

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

CONFERENCE OF EUROPEAN STATISTICIANS

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

Discussion on Volume 3

STATISTICAL DATA EDITING

Volume No. 3

IMPACT ON DATA QUALITY

Submitted Statistics Canada¹

¹ Prepared by John Kovar (kovar@statcan.ca).

STATISTICAL DATA EDITING

Volume No. 3

Impact on Data Quality

Contents

Preface

Introduction

Chapter 1 – Framework of Evaluation

Chapter 2 – Quality Measures

Section 2.1 – Indicators

Section 2.2 – Impact of the Edit and Imputation Processes on Data Quality and Examples of Evaluation Studies

Section 2.3 – Impact on precision

Section 2.4 – Data Users

Chapter 3 – Improving Quality

Section 3.1 – Continuous improvement

Section 3.2 – Effect of collection mode

Section 3.3 – Alternate data sources

Chapter 4 – Improving Surveys – Where does editing fit in?

Preface

This publication, *Statistical Data Editing: Impact on Data Quality*, is the third in the series of Statistical Data Editing publications produced by the members of the UN/ECE Statistical Data Editing Workshop (SDE). While the first two volumes dealt respectively with the topics of *what* is data editing and *how* is data editing accomplished, the principal focus of this volume is the impact of the activity on the quality of the outgoing estimates. The aim of the volume is to assist National Statistical Offices in assessing the impact of data editing process on data quality, that is, *how well* is the process working.

The material was prepared in the framework of the project on Statistical Data Editing in the programme of work of the Conference of European Statisticians. It comprises of selected papers presented at the Statistical Data Editing Workshops held in Prague (October 1997), Rome (June 1999), Cardiff (October, 2000), Helsinki (June, 2002), and Madrid (October 2003). Many of these papers were updated and the new versions were tabled at the Ottawa meeting (May 2005). In addition, a number of additional papers were solicited for the publication, and also presented at the Ottawa meeting of the Workshop. The volume was compiled and edited by the Statistical Division of the United Nations Economic Commission for Europe. It represents an extensive voluntary effort on the part of the authors. The editors express their thankfulness to all authors who contributed to the publication.

A special thanks is also extended to the members of the steering group namely, Leopold Granquist (*Statistics Sweden*), John Kovar (*Statistics Canada*), Carsten Kuchler (*Statistisches Bundesamt, Germany*), Pedro Revilla (*Instituto Nacional de Estadística, Spain*) Natalie Shlomo (*Central Bureau of Statistics, Israel and Southampton Social Statistics Research Institute, University of Southampton, UK*), Heather Wagstaff (*Office for National Statistics, U.K.*), and Paula Weir (*U.S. Energy Information Administration*), for coordinating and introducing the individual chapters, and for helping to compile and edit the material. Their joint efforts contributed significantly to the preparation of the publication as a whole.

Introduction– *John Kovar, Statistics Canada*

Data editing is an integral and expensive part of the entire survey process; its impacts are far reaching. In this volume we examine how editing affects data quality. We use the broadly accepted meaning of data quality, that is, its *fitness for use*, or *fitness for purpose*. In particular, we address issues of accuracy, relevance, coherence and interpretability. Furthermore, we are concerned with numerous stakeholders in the survey process, be they the users, respondents or data producers themselves.

To begin, we provide examples of evaluation frameworks of the editing process: What to plan for? What to measure? What to retain so we can learn from the process? Chapter 1 deals with these issues in detail, with a practical implementation example from the Italian Statistical Office.

In Chapter 2 we address specifically the issues of *what* to measure. Numerous quality indicators are proposed, both from a theoretical as well as practical point of view. We examine how editing impacts on data quality in a quantitative way. Considerations of the impact of imputation on the total survey error are provided. We end the chapter by suggesting what quality information should be provided to the data users and in what way.

Chapter 3 tries to address the problem of *improving* data quality. What can we learn from the editing process in order to be able to continuously improve the product? Specific examples of how the data editing process should fit in the statistical infrastructure are presented. More globally, we also address the problem of where and how editing should fit in the various modes of data collection, specifically as it relates to electronic data reporting and the use of alternate data sources.

Finally, in Chapter 4 we consider quality in its full breadth. We review what has been done and what was planned but not done in evaluating quality. We argue that editing should be considered in a wider perspective by providing information to other survey processes in a systematic way. Some future directions are outlined.

This volume is not intended to be an exhaustive literature review of the topic. Rather it is based on contributions to the state of the art made by the participants of the Statistical Data Editing Workshop over the past decade or so. The readers are also encouraged to visit the group's data editing knowledge base at <http://amrads.jrc.cec.eu.int/k-base/> for a complete documentation of the SDE efforts.

Chapter 1: Framework of Evaluation

Foreword – *John Kovar, Statistics Canada*

Three papers were selected for this introductory chapter which explores the question of how to plan and implement evaluation studies.. First, Svein Nordbotten explores the details of how one should approach the problem of evaluating the efficiency of the entire data editing problems by providing a general framework. He explores the issues of what to measure and how to measure it. Opportunities for systematic exploration and evaluation of relationships among statistical product quality and the editing process variables are considered with the goal of developing a causal model of the editing process in order to design a more efficient editing processes.

The paper by Ray Chambers concentrates more specifically on evaluation criteria of particular approaches with the goal of establishing a “best practice”. It is important to underscore that the proposed criteria assess performance of the methods when the “true” values are known. They may thus not be appropriate for assessing performance of methods with a specific real data set where the true values are not known. However, the described approaches are clearly valuable in providing guidance in finding effective edit and imputation methods.

The Italian paper examines the evaluation phase with a view to see if the edit and imputation process is in fact capable of achieving the dual objective of editing: improving data quality and as well as providing information on survey process (including the E&I process itself) with a view to continuous improvement. Particular evaluation objectives are explored, an extensive bibliography is presented and put in context and a discussion based on the Italian experience is provided.

Evaluation of statistical processes in general and the edit and imputation process in particular is relatively new domain that is in need of continuous study. Some tasks for future research in this area are identified at the end of the first paper.

- Svein Nordbotten: “Evaluating Efficiency of Statistical Data Editing: General Framework”, United Nations. The document was reviewed at the SDE Worksession in Rome, June 1999, and updated before the SDE Worksession in Cardiff, October 2000.
- Ray Chambers: “Evaluation Criteria for Editing and Imputation in Euredit”, Cardiff 2000, WP.5.
- Marco Di Zio, Orietta Luzi and Antonia Manzari: “Evaluating the Editing and Imputation Processes: The Italian Experience”, Helsinki 2002, WP.12, with minor updates tabled in Ottawa 2005, CRP.4

Chapter 2: Quality Measures

Foreword – *Paula Weir, U.S. Energy Information Administration*

Chapter 2 contains four sections that focus on the actual measurement and communication of quality with respect to edit and imputation. In the first section, a variety of performance indicators by purpose and data type are proposed. Approaches are provided on *how* to perform the measurement in order to effectively evaluate the quality of the E&I process and its relationship to overall survey quality. The second section moves on to explore the evaluation of methods and procedures from both the point of view of functionality, as well as empirically, to determine the impact of the process on the final data and the estimates produced. In the third section, precision of estimates is addressed by examining the impact of the E&I process on the variance. Section 4 concludes the chapter by examining what knowledge and metadata on E&I should be shared and disseminated to the data users and data producers to understand the attributes of the product (i.e., the statistics), as well as promote continuous improvement of the E&I process.

Section 2.1: Indicators

Foreword – *Carsten Kuchler, Statistisches Bundesamt, Germany*

When reasoning about indicators describing data quality, mainly two questions arise, namely *what* to measure in general and *how* to measure it particularly; or in other words: what are the basic concepts of quality that are to be reflected by the indicators, and how and to what extent can they be applied to a survey at hand. The contributions in this section tackle both questions with respect to the impact of data editing and imputation on the overall data quality.

The what-to-measure question is discussed by the first three contributions. Di Zio et al. (2005) derive some basic indicators referring to comparisons between true, raw and processed data, and subsequently discuss problems of measuring these indicators in a simulation approach. Thus, their focus is more on judging methods by means of synthetic data in order to provide a priori criteria for the selection of methods. On the other hand, Della Rocca et al. (2005) are geared to characterising a particular survey. Their starting point is a brief discussion of some rather fundamental criteria applied to the editing and imputation process (like preserving individual data, marginal and joint distributions, aggregates, etc.), that prepare the ground for a variety of specific measures to be applied to a particular data set under editing and imputation. Unlike the first two papers that chiefly address the bias due to editing and imputation, Rancourt (2002) deals with the impact of editing interventions in terms of variance measures. The basic idea is that editing needs to be considered a statistical process that causes missing data. Under this assumption Rancourt adopts well-established approaches for estimating the variance due to imputation to derive a general measure for reasoning about the increase of variance due to editing.

Beyond these rather abstract quality indicators, the contributions introduced so far point up the crucial role of software tools for relating the question of what to measure with the subsequent question of how to measure it. In this respect two problems have to be tackled by appropriate software tools: (a) how to judge editing and imputation methods by means of these indicators under controllable circumstances, and (b) how to integrate the involved evaluation of the indicators in a particular survey process. The first question is addressed by Rancourt and Di Zio et al., which both present simulation environments that allow for applying various editing and imputation methods to a given data set in order to account for the selection of a particular method. In addition Di Zio et al. discuss the generation of synthetic data that even bring the conditions of data production under the methodologists' control. Such synthetic data sets are required for the comparison of the true and processed data as proposed by Di Zio et al. for estimating the bias component due to specific editing and imputation procedures. The second problem refers to the costly evaluation of indicators and is addressed by Della Rocca et al. They present the IDEA system that generates the proposed indicators in one go while performing editing and imputation work steps. Instead of reducing costs by generating indicators more or less automatically in the application flow, Smith and Weir (2000) argue, that the set of indicators to be evaluated may be reduced to those that are meaningful for describing a survey at hand. These indicators are identified by means of Principal Component Analysis with respect to a reference survey. Indicators of low explanatory value can be left out in subsequent surveys. In two examples, Smith and Weir show the effect of this approach. After having derived a variety of data quality indicators in the past years, this approach opens a new view by providing a reasonable criterion for assessing indicators with respect to a particular survey.

- Marco Di Zio, Ugo Guarnera, Orietta Luzi, and Antonia Manzari: "Evaluating the Quality of Editing and Imputation: The Simulation Approach". A consolidated paper based on: Giulio Barcaroli and Leandro D'Aurizio: "Evaluating Editing Procedures: The Simulation Approach", Prague 1997, WP.17; and on Antonia Manzari and Giorgio Della Rocca, "A Generalized System Based on a Simulation Approach to Test the Quality of Editing and Imputation Procedures", Rome 1999, WP.13, tabled in Ottawa 2005, WP.50
- Eric Rancourt: "Using Variance Components to Measure and Evaluate the Quality of Editing", Helsinki 2002, WP.11.
- Giorgio Della Rocca, Marco Di Zio, Orietta Luzi, Marina Signore and Giorgia Simeoni. "Quality Indicators for Evaluating and Documenting Editing and Imputation". A revised paper based on Madrid 2003, WP.3, and Rome 1999, WP.3, tabled in Ottawa 2005, CRP.3

- Paul Smith and Paula Weir: “Characterisation of Quality in Sample Surveys Using Principal Component Analysis”, Cardiff 2000, WP.4

Section 2.2: Impact of the Edit and Imputation Processes on Data Quality and Examples of Evaluation Studies

Foreword – *Nathalie Shlomo, Central Bureau of Statistics, Israel and Southampton Social Statistics Research Institute, University of Southampton, UK*

This section includes five papers covering a wide range of evaluation studies on the impact of editing and imputation processes on the quality of the data. The Whitridge-Bernier paper focuses on economic business data and compares the edit and imputation carried out at both the data collection stage and the data processing stage and the impact on the quality of the final estimates. The analysis makes use of important metadata for each statistical unit describing the various stages of data processing and edit and imputation.

The Hoogland paper describes experiences in evaluating several important stages of edit and imputation processes for economic business data: the effectiveness of selective editing criteria for determining which statistical units need to be manually reviewed and which can be treated automatically, and the evaluation of automatic software for edit and imputation. The methodology and implementation of plausibility indicators for channelling the statistical units to the relevant phases of data processing is described. In addition, cut-off points for selective editing criteria and test statistics for comparing the quality of the statistical data at different stages of the edit and imputation processes are defined for the evaluation.

A wide range of test statistics for evaluating edit and imputation procedures on the quality of the data can be found in Charlton paper. The paper describes results from the EUREDIT Project funded under the EU Fifth Framework research program which brought together twelve partners from seven European countries to collaborate on the development and evaluation of edit and imputation methods. The aim of the project was to provide a statistical framework for comparing methods and for determining optimal methods according to the statistical units in the data set, the pattern of non-response and the scope and amount of erroneous and missing values in the data.

The final two papers evaluate software packages and algorithms of the error localization problem based on subjective and objective criteria. Error localization refers to the problem of changing as few fields as possible so that all edit checks are satisfied. The Poirier paper gives a thorough functional comparison between software packages and evaluates the packages with respect to their strengths and limitations, the edit and imputation method implemented, the type of statistical data that can be treated, online help and support and other subjective criteria. The final De Waal paper compares sophisticated error localization algorithms for their optimality and computational speeds. Four algorithms are evaluated and compared on real numerical data sets with different patterns and amounts of errors.

- Patricia Whitridge and Julie Bernier: “The Impact of Editing on Data Quality”, Rome 1999, WP.5.
- Jeffrey Hoogland “Selective Editing Using Plausibility Indicators and Slice”. A revised paper based on Madrid 2003, WP4, tabled in Ottawa 2005, CRP.2.
- John Charlton: “Evaluating New Methods for Data Editing and Imputation – Results from the Euredit Project”, Madrid 2003, WP.25.
- Claude Poirier: “A Functional Evaluation of Edit and Imputation Tools”, Rome 1999, WP.12.
- Ton de Waal: “Computational Results for Various Error Localisation Algorithms”, Madrid 2003, WP.22.

Section 2.3: Impact on precision

Foreword – *Pedro Revilla, Instituto Nacional de Estadística, Spain*

This section includes two papers covering the problem of variance estimation taking into account imputation and nonresponse. While one of the papers focuses more on the theoretical problems, the other one presents some empirical studies.

The paper by Rancourt presents an overview of the available theory and methods to measure and understand the impact of the imputation and nonresponse on variance estimation. Specific software to carry out these tasks is presented. It also proposes potential applications and research avenues.

The Spanish paper studies the performance of resampling variance estimation techniques under imputation, using data from the Structural Industrial Business Survey and the Retail Trade Index Survey. The jackknife variance estimation based on adjusted imputed values and the bootstrap procedures are applied. The performance of the two methods is measured through the Monte Carlo bias, the mean square error and the coverage rate of the 95 percent confident interval based on normal approximation.

- Eric Rancourt: : “Assessing and dealing with the impact of imputation through variance estimation”, Ottawa 2005, WP.10 – Loosely based on his Madrid paper: “Statistics Canada’s New Software to Better Understand and Measure the Impact of Non-Response and Imputation”, Madrid 2003, WP.29.
- Felix Aparicio-Pérez and Dolores Lorca: Performance of resampling variance estimation techniques with imputed survey data. Paper based on (1) Felix Aparicio and Dolores Lorca: “Performance of Bootstrap Techniques with Imputed Survey Data”, Madrid 2003, WP14, and (2) F. Aparicio and D. Lorca: “Performance of Jackknife Variance Estimation Using Several Imputation Methods”, Helsinki 2002, WP17, tabled in Ottawa 2005, WP.19.
- *Other papers from the Ottawa meeting may be candidates – if so, introduction will have to be modified*

Section 2.4: Data Users

Foreword – *Heather Wagstaff, Office for National Statistics, U.K.*

There are three papers in this section that discuss what knowledge and metadata about the primary source data and the data editing process should be disseminated to the data users. The first, by Nordbotten, discusses the sharing of knowledge between producers and users of statistics. Nordbotten argues that statistics can become commercial commodities. Hence, he explores how the statistical market mechanism might work in the future to meet the information needs of the users. The discussion is mainly confined to metadata about the accuracy of statistics and statistical data editing processes. It acknowledges that, in reality, the users may place higher priority on other dimensions of quality. Emphasis is placed on the importance of understanding the needs of the end users and how a statistical producer should provide these metadata. The paper concludes that further research should focus on which data is needed by end-users, how the data should be collected and the form in which the metadata about accuracy should be disseminated.

The second paper, by Luzi and Manzari, discusses the dissemination of knowledge within the National Statistical Institute. This is analysed from the perspective of information flows among the internal users of editing and imputation methods (the survey statisticians) and the centralised methodologists. The analysis is based on a functional model, the data editing method life cycle, as a conceptual framework. The paper concludes that there are significant gains to be made from having a good mechanism for information exchange. This in turn ensures that users understand the concepts and facilitates discussions on principles, techniques, methods and systems.

The final paper in this section, by Rouhuvirta, emphasises the direct link between the content management of register data and the quality of the resultant statistical data. The paper considers ways to present the metadata about statistical data which has been compiled from register sources to ensure a good understanding and accurate interpretation by users of the statistical data. There follows a discussion of methods to ensure that adequate metadata about data from registers is available at the data editing phase and how to preserve metadata from the editing process for utilisation in later stages of statistical production. A practical example is presented about the problem of structuring of taxation metadata and its utilisation.

- Svein Nordbotten: “Metadata About Editing for End Users”, Cardiff 2000, WP.14.
- Orietta Luzi and Antonia Manzari: “Data Editing Methods and Techniques: Knowledge to and from Users”, Cardiff 2000, WP.13.
- Heikki Rouhuvirta: "Conceptual Modelling of Administrative Register Information and XML - Taxation Metadata as an Example", Ottawa 2005, WP.3

Chapter 3: Improving Quality

Foreword – *Leopold Granquist, Statistics Sweden*

Chapter 3 contains 3 sections on themes how data editing could essentially improve data quality of surveys.

The first section covers 3 papers that all has an underlying principle: editing should primarily focus on identifying and eliminating error sources in advance of cleaning up the data. A key role plays Process Data System (PDS) that is outlined in the first paper, or Metadata systems as described in the second paper. Learning from editing is the key element of the approach. An application of the concept concludes the section showing that costs both for the producers and data providers can be reduced essentially as well as easing the respondent burden and at the same time real gains in data quality can be achieved.

The second section has 5 contributions in general focused on web surveys or computerized self administered questionnaires (CSAQs). Such data collections offer excellent opportunities of getting valuable information how respondents interpret questions and understand the survey concepts. The importance of interviewing respondents in this context cannot be underestimated in particular for CSAQs but in general for other modes of data collection. It should be noted that moving the editing to the reporting phase of data collection is beneficial for the respondents and for the ingoing data quality. An important issue is to convince the respondents that it is advantageous for them to use this mode of reporting when other alternatives are given, for example by giving them valuable feed back.

Section 3 presents 5 papers on topics using alternate data sources in data collection and improving data quality in both business and social subject matter areas. Of particular importance are:

- Concepts of incorporating administrative data into a framework of editing and imputation are defined for a multi-source data collection.
- The need for evaluating administrative records for their accuracy and completeness.
- Editing can be decreased by linking multiple data sources at the unit level.
- The quality can be improved by exploiting all available information, especially important metadata which contain information about the coverage of the frame, non respondents, field investigation and other data processing stages.
- An example of using administrative data throughout the entire data processing of a survey. An interesting feature of the survey is the need to develop a complete census for all units in the population using mass imputation procedures based on the sample and the administrative sources.

Section 3.1: Continuous improvement

Foreword – *Carsten Kuchler, Statistisches Bundesamt, Germany*

The commonplace that the best errors are those not occurring, perfectly describes the direction of data editing efforts since the mid 1990s. By inter-relating data editing to previous and subsequent stages of the survey process, background knowledge from fixing errors became valuable in identifying and suppressing potential error sources within the entire survey process. In addition enhancements in other sub processes were based on indicators and statistical measures derived during data editing, and vice versa. Finally data editing methods were selected and improved with respect to methodological requirements raised by adjacent processes. However, these cross-connected improvements remain ad hoc until they are embedded and reflected in feedback loops that relate sub processes of subsequent surveys.

The common question tackled by the papers collected in this section is, how to proceed from single improvements concerning methods and applications towards a continuous improvement of the entire survey process. The answers given in the papers shed light on this problem from different perspectives: Tackling continuous improvement requires conceptual orientation like given by data quality objectives, adaptable application flows, and –as referred to by any of the contributions– metadata for monitoring, evaluating and relating sub processes. In the succession of the papers the focus moves from basic methodological considerations towards their concrete implementation in National Statistical Agencies.

For this purpose the first paper by Engström and Granquist is an appropriate starting point. It shows one of the first and most influential attempts in consolidating different data editing approaches with respect to the idea of assuring data quality in an integrated and inter-related application flow. Several of the ideas proposed by Engström and Granquist re-emerge in the remaining papers of this section. They discuss concrete approaches like the identification of data error sources and the standardisation of editing processes by establishing current best practices, which all implicitly refer to the same issue: single improvements rely on metadata to become subject of continuous improvement. Based on general standards of how to monitor application flows, the authors outline so called Process Data Subsystems (PDS) that are intended to store quality characteristics, information about the methods applied and their particular performance.

It is a rare occasion for methodological considerations like these, to prove their relevance, when National Statistical Agencies revise their entire process and application flow. In the second contribution Pam Tate describes the way the data editing process is embedded into the Statistical Value Chain, underlying the ambitious modernisation programme of the Office for National Statistics (ONS). The basic message is that metadata are the lubricant of a complex and highly interdependent process flow. In describing the relation between the chain links of the survey process as considered by the ONS, the paper specifies types of metadata with respect to their application in processing a particular data set or managing and evaluating the entire survey process. In order to provide each survey sub process with the required metadata, the ONS established a Central ONS Repository for Metadata (CORM) that can be considered a descendant of the Process Data Subsystems introduced by Engström/Granquist.

The CORM (meta-) database indicates the crucial role of software for the requirements of planning, accomplishing, monitoring and evaluating editing processes. Implementing tailor-made software for specific survey processes opens the chance to link these processes by metadata streams that are provided by well-defined and general interfaces. The paper by Kuchler and Teichmann presents the implementation of such a metadata stream in the Federal Statistical Office Germany. Point of reference is the embedding of an IT-tool for data editing with respect to an office wide software standard. The so called PL-Editor is considered a gateway for metadata utilised in previous and subsequent sub processes. Beside a brief description of the software standard and the technical realisation of the underlying interfaces, the paper mainly deals with the question of how methodological enhancements can easily be integrated into an existing application flow. Thus a flexible software standard is shown to be necessary to allow for continuous improvement of survey processes.

Starting from rather general methodological considerations this section touched the implementation of actions assuring continuous improvement to finally tackle the question of how these ideas may be applied in a specific survey. Thus the last paper by Colleen Martin and Claude Poirier describes a concrete process of continuous improvement the Unified Enterprise Survey (UES) conducted by Statistics Canada ran through since 2002. The authors refer to both the collection and the post-collection sub process. In particular they introduce the effects of setting up quality control processes in the collection phase and show how the use of external data

- Per Engström and Leopold Granquist: “Improving Quality by Modern Editing”, Rome 1999, WP23.
- Pam Tate: Promised to prepare a more targeted version of her paper based on “The Data Editing Process within the New Statistical Infrastructure of the Office for National Statistics”, Madrid 2003, WP9, tabled in Ottawa 2005, WP.3.
- Carsten Kuchler and Corina Teichmann: Linking Data Editing Processes by IT-Tools, a paper based on Carsten Kuchler: “IT Tools for an Integrated Data Editing Concept”, Madrid 2003, WP19, tabled in Ottawa, 2005, WP.15
- Jean-Sébastien Provençal: An update of “Analysis of Data Slices and Metadata to Improve Survey Processing”, by Colleen Martin and Claude Poirier, Helsinki 2002, WP2, with references to Poirier, Phillips and Pursey, Madrid 2003, and to Hazelton, Madrid 2003. Tabled in Ottawa under topic (v), WP.51

Section 3.2: The effect of data collection on editing and data quality.

Foreword – Pedro Revilla, Instituto Nacional de Estadística, Spain

This section includes five papers covering the issue of the effects of data collection methods on data quality. The main focus of the papers is the impact of Electronic Data Reporting on the design of the editing strategy.

The paper by Nichols et al. presents the U.S. Census Bureau’s experience with data editing strategies used in business surveys. It describes the interactive editing approach currently incorporated into computerised self-administered questionnaires (CSAQs), which are delivered to the respondent by downloadable files or transmitted electronically over the Internet.

The paper by Paula Weir presents the change in electronic data reporting options and usage for the Energy Information Administration of US. One fully web-based survey that recently implemented an editing module is examined in more detail to better understand the respondents’ views and use of the edit feature. Respondents were asked how they used the edit function and how clear and useful the information provided for edit failures was. The responses regarding the edit function for this survey is further compared to a study of the edit log which records information each time the edit function is invoked.

The Spanish paper explores the possibilities of Web questionnaires in order to improve editing. It discusses the possibilities and challenges that Web questionnaires provide to the editing tasks, in particular the combination of built-in edits and selective editing approach. Some practical experiences in the Spanish Monthly Turnover and New Orders Survey are presented.

The paper by Laroche is focused on the evaluation of the Internet option offered to the respondents in the 2004 Census test. It is planned that all the households will be offered with the possibility of using Internet in the 2006 Census. A satisfaction survey of the respondents using the Internet and a follow up survey of the respondents not using the Internet are described. Comparisons between the two ways of data collection (paper and Internet) are presented.

Finally, the paper by Ceccarelli and Rosati describes the data editing method used for the Italian Labour Force Survey. It presents the data edit and imputation strategy implemented for the whole survey process. A discussion on the main outcomes of the effect of using a combination of CAPI and CATI for data collection is also presented.

- Amy E. Anderson et al: Designing edits for electronic economic surveys and censuses: Issues and guidelines, Ottawa 2005, WP.22
- Pedro Revilla: EDR Impacts on editing, Ottawa 2005, WP.27
- Paula Weir: Electronic data reporting and the impact on editing – a summary and a case study, a paper based on Paula Weir: “Electronic data reporting – moving editing closer to respondents”, Madrid 2003, WP37, tabled in Ottawa 2005, WP.28
- Danielle Laroche and Laurent Roy: Evaluation of data collection via internet for the 2004 Census of population test, Ottawa 2005, WP.23

- Claudio Ceccarelli and Simona Rosati: Editing and imputation strategy in the Italian Labour Force Survey, Ottawa 2005, WP.24

Section 3.3: Making use of alternate data sources

Foreword - *Nathalie Shlomo, Central Bureau of Statistics, Israel and Southampton Social Statistics Research Institute, University of Southampton, UK*

The focus in this section is twofold: effective edit and imputation procedures which ensure consistent, logical and high quality records when combining multiple data sources for producing statistical data; and the enhancement and augmentation of survey data when incorporating administrative data into the edit and imputation phase of data processing. The section contains five papers on these topics in both business and social subject matter areas.

The Blum paper elaborates and defines the concepts of incorporating administrative data into the framework of edit and imputation, in particular for a multi-source data collection. The benefits from using multiple sources of data are outlined as well as the impact on the quality dimensions. Administrative records are used for enriching information to obtain better imputation models, for creating a reference file for error localization and for continuous quality assurance during the data processing. To obtain high quality output from the integration of several data sources, the administrative records need to be evaluated for their accuracy and completeness.

The Gasemyr paper also emphasizes the need for prior knowledge of the quality of each administrative data source in order to ensure overall high quality for the final records of the statistical data. Methods for editing a single administrative source are similar to survey data. However, when linking multiple data sources at the unit level, more errors and inconsistencies may occur, especially when data sources have common variables with conflicting values. The paper demonstrates methods for editing and imputation to obtain a complete job file based on linking multiple sources of administrative data.

The Shlomo paper is more technical than the previous papers. A probabilistic method is developed for linking multiple data sources at the unit level which ensures consistent and accurate records and less need for editing and imputation. The method determines the correct values of variables which a priori pass edit constraints, taking into account the quality of each data source. The method can also be used to enhance and improve the edit and imputation phase of survey data by enabling a cold deck imputation module for basic demographic and geographic variables from linked administrative data at the unit level. More sophisticated modeling can then be carried out on survey target variables and indicators.

The Laaksonen paper discusses the need for high level auxiliary data to improve the quality of survey data by exploiting all available information, especially important metadata which contain information about the coverage of the frame, non respondents, field investigation and other data processing stages. The three main tasks in editing and imputation according to the author are the following: model building for a pre-imputation phase in order to obtain preliminary values which help in the editing process; error localization; and imputation to obtain predicted values of missing and erroneous variables.

The final Mathews and Yung paper describes the use of administrative data in business surveys. Administrative data can replace some of the statistical data that are collected through surveys for some parts of the population. An example is shown on Statistics Canada's Annual Survey of Manufactures where administrative data is used throughout the entire data processing of the survey. An interesting feature of this survey is the need to develop a complete census for all units in the population using mass imputation procedures based on the sample and the administrative sources. An analysis is presented on the impact of combining administrative and survey data on the quality of the estimates and their variances.

- Olivia Blum "Evaluation of Editing and Imputation Supported by Administrative Files". A paper based on Prague 1997, WP16, tabled in Ottawa 2005, WP.7

- Svein Gasemyr: Editing and imputation for the creation of a linked micro file of base registers and other administrative data, Ottawa 2005, WP.8
- Natalie Shlomo: “The Use of Administrative Data in the Edit and Imputation Process”, Madrid 2003, WP30, an updated version tabled in Ottawa 2005, CRP.1
- Seppo Laaksonen: “Need for a High-Level Auxiliary Data Service to Improve the Quality of Editing and Imputation”, Helsinki 2002, WP.8
- Steve Matthews and Wesley Yung: The use of administrative data in estimation for the Annual Survey of Manufactures”, Ottawa 2005, WP.2

Chapter 4: Looking forward

- Leopold Granquist, John Kovar and Svein Nordbotten “Improving Surveys: Where does Editing Fit In?”

Abstract:

This paper considers a broad view of improving surveys as it relates to four groups of stakeholders: users, respondents, financiers and producers. From the user perspective, data quality is the main issue – fitness for use, including data quality in all its dimensions, but in particular, timeliness, accuracy, relevance, coherence, interpretability, and accessibility. From the respondents’ point of view, response burden is likely the greatest issue while from the treasury / taxpayers’ point of view, value for money, efficiency is the issue. From the statistical producers’ point of view, professional knowledge, modern equipment and efficient planning and production, employee job satisfaction, or from the negative side, misallocation of resources is a concern.

It can be argued that editing impacts on all of the above aspects. Traditionally we concentrated narrowly on issues of reducing nonresponse bias and hopefully correcting response and some processing errors. In this chapter we review what has been done, and what was planned but not done, in evaluating quality aspects of editing in the last decade or so, and argue that editing should be considered in a wider perspective, specifically as it should provide valuable information to other survey processes in a systematic way.
