the respondent may not know the answer, may not be willing to respond or may have simply missed a question.

Model based imputation

Imputation based on an explicitly described statistical model. E.g. use of averages, medians, regression equations, etc. to impute a value.

Multiple imputation

An observation with failing and/or missing values is imputed several times stochastically. Multiple imputation allows under certain conditions the correct estimation of the variance due to imputation. This estimation is based on a combination of the within and the between variance of the multiply imputed data.

Multivariate outlier detection

Outlier detection based on multivariate models for several variables simultaneously.

Nearest neighbor imputation (NNI)

In nearest neighbor imputation the donor is chosen in such a way that some measure of distance between the donor and the recipient is minimized. NNI is a deterministic imputation method in its basic form, but can be made stochastic.

Nonresponse

Nonresponse is a form of non-observation present in most surveys. Nonresponse means failure to obtain a measurement on one or more study variables for one or more elements selected for the survey. Nonresponse causes both an increase of variance, due to the decrease in the effective sample size and/or due to the use of imputation, and may cause a bias if the nonrespondents and respondents differ with respect to the characteristic of interest.

Non-sampling error

An error in sample estimates which cannot be attributed to sampling fluctuations. Such errors may arise from many different sources such as defects in the frame, faulty demarcation of sampling units, mistakes in the collection of data due to personal variations or misunderstandings or bias or negligence or dishonesty on the part of investigators or of the interviewee, mistakes at the stage of the processing of the data, or during the E&I process etc.

Outlier

An outlier is an observation which is not fitted well by a model for the majority of the data. For instance, an outlier may lie in the tail of the statistical distribution or “far away from the center” of the data.

Over-editing

The share of resources and time dedicated to editing is not justified by the resulting improvements in data quality (Granquist, 1995).

Phase

A phase is a compact set of procedures.

Predictive mean matching (PMM)

Method that can be considered a variant of the NNI where a model (typically the normal model with continuous numerical variables) is used only to define a suitable distance function. In particular for each missingness pattern, the predictive means of the missing values conditional on the observed ones are computed based on the assumed model. Then each recipient is matched to the donor with the closest predictive mean.

Procedure of the E&I process
A procedure is an implemented method or technique.

Processing error
Once data have been collected, they pass through a range of processes before the final estimates are produced: coding, keying, editing and imputation, weighting, tabulating, etc. Errors introduced at these stages are called processing errors.

Query edit
An edit rule whose failure indicates an error with probability less than 1. For example, a value that, compared to historical data, seems suspiciously high. Also known as soft edit or statistical edit.

Query errors
Errors identified by query edits.

Random donor imputation
Random donor imputation is the stochastic version of nearest neighbor imputation usually performed by randomly choosing a donor from the donor pool with distance below a threshold.

Range check
A range check verifies whether a data item value is in a previously specified interval.

Random error
Random errors are errors that are not caused by a systematic reason, but by accident.

Ratio edit
An edit rule determining the acceptable bounds for a ratio of two variables.

Recipient
A recipient is an observation with at least one missing or inconsistent value which will be imputed by donor imputation.

Repetition
If the outcome of a phase is not satisfactory then a repetition of the phase, i.e. a repetition of procedures with adapted parameters inside the phase may be decided.

Residual error
Residual errors may arise during the third phase of an E&I process because some detection methods could only be applied in this phase, e.g. comparison with the last surveys estimates or with external data or multivariate analysis. These errors may also have been introduced during the E&I process for example by treatment methods which do not take edit rules into account.

Response Fraction
Used to calculate economically weighted quality indicators. The weight is a product of the sampling

weight and a size indicator.

Rule-based imputation

An imputation method where the procedure is based on rules defined on the values of the other fields and/or the values to be replaced. Rule-based imputation is usually performed through imputation rules having the form "IF *error condition* THEN *imputation action*". Example: "IF flag of inconsistency for `number of employees=1` and `wages=0` THEN `employees=0`".

Score function

A function assigning a score to an observation which correlates well with the potential effect the observation may have on a predefined set of estimates (Latouche and Berthelot, 1992). Scores are calculated with the help of anticipated values for the data. Score functions are used in selective editing to prioritize observations for interactive treatment.

Selective editing

Selective editing is a procedure which, based on a score function, splits the data in a critical stream, with potential influential errors, and a non-critical stream. The critical stream goes through thorough editing and imputation whereas the non-critical stream goes through minimal or even no further editing and imputation. Selective editing allows to obtain large gains in cost without losing much quality.

Short-term statistic

Short-term statistics collect information (variables) necessary to provide a uniform basis for the analysis of the short-term evolution of supply and demand, production factors and prices (Council Regulation (EC) No 1165/98 of 19 May 1998). All variables are to be produced more frequently than annually.

Soft edit

See Query edit.

Statistical edit

See Query edit.

Statistical unit

An object of statistical survey and the bearer of statistical characteristics. The statistical unit is the basic unit of statistical observation within a statistical survey.

Structural statistics

Structural statistics shall have as its purpose, to analyze, for example, the structure and evolution of the activities of businesses, the regional, national, Community and international development of businesses and markets, the business conduct, the small and medium-sized enterprises, specific characteristics of enterprises related to particular groupings of activities (Council Regulation (EC) No 58/97 of 20 December 1997).

Structurally missing value

A special case of nonresponse arises due to filter questions which lead to unanswered questions in according parts of the questionnaire.

Study domains

Statistics are presented for different sub-groups of the population, so called study domains. These study domains are usually defined according to some classification (e.g. territorial units, economic activity, etc.).

Stochastic imputation

In stochastic imputation the imputed value contains a random component. Repetition of the imputation leads to a different result.

Systematic error

A systematic error is a type of error for which the error mechanism and the imputation procedure are known. A well-known systematic error is the so called unity measure error, e.g. values reported in Euros instead of 1'000 Euros.

Target population

The target population is the population we wish to study, that is, the set of elements about which estimates are required.

True value

The actual population value that would be obtained with perfect measuring instruments and without committing any error of any type, both in collecting the primary data and in carrying out mathematical operations.

Unit nonresponse

Unit nonresponse is a a complete failure to obtain data from a sample unit. The term encompasses a wide variety of reasons for non observation: "impossible to contact", "not at home", "unable to answer", "incapacity", "hard core refusal", "inaccessible", "unreturned questionnaire", and others. In the first two cases contact with the selected element is never established.

Unit response rate

Unweighted unit response rate (see indicator D.38) and weighted unit response rate (see indicator D.39) as defined by EUROSTAT.

Unity measure error

This error occurs when respondents report the value of a variable in a wrong unity measure. For example, if total turnover is required to be reported in thousands of Euros, but the amount is erroneously declared in Euros.

Variance

Variance is the mean square deviation of the variable around the average value. It reflects the dispersion of the empirical values around its mean.

Variance estimation

The task of estimating the value of the variance of an estimate. The method employed are commonly classified in: analytic methods: use and compute the proper formulas of the variance; approximate methods: methods which use approximations for complex and multi-stage sample design.

Bibliography

C. Béguin and B. Hulliger. Multivariate oulier detection in incomplete survey data: The epidemic algorithm and transformed rank correlations. *J.R.Statist.Soc.A*, 167(Part 2.):275–294, 2004.

L. Breiman, J. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees*. CRC Press, 1984. ISBN 0412048418.

R. Chambers, A. Hentges, and X. Zhao. Robust automatic methods for outlier and error detection. *Journal of the Royal Statistical Society, Series A*, 167(2):323–339, 2004.

R.L. Chambers. Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81(396):1063–1069, 1986.

J. Chen and J. Shao. Nearest neighbor imputation for survey data. *Journal of Official Statistics*, 16: 113–131, 2000.

T. De Waal and W. Coutinho. Automatic editing for business surveys: An assessment of selected algorithms. *International Statistical Review*, 73:73–102, 2005.

T. De Waal, F. Van de Pol, and R. Renssen. Graphical macro editing: Possibilities and pitfalls. In *Proceedings of the Second International Conference on establishment Surveys*, 2000.

G. Della Rocca, O. Luzi, E. Scavalli, M. Signore, and G. Simeoni. Evaluating, monitoring and documenting the effects of editing and imputation in istat surveys. Technical report, Working paper No. 3 presented a the Conference of European Statisticians, Work Session on Statistical Data Editing, Madrid, Spain, 2003.

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood estimation from incomplete data via the em algorithm. *Journal of the Royal Statistical Society Series B*, 39(1):1–38, 1977.

M. Di Zio, U. Guarnera, and O. Luzi. Editing systematic unity measure errors through mixture modelling. *Survey Methodology*, 31:53–63, 2005a.

M. Di Zio, U. Guarnera, O. Luzi, and A. Manzari. Evaluating the quality of editing and imputation: the simulation approach. *Work Session on Statistical Data Editing, UN/ECE*, 2005b.

M. Di Zio, U. Guarnera, and R. Rocci. A mixture of mixture models for a classification problem: the unity measure error. *Computational Statistics and Data Analysis*, 51:2573–2585, 2007.

P. Duchesne. Robust calibration estimators. *Survey Methodology*, 25:43–56, 1999.

EUREDIT Project. *Methods and Experimental Results from the Euredit Project*, volume 2. http://www.cs.york.ac.uk/euredit/results/results.html, 2004a.

EUREDIT Project. *Towards Effective Statistical Editing and Imputation Strategies - Findings of the Euredit project*, volume 1. http://www.cs.york.ac.uk/euredit/results/results.html, 2004b.

Eurostat. Working group on quality: Assessment of the quality in statistics: methodological documents - glossary, 6th meeting. Technical report, EUROSTAT, October 2003.

K. Farwell and M. Raine. Some current approaches to editing in the abs. In *Proceedings of the Second International Conference on Establishment Surveys*, pages 529–538, 2000.

P.I. Fellegi and D. Holt. A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association, Applications Section*, 71:17–35, 1976.

S. Franklin and M. Brodeur. A practical application of a robust multivariate outlier detection method. In *ASA Proceedings of the Section on Survey Research Methods*, pages 186–191. American Statistical Association, 1997.

W.A. Fuller. Simple estimators for the mean of skewed populations. Technical report, U.S. Bureau of the Census, 1970.

L. Granquist. Improving the traditional editing process. In *Business Survey Methods*, pages 385–401. John Wiley and Sons, 1995.

L. Granquist and J. Kovar. Editing of survey data: How much is enough? In *Survey Measurement and Process Quality*, pages 415–435. John Wiley & Sons, 1997.

L. Granquist, J. Kovar, and S. Nordbotten. Improving surveys - where does editing fit in? In *Statistical Data Editing Volume No. 3, Impact on Data Quality*, pages 355–361. United Nations, New York and Geneva, 2006.

J.P. Gwet and L.P. van Rivest. Outlier resistant alternatives to the ratio estimator. *Journal of the American Statistical Association*, 87:1174–1182, 1992.

D. Haziza. Imputation classes. *The Imputation Bulletin*, 2(1):7–11, 2002.

D. Hedlin. Score functions to reduce business survey editing at the U.K. office for national statistics. *Journal of Official Statistics*, 19(2):177–199, 2003.

M.A. Hidiroglou and J.M. Berthelot. Statistical editing and imputation for periodic business surveys. *Survey Methodology*, 12(1):73–83, June 1986. Statistics Canada.

M.A. Hidiroglou and K.P. Srinath. Some estimators of a population total from simple random samples containing large units. *Journal of the American Statistical Association*, 76:690–695, 1981.

G. Houston and A.G. Bruce. gred: Interactive graphical editing for business surveys. *Journal of Official Statistics*, 9(1):81–90, 1993.

P.J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35:73–101, 1964.

B. Hulliger. Outlier robust Horvitz-Thompson estimators. *Survey Methodology*, 21(1):79–87, 1995. Statistics Canada.

B. Hulliger. Simple and robust estimators for sampling. In *Proceedings of the Section on Survey Research Methods*, pages 54–63. American Statistical Association, 1999.

K.N. Javaras and D.A. van Dyk. Multiple imputation for incomplete data with semicontinuous variables. *Journal of the American Statistical Association*, 2003.

G. Kalton and D. Kasprzyk. Imputing for missing survey responses. In *Proceedings of the section on Survey Research Methods*, pages 22–31. American Statistical Association, 1982.

M. Latouche and J.M. Berthelot. Use of a score function to prioritize and limit recontacts in editing business surveys. *Journal of Official Statistics*, 8(3):389–400, 1992.

D. Lawrence and R. McKenzie. The general application of significance editing. *Journal of Official Statistics*, 16(3):243–253, 2000.

H. Lee. Outliers in business surveys. In *Business Survey Methods*, pages 503–526. John Wiley and Sons, 1995.

J. Little and D. Rubin. *Statistical Analysis with Missing Data*. Wiley, New York, 2002.

R.J.A. Little. Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, 6(3):287–296, 1988.

S. Lundström and C.-E. Särndal. Estimation in presence of imputation. Technical report, Statistics Sweden, 2001.

J. Pannekoek and T. De Waal. Automatic edit and imputation for business surveys: The dutch contribution to the euredit project. *Journal of Official Statistics*, 21(2):257–286, 2005.

J.N.K. Rao and J. Shao. Jackknife variance estimation with survey data under hot-deck imputation. *Biometrika*, 79:811–822, 1992.

P.J. Rousseeuw and K. van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41:212–223, 1999.

D. Rubin. *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York, 1987.

J.L. Schafer. *Analysis of Incomplete Multivariate Data*. Chapmann and Hall/CRC, New York, 2000.

D.T. Searls. An estimator for a population mean which reduces the effect of large observations. *Journal of the American Statistical Association*, 61:1200–1204, 1966.

J. Shao and R.R. Sitter. Bootstrap for imputed survey data. *Journal of the American Statistical Association*, 91:1278–1288, 1996.

C.J. Skinner, D. Holt, and T.M.F. Smith. *Analysis of Complex Surveys*. John Wiley and Sons, 1989.

C.E. Särndal and S. Lundström. *Estimation in Surveys with Nonresponse*. Wiley, 2005.

Statistics Canada. *Statistics Canada Quality Guidelines*. 2003.

United Nations. *Glossary of terms of statistical data editing*. United Nations, Geneva 2000, 2000.

P. Weir, R. Emery, and J. Walker. The graphical editing analysis query system. In *Statistical Data Editing Methods and Techniques, Volume No. 2*, pages 51–55. United Nations Statistical Commission and Economic Commission for Europe, 1997.

P. Whitridge and J. Kovar. Applications of the generalised edit and imputation system at statistics canada. In *Proceedings of the section on Survey Research Methods*, pages 105–110. American Statistical Association, 1990.

Working group on quality. Assessment of the quality in statistics: standard quality report. Technical report, EUROSTAT, 2003.

Working group on quality. Assessment of the quality in statistics: standard quality indicators, 7th meeting. Technical report, EUROSTAT, May 2005.