Statistical Policy
Working Paper 18

Data Editing in Federal Statistical Agencies

Prepared by
Subcommittee on Data Editing in Federal Statistical Agencies
Federal Committee on Statistical Methodology

Statistical Policy Office
Office of Information and Regulatory Affairs
Office of Management and Budget

May 1990

MEMBERS OF THE FEDERAL COMMITTEE ON

STATISTICAL METHODOLOGY

(April 1990)

Maria E. Gonzalez (Chair)

Office of Management and Budget

Yvonne M. Bishop
Energy Information Administration

Daniel Kasprzyk
Bureau of the Census

Warren L. Buckler
Social Security Administration

Daniel Melnick
National Science Foundation

Charles E. Caudill
National Agricultural Statistical Service

Robert P. Parker
Bureau of Economic Analysis

John E. Cremeans
office of Business Analysis

David A. Pierce
Federal Reserve Board

Zahava D. Doering
Smithsonian Institution

Thomas J. Plewes
Bureau of Labor Statistics

Joseph K. Garrett
Bureau of the Census

Fritz J. Scheuren
Internal Revenue Service

Robert M. Groves
Bureau of the Census

Monroe G. Sirken
National Center for Health Statistics

C. Terry Ireland
National Computer Security Center

Robert D. Tortora
Bureau of the Census

Charles D. Jones
Bureau of the Census

Preface

The Federal Committee on Statistical Methodology was organized by OMB in 1975 to investigate issues in Federal statistics. Members of the committee, selected by OMB on the basis of their individual expertise and interest in statistical methods, serve in their personal capacity rather than as agency representatives. The committee conducts its

work through subcommittees that are organized to study particular issues and that are open to any Federal employee who wishes to participate in the studies. working papers are prepared by the subcommittee members and reflect only their individual and collective ideas.

The subcommittee on Data Editing in Federal Statistical Agencies was formed in 1988 to document, profile, and discuss the topics of data editing in Federal surveys. In preparing this report, the subcommittee walked in uncharted territory. Unlike many other survey process topics, such as design and estimators, where there is substantial literature, textbooks, and documentation, the formal literature pertaining to data editing is quite limited. It is hoped that this report will further the awareness within agencies of each other's data editing practices, as well as of the state of the art of data editing, and thus lead to improvements in data quality throughout Federal statistical agencies. A key ingredient in this effort is a profile of current data editing practices constructed from an editing questionnaire designed by the subcommittee and covering 117 Federal surveys. The report also describes current and recent research

developments that may aid agencies in evaluating their current data editing practices, as well as in planning for future data editing systems.

The subcommittee report is presented in a format and style that aims to increase awareness of Federal survey managers and subject matter specialists (statisticians, economists, computer programmers, statistical assistants, and clerks, etc.) on survey data editing practices. When possible, observations are made in this report that may aid in the evaluation of current editing practices and in the planning of future editing systems. In fact, this goal provided the subcommittee with the incentive to also investigate the methodology for software, technology, and research developments beyond the profile of current editing practices.

This subcommittee on Data Editing in Federal Statistical Agencies was chaired by George Hanuschak of the National Agricultural Statistics Service, U.S. Department of Agriculture.

MEMBERS OF THE SUBCOMMITTEE ON DATA EDITING IN

FEDERAL STATISTICAL AGENCIES

George Hanuschak, (Chair)

National Agricultural Statistics Service

Yahia Ahmed

Internal Revenue Service

Laura Bauer

Federal Reserve Board

Charles Day

Internal Revenue Service

Maria Gonzalez

Office of Management and Budget

Brian Greenberg

Bureau of the Census

Anne Hafner

National Center for Education Statistics

Gerry Hendershot

National Center for Health Statistics

Rita Hohenbrink

National Agricultural Statistics Service

Renee Miller

Energy Information Administration

Tom Petkunas

Bureau of the Census

David Pierce

Federal Reserve Board

Mark Pierzchala

National Agricultural Statistics Service

Marybeth Tschetter

Bureau of Labor Statistics

Paula Weir

Energy Information Administration

ACKNOWLEDGMENTS

This report represents an intensive voluntary effort on the part of dedicated subcommittee members and outside reviewers over an eighteen month period. It is truly a collective effort on the part of the subcommittee members, who worked very well as a team. While maintaining their full time Federal job responsibilities, the fifteen subcommittee members worked diligently on this challenging mission.

The subcommittee expresses its appreciation to Cathy Mazur of the National Agricultural Statistics Service, U.S. Department of Agriculture for her timely summarization of the current editing practices survey, to Dale Bodzer and Howard Magnus of the Energy Information Administration, and Cathy Cotton of Statistics Canada and John Monaco of the U.S. Bureau of the Census for demonstrating editing software packages to the subcommittee. Jelke Bethlehem of the Netherlands Central Bureau of Statistics and John Kovar of Statistics Canada provided substantial aid to the subcommittee by answering numerous questions about editing systems and providing software systems for subcommittee review.

The subcommittee extends its thanks to David Pierce, Federal Reserve Board, Terry Ireland, National Security Agency, and Fritz Scheuren, Internal Revenue Service of the Federal Committee on Statistical Methodology for their reviews of the report. The subcommittee extends its appreciation to Maria Gonzalez, Chair of the Federal Committee on Statistical Methodology, for her guidance, encouragement and advice throughout the eighteen months.

Last but not least. the subcommittee extends its appreciation to

Sherri Finks, Amy Curkendall, and Jennifer Kotch of the National Agricultural Statistical Service who diligently and patiently did the fine word processing using a desktop publishing package under rather severe time constraints.

TABLE OF CONTENTS

Page

Chapter 1. EXECUTIVE SUMMARY 1

A. Introduction 1

B. Key Findings 2

C. Recommendations 3

D. Implementation of Recommendations 4

E. Structure of Report 4

Chapter II. BACKGROUND 5

A. Scope, Audience and Objectives 5

B. Subcommittee Approach to Accomplishing Mission 5

C. Subcommittee Work Groups 7

D. Practices and Issues in Editing 7

Chapter III. CURRENT EDITING PRACTICES IN

FEDERAL STATISTICAL AGENCIES 11

A. Profile on Editing Practices 11

B. Case Studies 17

Chapter IV. EDITING SOFTWARE 21

A. Introduction 21

B. Software Improving Quality and Productivity 22

C. Descriptions of Editing Software 26

Chapter V. RESEARCH ON EDITING 31

A. Introduction 31

B. Areas of Edit Research 31

C. Editing Research in Other Countries 33

D. Case Studies 34

E. Summary 36

F. Bibliography 38

APPENDIX A: Results of Editing Practices Profile From Questionnaire Responses 39

APPENDIX B: Cast Studies 47

APPENDIX C: Checklist of Functions of Editing Software Systems 65

APPENDIX D: Annotated Bibliography of Articles on Editing 77

APPENDIX E: Glossary of Terms 87

CHAPTER I

EXECUTIVE SUMMARY

A. INTRODUCTION

The Subcommittee on Data Editing in Federal Statistical Agencies was

established by the Federal Committee on Statistical Methodology in November 1988 to document, profile and discuss data editing practices in Federal surveys. The subcommittee had the following mission statement:

> The objective is to determine how data editing is currently being done in Federal statistical agencies, recognize areas that may need attention and, if appropriate, to recommend any potential improvements for the editing process.

To accomplish its mission, the subcommittee first addressed the definition of data editing - what was it? No universal definition of survey data editing exists. The following working definition of editing was developed and adopted by the subcommittee:

> Procedure(s) designed and used for detecting erroneous and/or questionable survey data (survey response data or identification type data) with the goal of correcting (manually and/or via electronic means) as much of the erroneous data (not necessarily all of the questioned data) as possible, usually prior to data imputation and summary procedures.

Data editing can be seen as a data quality improvement tool by which erroneous or highly suspect data are found, and if necessary corrected. The subcommittee members realize that the boundaries of editing (where it begins and ends) is not absolute. The subcommittee was instructed by the Federal Committee on Statistical Methodology to concentrate on the front end of the editing process and not to duplicate the extensive work on imputation done by the Panel on Incomplete Data, Incomplete Data in Sample Surveys, Volumes I, II and III, Academic Press, 1983. Therefore, the rest of the document is based on the subcommittee's working definition of editing.

In order to gather the necessary information related to Federal survey editing practices, the subcommittee used a combination of information gathering techniques: a profile on editing practices using a subcommittee prepared questionnaire (6 pages and 41 questions) for 117 Federal surveys in 14 Agencies, an extensive literature search and review, case studies of 8 Federal surveys, editing system software evaluation for several recently developed generalized editing systems, and a search and review of current research efforts, including a few

case studies. on the editing process. These information-gathering techniques contributed to the development of an extensive editing information base for this report.

In summary, data editing is considered to be an important component of Federal statistical agencies. Key findings from the survey on editing practices conducted by the subcommittee follow, along with recommendations. In some cases. detailed discussions of recommendations are handled in the text of the document. A glossary of terms used in this report is in Appendix E.

B. KEY FINDINGS

Key findings from the profile on editing practices follow.

- About 60 percent of Federal survey managers reported that they refer all data that fad edit checks (not only critical

errors or severe outliers) to subject matter specialists or editors (economists, statisticians, clerks, etc.) for review and resolution. The role of the subject matter specialist is often valued as somewhat indispensable, as their expert knowledge and judgment are key ingredients in the survey editing process. Two key questions are:

1. at is the cost/benefit relationship of this extensive manual review?
2. How consistent are the actions of different subject matter specialists?

- Editing costs are reported to have a median value of about 35 percent of the total survey cost; however, the mode is 10 percent and the distribution is quite skewed to the right. Administrative records systems, such as those used by the Federal Reserve Board and the Internal Revenue Service, are on the skewed right-hand tail of the cost distribution. The reason is that compared to censuses or sample surveys, data collection costs are a much smaller

portion of total costs. This finding points to the importance of improvements in the cost efficiency of the editing process as a target for all Agencies in the next decade.

- Federal survey managers report that in over 80 percent of their surveys there is a good internal documentation of the editing system. Federal survey managers appear to take the editing process very seriously and recognize its importance in the overall survey performance. Under tight resource constraints, the level of documentation on the editing process is impressive.

- There is a strong desire by many of those involved in the editing process to combine or replace "batch-oriented" systems with "on-line" or quick-turnaround systems.

- Another desire expressed by respondents is for continued research and development and implementation of more efficient, well targeted, consistent, and accurate methods to detect potentially erroneous survey data.

- Integration of survey tasks (e.g., computer-assisted data collection, data entry, data editing, imputation, and summary) is important for improving data quality and productivity in survey processing.

- Several major developments in generalized editing software are taking place in domestic and international statistics agencies. Three major ones covered in some detail in this report (Chapter IV) are: Statistics Canada's Generalized Edit and Imputation System (GEIS): the Netherlands Central Bureau of Statistics Blaise system (named after Blaise Pascal, the well known mathematician of the 1600's); and the U.S. Census Bureau's Structured Program for Economic Edit and Referral (SPEER).

- Some agencies are currently conducting research on the editing process, and several case

studies are presented in Chapter V and Appendix B.

- There is substantial potential for several related technology and data systems developments to contribute to more efficient and consistent editing systems in the next decade. These include data base systems, expert systems, electronic data collection such as computer assisted telephone interviewing (CATI), computer-assisted personal interviewing, (CAPI) and touchtone surveys, major generalized edit systems, and artificial intelligence systems.

- If the cost of data processing continues to drop at its current rapid pace, the analysis of multi-variate statistical relationships among survey variables can be more widely used for editing (and imputation) if appropriate.

- The major challenge in software development lies in the

reconciliation of two goals: the increased use of computers for certain tasks and the more intelligent use of human expertise.

C. RECOMMENDATIONS

Based upon the findings in the Subcommittee's editing information base, we present the following recommendations.

Federal survey managers and administrators should:

- Evaluate and examine the cost efficiency, timeliness, productivity, repeatability, statistical defensibility, and accuracy of their current editing practices versus alternative systems. The checklist of editing software features provided in Appendix C and the remainder of this report is an aid in such an effort. Such an effort can also be part of a Total Quality Management system for surveys and agencies.

- Review and examine the implications, for their editing situation, of important developments in data processing such as powerful microcomputers and scientific workstations, local area networks (LAN's) and data base software that provide electronic communication links from microcomputers and LAN's to mainframe computers.

- Follow the research and applications developments in the use of CATI, CAPI, touchtone, and other electronic means of data capture with potential for improving the editing and/or data processing flow.

- Continue to share information on research and development and software systems efforts in the editing process with other Federal and international statistical agencies.

- Stay attuned to research and developments in the use of expert systems and/or artificial intelligence software for survey data editing.

- Evaluate both the role and the effectiveness of editing in reducing nonsampling errors for their surveys.

- Evaluate the extensive relationship of extensive manual review on resulting estimates.

- Explore development of a catalog of situations in which various techniques work well or not; e.g., research has indicated that exponential smoothing does not work well when data are erratic.

- Recognize the value of editing research and place a high priority on devoting resources to their own research, to monitoring developments in data editing at other agencies and to implementing improvements when they are found to be desirable.

- Explore integration of functions in a survey; e.g., data entry, data editing and computer assisted data collection.

- Give attention to the future roles of the subject matter specialist and the methodologist and to the tools and consistency with which they perform their jobs.

D. IMPLEMENTATION OF RECOMMENDATIONS

An interagency working group should be formed to continue the mission of the subcommittee and work on the implementation of the subcommittee's recommendations.

E. STRUCTURE OF REPORT

The structure for this report is executive summary, followed by Chapter II which is introductory, Chapter III on the editing profile and the case studies, Chapter IV on the role of software in editing, and Chapter V on the role and status of research in editing.

Supporting appendices include:

A. Results of Editing Practices Profile From Questionnaire Responses

B. Case Studies

C. Checklist of Functions of Editing Software Systems

D. Annotated Bibliography of Articles on Editing

E. Glossary of Terms

CHAPTER II

BACKGROUND

A. SCOPE, AUDIENCE, AND OBJECTIVES

The Subcommittee on Data Editing in Federal Statistical Agencies was established by the Federal Committee on Statistical Methodology in November 1988 to document, profile and discuss data editing practices for Federal surveys. The Subcommittee had the following mission statement:

> The objective is to determine how data editing is currently being done in Federal statistical agencies, recognize areas that may need attention and, if appropriate, to recommend any potential improvements for the editing process.

> The project will obtain information on current data editing practices. The information on editing will include the role of subject matter specialists; hardware, software, and data base environment; new technologies of data collection (and editing) such as CATI and CAPI. and current research efforts in the Agencies and some recent developments, in generalized editing systems, from the U.S. Census Bureau, Statistics Canada, and the Netherlands Central Bureau of Statistics.

B. SUBCOMMITTEE APPROACH TO ACCOMPLISHING MISSION

A number of paths were followed by the subcommittee in accomplishing its goals as set forth in the preceding mission statement, including developing a questionnaire on survey editing practices. assembling several detailed case studies, investigating alternative editing systems and software, exploring research needs and practices, and compiling an annotated bibliography of literature on editing.

The editing profile questionnaire (6 pages and 41 questions) was developed and administered to 117 Federal surveys covering 14 different agencies. The 117 surveys were selected by Subcommittee members and thus were not a scientific sample of all Federal surveys. The subcommittee members felt that the 117 surveys represented a broad coverage of agencies and types of surveys or censuses that might have different editing circumstances or situations.

The two major purposes of the editing questionnaire were to provide an adequate profile of current editing practices and to aid in

developing a typology of surveys to be used for selecting case studies. The typology is a classification of surveys according to a number of criteria such as frequency of the survey, number of respondents. degree of automation and judgmental review of the edits, whether respondents are contacted regarding questionable items, whether historic data are used in the editing of current data, and so forth. This information is of general interest, and was useful to the subcommittee in selecting a representative group of surveys to serve as case studies

of editing practices. The questionnaire and a tabular summary of the results are presented for the reader in Appendix A.

Chapter III of this report contains the analysis of the questionnaire and a description of the case studies. For each different editing environment, a case study was conducted. The case studies provide more detailed information for the selected cases than just the editing questionnaire . The case studies are published in two forms (long and

short) in Appendix B to give descriptions of the varied editing practices and situations.

Another important area of the subcommittee's work was the investigation and evaluation of some recently developed generalized editing systems and software packages. Several major editing systems were studied and a profile of features was developed and is presented in Chapter IV. The editing systems reviewed were the U.S. Census Bureau's Structured Program for Economic Edit and Referral (SPEER), the Netherlands Central Bureau of Statistics Blaise system, and Statistics Canada's Generalized Edit and Imputation System (GEIS). Also, several recent application specific editing systems at the U.S. Department of Energy and the Bureau of Labor Statistics were reviewed. These systems were developed under different conditions and applications, so direct comparisons are not feasible. However, the subcommittee believes that a description of these systems features and capabilities is of substantial value to Federal statistical agencies. Appendix C gives the reader a detailed checklist of editing software system features. This checklist will be a valuable tool to editing system developers.

The remaining major activity of the subcommittee was a review of historic and ongoing research. This review consisted of a literature search that enabled the subcommittee to develop an annotated bibliography, presented in Appendix D. This appendix provides a valuable source of information on editing literature. In addition, case studies of ongoing or recent editing research were conducted. Also information about editing research and research needs on the editing process were gleaned from the editing profile. A more detailed description of editing research is provided in Chapter V. A short glossary of editing terms is given in Appendix E.

C. SUBCOMMITTEE WORK GROUPS

To effectively accomplish its mission, the subcommittee was divided into four major groups.

I. Editing Profile Group

Charles Day, Leader

Yahia Ahmed

George Hanuschak

Rita Hohenbrink

Renee Miller

II. Case Studies Group

Anne Hafner, Leader

Yahia Ahmed

III. Editing Software Group

Mark Pierzchala, Leader

Charles Day

Gerry Hendershot

Rita Hohenbrink

Tom Petkunas

Marybeth Tschener

IV. Editing Research Group

Brian Greenberg, Leader

Yahia Ahmed

Laura Bauer

Renee Miller

David Pierce

Paula Weir

D. PRACTICES AND ISSUES IN EDITING

Description of the Process

Pre-survey editing tasks include the writing and evaluation of editing programs, evaluation of tile edits themselves, and writing instructions for the inspection of questionnaires by interviewers. field supervisors, clerks, and subject matter specialists. These activities influence how well editing is done, as well as how many resources will be expended on editing once data are collected. During the survey itself, editing may occur in many ways and at many stages, from data collection to publication, and even after publication in some cases.

In paper and pencil interviewing, the interviewer is the first to inspect the questionnaire for errors. Optimally, this should be done immediately after the interview so that the respondent can easily be contacted to clarify responses. If questionnaires are channeled through a supervisor, then a second

inspection can be done. Not only can recontacts be made shortly after the interview, but the supervisor can provide feedback to the interviewers on how they are doing. Once the questionnaires reach an office, they may be edited manually by clerks, subject matter specialists, or both. In some organizations, this manual edit may include a significant amount of coding. It can also include a visual check that answers are given in correct units, that routing instructions have been followed correctly, and consideration of notes written by either the respondent or the enumerator. In most cases a computer edit is then performed. Error signals (flags) and messages are presented to a reviewer either on printouts or a screen in an

interactive session. If the program output is printed, then the review tends to be cyclical as the computer must then re-edit, in batch, all of the changes. If the output is on a screen, (microcomputers or terminals hooked to a larger computer), then questionnaires are usually processed one at a time until they pass review.

All of the above editing activities relate to reviewing data at the record (or questionnaire) level. This is often referred to as micro-editing. Editing of data at an aggregate level will then take place even if it is not explicitly recognized as such. This macro-editing may be by cells in an economic table, or by some other aggregation such as a stratum. The cells in a table may be edited against themselves (one can visualize some sort of super-questionnaire) or against similarly defined cells from previous surveys. This macro-editing may be done by hand or through specially designed software. Depending on the degree of automation, it may or may not be possible to trace inconsistencies at the aggregate level to the offending questionnaires. If the macro-editing program can trace inconsistencies back to the micro-level, then macro-editing can in theory be used to direct the micro-editing.

If Computer Assisted Data Collection is used, then much of the editing process is formally introduced and enforced at the time of data collection. Not only are most major errors corrected at the time of the interview. but the subject matter specialists may have greater confidence in the data after collection and be more likely to let the data pass without change. Thus, Computer Assisted Data Collection has enormous potential for reducing the costs of data editing after data collection. By introducing edits into data collection, it will also improve the data themselves. Currently, Computer Assisted Data Collection is becoming more common in the survey world. However, for the foreseeable future, many surveys will still be collected by mail or by paper and pencil interviewing. In any case, the need for editing after data collection will never be totally eliminated.

Issues in Editing

Costs and Benefits

The importance of data editing in reducing non-sampling errors has

been questioned. Granquist (1984) questions whether the editing process can essentially improve data quality after data are collected. He states that there are three purposes of editing: to give more detailed information about the quality of the survey, to provide basic data for the improvement of the survey. and to tidy up the data so that further processing can be done. Further, Granquist considers the sources and types of survey errors, and questions the ability of most generalized editing systems to address all kinds of errors including systematic response errors. If data are considered to have a timely quality. that is, the value of data deteriorate as time goes along, then editing can reduce die value of the data. Pullum, Harpham, and Ozsever (1986) describe a situation where the editing of survey data

had no discernible effect on the estimates other than to delay their release by about one year. One common question that many organizations have is when to declare that editing is finished.

"Over-editing" of data has long been a concern. In order to make sure that all errors are flagged, often many unimportant error signals (flags) are generated. These extra signals not only take time to examine but also distract the reviewer from important problems. These extra signals are generated because of the way that error limits are set. One way that researchers are trying to reduce the number of error signals, while at the same time ensuring that the important cases are flagged, is through the development of statistical editing techniques. For example, time series techniques can be used in repetitive surveys on a record-by-record basis. Alternatively, cross record statistical editing can be done on either a univariate or multivariate basis. This may include the graphical inspection of data.

Data editing often requires considerable resources in federally conducted surveys both in terms of staff time and dollars. These expenditures are themselves reason enough to re-evaluate the editing process. In addition, there are often external economic incentives in the form of reduced budgets for statistical agencies. The

combination of rapidly decreasing computing costs, rapidly increasing computing capabilities, and steady or increasing staff costs, is changing the economics of the process vis-a-vis the proper mix of human and computer processing. Another cost that is not considered much is the increase in respondent burden. In some surveys, edits are so tightly set that few if any records pass the edits. As a result respondents are called back, some many times, in order to clear up suspicious situations. There is also an opportunity cost to editing. Any time spent in editing is time that is not being used for other activities, some of which may have greater potential for reducing nonsampling errors.

Statistical and Quality Concerns

Statistical considerations will impact the development of new editing systems and may even lead to their development. Defensibility of the process is a concern because data are changed after data collection and before summary. The ability of an agency to defend itself from criticism is enhanced by implementing methodologically sound procedures, by capturing data electronically as they are reported, and by auditing all changes made during the edit. The effect of editing

can then in principle be known, and feedback for the improvement of the survey can be given. Conceptually, the edit process should be repeatable (or reproducible). This means that the same data run through a system twice should lead to the same results. Editing should not change survey results in an unpredictable manner.

Integration of Survey Tasks

Integration of survey tasks is important for improving both data quality and productivity in survey processing. Consider the functions of Computer Assisted Data Collection, data entry, and data editing. By integrating these functions, data quality can be improved by injecting the editing function into collection, and also by reducing transcription errors by eliminating the need for in office data entry. Given the proper software, pre-survey activities may be done more productively by reducing the need for multiple specification of the data. For example, if a particular variable can take only the values of 1, 2, and 3, then the program for each of the three functions should have

specified this limitation. Time is saved, and potential for inconsistencies are reduced, if all three programs derive from one specification. Usually, routing instructions and edits are common between a data collection instrument and an editing program. If both functions derive from the same program, then double programming can be eliminated. Also it is easier to make more explicit and purposeful the differences between the collection and editing instruments.

Constraints

Constraints (other than economic constraints already considered) on the organization or on the survey itself often adversely affect the quality of editing. Some large federal surveys (e. g., in the National Agricultural Statistics Service or in the Bureau of Labor Statistics) are conducted under extremely tight deadlines. Data are collected, processed, analyzed, and published under a rigid schedule.

Detecting all of the major problems in such a large data set under the time allowed becomes enormously difficult. Computer hardware and software have their own constraints. For example, access to a mainframe may be limited and editors may have to review error signals on paper printouts because of costs. Software, may not be easy to use, and it may be extremely difficult to coordinate disparate programs. Data editors may not have sufficient knowledge of the subject matter or survey procedures nor sufficient training. High turnover of editors may be a problem in some surveys. The challenge then is in providing the inadequate staff with enough, effectively presented information to allow the job to be done correctly.

There may be resistance to change or a questioning of its need in the implementation of new editing systems and methodologies. People may wonder how their job will be changed or if it will be eliminated. Some problems may be easy to identify (e.g., the amount of resources consumed is too large) but others may require special studies (e. g., how much is spent on each task and how much do we get from it?). In considering either the development of a new editing system or the purchase of one, it is often difficult to know which editing system

features are necessary, and their relative performance. Evaluation of editing software is difficult and time consuming. Another consideration is who should be on the evaluation team.

CHAPTER III

CURRENT EDITING PRACTICES

A. PROFILE ON EDITING PRACTICES

To obtain an adequate profile of current editing practices, the subcommittee developed an editing profile questionnaire which was administered to 117 Federal surveys covering 14 different if agencies. The 117 surveys were selected by Subcommittee members and thus were not a scientific sample of all Federal surveys; however, the subcommittee members felt that the 117 surveys represented a broad coverage of agencies and types of surveys or censuses that might have different editing situations.

This section describes how the questionnaire was designed and administered, and summarizes the findings. While this section focuses on the highlights of the profile, tallies of responses to all of the questions appear in Appendix A.

Designing and Administering the Questionnaire

The subcommittee designed a six-page questionnaire containing general descriptive questions about a particular survey as well as specific questions on editing practices. (See Appendix A for a copy of the questionnaire) Each subcommittee member pretested the editing questionnaire by answering the questions for a survey with which each was familiar. Although a scientific sample was not drawn, the goal was to select a group of surveys that would be representative of the surveys conducted by Federal statistical agencies. Each subcommittee member sought to obtain information for ten to twenty surveys that represented their agencies surveys. In addition, they obtained information from several agencies not represented on the subcommittee.

Some subcommittee members reviewed the completed questionnaires for consistency by contacting the agency respondents prior to submitting them. A small number of consistency edit checks were performed for the questionnaires; however, the editing was limited.

Characteristics of Surveys in Editing Practices Study

Illustrating the wide range of surveys in the study, large and small surveys were represented. The smallest survey in the study had 22 units, while the largest had about 1 million units. As shown in Figure 1:

About three-fourths of the surveys in the study are sample surveys.

Various frequencies of collection are represented (annual, quarterly, monthly, and weekly).

About three-quarters of the surveys are filed by establishments, institutions, farms and other entities, and the remaining quarter by households or individuals.

Traditional means of data collection such as mail, personal, and telephone interviews were the most common. Only a small proportion used computer assisted telephone interview (CATI); and no survey respondents reported using computer-assisted personal interview (CAPI) as their primary method of data collection, although a few did report using CAPI as a secondary method. About sixteen percent of the surveys in the study use administrative records.

The remainder of this section discusses editing practices. As part of the analysis, data on editing practices were cross-classified by the characteristics just discussed (sample versus census, frequency of collection, type of unit surveyed and mode of collection) to determine whether editing practices varied by these characteristics. If these characteristics do in fact affect editing practices, and the surveys in the study are not representative of all surveys on these

characteristics, then the aggregated results of this study would not be applicable to all surveys. Results are presented for all of the surveys in this study, but situations in which editing practices differed greatly are highlighted.

Editing Practices

The questionnaire covered the following areas with respect to editing practices: cost of editing, when and how editing occurs. type of edits used, degree of satisfaction with current system, and future applications.

Cost of Editing

The survey respondents were instructed to include all aspects of editing in their cost figures, such as edits made at data entry. clerical work, computer time, design, testing, monitoring, analyst review, call-backs, and summary table review. However, in all of the following information on editing costs, about two-fifths of the respondents reported that information on the cost of data editing was available. The subcommittee does not claim that this data is totally

free of nonsampling errors. Therefore. all conclusions are subject to this constraint. Editing costs representing at least 20 percent of the total cost of the survey were reported for four-fifths of these surveys. A similar pattern was observed for the surveys for which cost information was not available. Of the 73 surveys where no cost data were available, cost estimates were provided for about two-thirds (49 surveys). About three-quarters of these surveys had editing costs representing at least 20 percent of the total cost of the survey.

The median editing cost as a percentage of the total survey cost was 35 percent. While an attempt was made in the instructions to the survey to standardize the activities to be included in the cost of editing (see question 20 in Appendix A), record-keeping practices vary. As a result, estimates may not represent the same activities from survey to survey. However, the data still proved useful in determining the survey characteristics that most effect the cost of editing.

Editing costs as a percentage of the total survey cost varied greatly by the type of survey. Demographic surveys (surveys of individuals

and households) had a far lower median than economic surveys (surveys of farms, establishments or firms, institutions, and others). The median for demographic surveys was 20 percent compared with 40 percent for economic surveys. Among the economic surveys, the surveys that used administrative records had the highest median of percentage of cost, 60 percent. This high percentage does not necessarily indicate a high absolute editing cost, it could indicate a low total survey cost because no new survey data are collected. As discussed in the next section, these surveys have more extensive involvement of subject matter analysts than demographic surveys have.

Overall, surveys in which all error correction was done by clerks or analysts were more likely to have editing costs that represent over 40 percent of the total survey cost than were surveys in which only unusual cases were referred to analysts . Almost one-half of the former group had editing costs in the category "40 percent or greater"

compared with one-third of the latter group. Reversing the perspective, only 6 percent of the former group (all error correction by clerks or analysts) had editing costs in the category, "under 10 percent" compared with about one-third of the latter group (unusual cases by analysts).

When and How Editing and Follow-up Occur

For about two-thirds of the 117 surveys studied, the majority of data editing takes place after data entry. Subject matter analysts play a large role in almost all of the surveys. In about three-quarters of the surveys, subject matter analysts review all unusual or large cases after automated or clerical editing. Only seven surveys have little or no intervention from subject matter specialists. Of these, only four are completely automated (i.e., edit checking and error correction are done without referral to analysts).

Surveys of farms, establishments and institutions tend to have heavier involvement from subject matter analysts than surveys of individuals and households (i.e., higher proportions of the study respondents

report that all data editing is done by subject matter analysts for these surveys than for the others). This could explain the relatively higher editing costs as a percentage of the total survey cost reported for the surveys of farms, establishments and institutions.

The degree of automation varies considerably among the surveys in the study. About three-fifths of the survey managers note that automated edit checking is done, but error correction is performed by clerks or analysts (Figure 2). In about 62 percent of the cases, there is no analysis of the effect of editing practices on the estimates produced.

Types of Edits

Almost all the surveys in the study use validation editing which detects inconsistent data within a record. A large proportion (83 percent) also use macro-editing where aggregated data are examined to detect inconsistencies. In addition to these two types of edits, 57 percent of the survey managers report using other edits.

In response to an open-ended request to describe "other" edits, "range edits" were mentioned most frequently. followed by procedures that

used historical data. "Ratio edits" were another common response. These three groups may not be distinct. Because responses we re not detailed, it was difficult to determine exactly what these edits involved. Other types of edits and analyses mentioned include: comparisons with other surveys. comparing the current value to a value estimated by regression analysis, using interquartile measures, and listing the ten highest and ten lowest values before and after expansion factors were applied.

Satisfaction With Current Edit System

The study respondents were split in the level of satisfaction with their current system. About 47% were satisfied, while about half thought that at least minor changes were required. A small proportion said it was not possible to determine what changes were required at this point (Figure 4).

Future Applications

Among those expressing a need for change, an on-line system topped the list of desired improvements. Other changes that were mentioned frequently (as a result of an open-ended question) included:

- The use of prior years' data to test the current year,
- More statistical edits, and
- More sophisticated and more extensive macro and validation editing.

An audit trail, more automation in general. and a user-friendly system were also mentioned several times. In addition, the following enhancements were mentioned: automated error correction, incorporation of imputation into the editing package. evaluation of the effect of data editing, reduction of the number of edit flags to follow-up, incorporation of information on auxiliary variables, multivariate editing. use of an expert system approach for criteria which require judgment, and editing using micro-computers.

In summary our questionnaire revealed wide diversity in current editing practices and in user satisfaction with them. To present more of an in-depth picture, we now describe the development of the case studies.

B. CASE STUDIES

Federal government surveys, censuses, and administrative records systems create a broad range of data editing situations. In addition to the statistical profile on editing practices, it was felt that a further description of several of the surveys in case study form would reveal in greater detail the complexity of the different editing practice situations in operation.

A typology of editing situations was developed by the subcommittee to be used for selecting case studies (Figure 5). The typology was developed through extensive subcommittee discussion and from analysis of responses to the editing practices questionnaire.

The grouping variables included in Figure 5 are:

1. Census or sample survey approach

2. Longitudinal or cross sectional approach

3. Frequency of census or sample survey

4. Size of census or sample survey

5. Continuous and/or categorical data

6. Administrative records used (Yes or No)

7. Mode(s) of data collection used (mail, telephone, CATI, CAPI, touchtone, personal, etc.)

8. Use of historical data in the edit process (Yes or No)

There were also other grouping variables that were considered and then

discarded, for example, the level of clerical knowledge of subject matter when editing. The major reason for elimination was subjectivity involved in measuring those variables.

In order to represent the range of different editing situations, the subcommittee picked eight case, studies that covered the different values of the eight grouping variables. Four were chosen to develop brief case studies which represent different survey situations and are presented in short abstract form in Appendix B. These are:

- BLS:CPI: Commodities and Services

- IRS: US Corporation Income Tax Returns

- NCES: National Education Longitudinal Study of 1988

- Federal Reserve Board: Edited Deposits Data System

The first paragraph of each abstract describes the environment in which the survey takes place, including type of survey and size. The second paragraph includes a brief description of editing practices used.

Four additional surveys are described in greater detail (in Appendix B) to give the reader a flavor of the range of editing practices and situations. Surveys covered are:

- NCHS: National Health Interview Survey;
- Census: Enterprise Summary Report;
- NASS: Quarterly Agricultural Survey; and
- EIA: Monthly and Weekly Refinery Reports.

The first section of the in-depth case studies describes the

environment in which the survey takes place. The second section describes editing practices - used, including data processing environment, audit trail, micro, macro and statistical editing, prioritizing of edits, imputation process, standards, costs, role of subject matter specialists, measures of variation, and current and future research.

The wide variation in editing situations makes it impossible to recommend any one editing system or methodology for all Federal statistical agencies, surveys, administrative records systems, or censuses.

CHAPTER IV

EDITING SOFTWARE

A. INTRODUCTION

For most surveys, large parts of the editing process are carried out

through the use of computer systems. The task of the Software Subgroup has been to investigate software that in some way incorporates new methodologies, has new ways of presenting data, operates in recently developed hardware environments, or integrates editing with other functions. In order to fulfill this charge, the Subgroup has evaluated or been given demonstrations of new editing software. In addition, the Subgroup has developed an editing software evaluation checklist that appears in Appendix C. This checklist contains possible functions and attributes of editing software, which would be useful for an organization to use when evaluating editing software.

Extremely technical jargon can be associated with new editing systems; and new approaches to editing may not be familiar to the reader. The purpose of section B is to explain these approaches and their associated terminology as well as to discuss briefly the role of editing in assuring data quality.

A distinction must be made between generated systems and software meant for one or a few surveys. The former is meant to be used for a variety of surveys. Usually there is an institutional commitment

to spend staff time and money over several years to develop the system. It is hoped that the investment will be more than recaptured after the system is developed through the reduction in resources spent on editing itself and in the elimination of duplication of effort in preparing editing programs. Some software programs have been developed that address specific problems in a particular survey. While the ideas inherent in this software may be of general interest, it may not be possible to apply the software directly to.other surveys. Section C describes three generalized systems in some detail, and then briefly describes other systems and software. These three systems have been used or evaluated by Subgroup members in their own surveys.

New and exciting statistical methodology is also improving the editing process. This includes developments in detecting outliers, aggregate level data editing, imputation strategy, and statistical quality control of the process itself. These activities are covered more fully in Chapter V. The Implementation of these activities, however, requires that the techniques be encoded into a computer program or system.

B. SOFTWARE IMPROVING QUALITY AND PRODUCTIVITY

Reasons for the Development of New Editing Software

Traditional editing systems do not fully utilize the talents or expertise of subject matter specialists. Much of their time may be spent in dealing with unimportant or spurious error signals and in coping with system shortcomings. As a result, the specialist has less time to deal with important problems. In addition, editing systems may be able to give feedback on the survey itself. For example, a pattern of edit failures may suggest misunderstandings by the respondent or interviewer. If this is recognized, then the expertise of the specialist may then be used to improve the survey itself.

Labor costs are a large part of the editing costs and are either steady or increasing, whereas the cost of computing is decreasing.

In order to justify the heavy reliance on people in editing, their productivity will have to be improved through the use of more powerful tools. However, even if productivity is improved, different people may do different things in similar situations. If so, this makes the process less repeatable (reproducible) and more subject to criticism. When work is done on paper, it is hard to track, and it is impossible to estimate the effect of editing actions on estimates. Finally, some tasks are beyond the capability of human editors. For example, it may be impossible for a person to maintain the multivariate frequency structure of the data when making changes.

These reasons and several others are commonly given as explanations for the increased use of computer software to improve the editing process. It is in the reconciliation of these two goals, (the increased use of computers for some tasks and the more intelligent use of human expertise), that the major challenge in software development lies. There will always be a role for people, but it will be modified. One positive feature of new editing software is that it can often improve the quality of the editing process and productivity at the same time.

Ways that Productivity Can Be Improved

One way to improve productivity is to break the constraints imposed by computer systems themselves. The use of mainframe systems for editing data is widespread. In some cases, however, an editor may not use the system directly. For example, error signals may be presented on paper printouts, and changes entered by data typists. Processing costs may dictate that editing jobs are run at low priority, overnight. or even less frequently. The effect of the changes made by the editor may not be immediately known: thus, paper forms may be filed, taken from files, and refiled several times.

The proliferation of microcomputers promises to eliminate many of these bottlenecks. while at the same time it creates some challenges in the process. The editor will have direct access to the computer, and will be able to prioritize its use. Once the microcomputer is acquired. user fees are eliminated, thus resource-intensive programs such as interactive editing can be employed, provided the microcomputers are fast enough. Moving from a centralized environment (i. e., the main frame) to a decentralized environment (i. e.,

microcomputers) will present challenges of control and consistency. In processing a large survey on two or more microcomputers,

communications will be necessary. This will best be done by connecting them into a Local Area Network (LAN).

New systems may reduce or eliminate some editing tasks. For example, where data are edited in batch and error signals are presented on printouts, a manual edit of the questionnaires before the machine edit may be a practical necessity. Editing data and error messages on a printout can be a hard, unsatisfactory chore because of the volume of paper and the static and sometimes incomplete presentation of data. The purpose of the manual edit in this situation is to reduce the number of machine-generated error signals. In an interactive environment, information can be efficiently presented and immediately processed. The penalty associated with machine-generated signals is greatly reduced. As a result, the preliminary manual edit may be eliminated. In addition, questionnaires are handled only once,

further reducing filing and data entry tasks.

Productivity may be increased by reducing the need for editing after data are collected. Instruments for Computer Assisted Telephone Interviewing (CATI), Computer Assisted Personal Interviewing (CAPI), and on-site data entry and editing programs are gaining wider use. Routing instructions are automatically followed, and other edit failures are verified at the time of the interview. There may still be many error signals from suspicious edits, however, the analyst has more confidence in the data and is more likely to let them pass.

There are two major ways that productivity can be improved in the programming of the editing instruments. First is to provide a system that will handle all, or an important class, of the agency's editing needs. In this way the applications programmer need not worry about systems details. For example, in an interactive system, the programmer does not have to worry about how and where to flag edit failures as it is already provided. The programmer only codes the edit specification itself. In addition, the end-user has to learn only one system when editing different surveys. Second is the

elimination of multiple specification and programming of variables and edits. For example, if data are collected by CATI, and edited with another system, then essentially the same edits will be programmed twice, possibly by two sets of people. If the system integrates several functions, e. g., data entry, data editing, and computer assisted data collection, then one program may be able to handle all of these tasks. This integration would also reduce time spent on data conversion from one system to another.

Systems that Take Editing and Imputation Actions

Some edit and imputation systems take actions usually reserved for people. They choose fields to be changed and then change them. The human element is not removed. rather this expertise is incorporated into the system. One way to incorporate expertise is to use the edits themselves to define a feasible region. This is the approach outlined in a famous article by Fellegi and Holt (1 976). Edits that are explicitly written are used to generate implied edits. For example if $100 < x/y < 200$, and $3 < y/z < 4$, are explicit edits. then an implied edit obtained algebraically is $300 < x/z < 800$. Once all

implied edits are generated the set of complete edits is defined as the union of the explicit and implied edits. This complete set of edits is then used to determine a set of fields to be changed for every possible edit failure. This is called error localization. An essential aspect to this method is that changes are made to as few fields as possible, or alternatively, to the least reliable set of fields which are determined by weights given to each field.

The analyst is given an opportunity to evaluate the explicit edits. This is done through the inspection of the implied edits and external records (the most extreme records that can pass through the edits without causing an edit failure). In inspecting the implied edits, it may be determined if the data are being constrained in an unintended way. In inspecting external records, the analyst is presented with combinations of the most extreme values possible that can pass the edits. The human editor has several ways to inject expertise into this kind of a system: (1) the specification of the

edits; (2) the action of implied edits and external records and then the respecification of edits; (3) the weighting of variables according to their relative reliability.

There are some constraints in systems that allow the computer to take editing actions. Fellegi and Holt systems cannot handle certain kinds of edits, notably nonlinear and conditional edits. Also algorithms that can handle categorical data cannot handle continuous data and vice versa. Within these constraints (and others), most edits can be handled. For surveys with continuous data, a considerable amount of human attention may still be necessary, either before the system is applied to data or after.

Another way that computers can take editing actions is by modeling human behavior. This is the expert system" approach. For example, if typically maize yields average 100 bushels per acre, and the value 1.000 is entered. then the most likely correction is to assume that an extra zero was typed. The computer can be programmed to substitute 100 for 1,000 directly and then to re-edit the data.

Ways that Data Quality can be Improved or Maintained

It is not clear that editing done after data collection can always improve the quality of data by reducing non-sampling errors. An organization may not have the time or budget to recontact many of the respondents or may refrain from recontacts in order to reduce respondent burden. Additionally, there may be cognitive errors or systematic errors that an edit system cannot detect. Often, all that can be done is to maintain the quality of the data as they are collected. To use the maize yield example again, if the edit program detects 1,000 bushels per acre, and sets the value to 100 bushels per acre, then the edit program has only prevented the data from getting worse. Suppose the true value was really 103 bushels per acre. The edit and imputation program could not get the value closer to the truth in this case. Detecting outliers is usually not the only problem. The proper action to take after detection is the more difficult problem. One of the main reasons that Computer Assisted Data Collection is employed is that data are corrected at the time of collection.

There are a few ways that an editing system may be able to improve data quality. A system that captures raw data, keeps track of changes, and provides well conceived reports, may provide feedback on the performance of the survey. This information can be used to improve the survey in the future. To take another agricultural example, farmers often harvest corn for silage (the whole plant is harvested, chopped into small pieces, and blown into a silo). Production of silage is requested in tons. Farmers often do not know their silage production in tons. Instead, the farmer will give the size (diameter and height) of all silos containing silage. In the office, silo sizes are converted into tons of production. If this conversion takes place before data are entered, then there is no indication from the machine edit of the extent of this reporting problem.

Another way that editing software can improve the quality of the data is to reduce the opportunity cost of editing. The time spent on editing leaves less time for other tasks, such as persuading people

to participate, checking overlap of respondents between multiple frames, and research on cognitive errors.

Ways that Quality of the Editing Process can be Defended or Confirmed

There is a difference between data quality and the quality of the editing process itself. To refer once again to the maize yield example, a good quality process will have detected the transcription error. A-poor quality process might have let it pass. Although neither process will have improved data quality, the good quality process would have prevented their deterioration from the transcription error. Editing and imputation have the potential to distort data as well as to maintain their quality. This distortion may affect the levels of estimates and the univariate and multivariate distributions. A high quality process will attempt to e distortions. For example, in Fellegi and Holt systems, changes to the data will be made to the fewest fields possible and in a way such that distributions are maintained.

A survey organization should be able to show that the editing process is not abusing the data. For editing after data collection, this may

be done by capturing raw (unedited) data and keeping track of changes and the reasons for change. This is called an audit trail. Given this record keeping, it will be possible to estimate the impact of editing and imputation on expansions and on distributions. It will also be possible to determine the editor effect on the estimates. In traditional batch mode editing on paper printouts, it is not unusual for two or more specialists to edit the same record. For example, one may edit the questionnaire before data entry while another may edit the record after the machine edit.. In this case, it is impossible to assign responsibility for an editing action. In an on-line mode one person handles a record until it is done. Thus all changes can be traced to a person. For editing at the time of data collection, (e.g., in CATI), it may be necessary to conduct an experiment to see if either the mode of collection, or the edits employed, will lead to changes in the data.

A high quality editing process will have other features as well. For example, the process should be repeatable, in time and in space. This means that the same data passed through the same process in two different locations, or twice in one location, will look (nearly) the

same. The process will have recognizable criteria for determining when editing is done. It will detect real errors without generating too many spurious error signals. The system should be easy to program in and have an easy user interface. It should promote the integration of survey functions such as micro- and macro-editing. Changes made by people should be on-line (interactive) and traceable. Database connections will allow for quick and easy access to historical and sampling frame data. An editing system should be able to take actions of minor impact without human intervention. It should be able to accommodate new advances in statistical editing methodology. Finally, quality can be promoted by providing statistically defensible methods and software modules to the user.

C. DESCRIPTIONS OF EDITING SOFTWARE

Three Generalized Editing Software Systems

Blaise

The Blaise system has been developed by the Netherlands Central Bureau of Statistics and is its standard data processing system. It is intended for use on microcomputer Local Area Networks (LANS) but can work on stand-alone machines as well. The required operating system for the microcomputers is MS DOS. The preferred LAN protocol is Novell, though Blaise will work with others as well. Turbo Pascal is required to compile applications programs; however, it is not needed by the end user. Development of applications in Blaise can be done in Dutch, English, Spanish, and French.

Blaise can handle categorical, continuous, and character data. It has been used for economic, agricultural, social, and demographic surveys. It handles edits of all types. In Blaise, the human editor is not replaced as the primary reviewer of data. Rather, the individual is given a more powerful, interactive tool with which to work.

Blaise is used to perform CATI, CAPI, and data entry as well as editing. Herein lies the strength of the system. Since it can perform these related functions, it can also integrate them. This

integration is done through the creation of a "Blaise Questionnaire". This questionnaire is not a survey instrument itself, rather it is a "specifications generator". In it, data are defined, routes are described. and edits are written. From this one specification, the Blaise system can generate two related modules. The first, for data collection, can be used for both CATI and CAPI applications. The second is used for data entry and data editing. Since Blaise integrates these related survey tasks, multiple specification of the data and edits is avoided.

Blaise does not perform data analysis because there are already many packages that can perform this Job. Blaise does generate dataset specifications for SPSS and Stata statistical packages and for the Paradox database system. Users can also specify their own specialized setups. Blaise can read in data from other sources as long as they are in (virtually any) ASCII format. A related tabulation module, which is part of the Blaise system, is called ABACUS. It can generate tables from Blaise data sets. These can be used for, among other things, some survey management functions. Weighted data can be tabulated in ABACUS.

Interactive editing in Blaise can be approached in several different ways. For example, data can be entered either by the analyst or by high speed data entry operators. In the first case. data are edited as they are entered. In the second case, the editor has several different ways of approaching the task. A batch edit can be performed. In the batch edit, records are marked as clean. suspicious. or dirty. The editor can retrieve the records based on their status. Also, the editor can access any record by its identification number or call up records based on certain criteria such as stratum designation, or the value or range of values of designated variables.

Generalized Edit and Imputation System (GEIS)

The GEIS system, developed by Statistics Canada, is based on the work of Fellegi and Holt. A predecessor, the Numerical Edit and Imputation System (based on the ideas of Sande), has been used as a prototype for GEIS. GEIS has been developed as part of the Business Survey Redesign

Project.

GEIS is intended to be applied to continuous data in economic surveys. Editing and imputation are considered to be part of the same process. In GEIS, data review and change are performed primarily in batch. GEIS performs edit analysis, error localization, and imputation. The system can be used on mainframes as well as on microcomputers. The database system ORACLE is required for all stages of processing, (GEIS is not part of the ORACLE system).

GEIS handles linear edits and variables that take positive values. Within these constraints, most situations can be handled. Non-linear edits can often be transformed to linear edits or can be restated keeping in mind the intent of the original edit.

It is intended that the system be used by a subject matter specialist working with a methodologist. Edits are specified interactively through specially designed screens. After specification, feedback is provided in the form of implied edits, extremal records, and edit analysis such as checks for consistency and redundancy. Data are

edited in batch. Fields are automatically selected for change under the principle that the smallest (weighted) set of fields is changed. Next, imputation is performed in a manner that the edits are satisfied. The primary method of imputation is hot-deck imputation where good records are used to donate values for incomplete records. Other model-based methods can also be specified.

Since GEIS is embedded in the ORACLE system, the edit and imputation process can be easily monitored. Many different kinds of reports can be generated. For example, the frequency of imputation by field, and the number of times each donor record has been used in imputation are two reports that can be generated. Through these reports, it is possible to measure the impact of the process on the estimates. Defensibility of the edit and imputation process is a priority in GEIS. This is done not only through the tracking of records as they proceed through the system, but also by providing the user with statistically defensible methods.

Data are held in an ORACLE database. Before they are edited in GEIS, they are treated in a preliminary edit. For example, all coding and

respondent follow-up would be done in this preliminary edit.

Unresolved records from the preliminary stage are sent to GEIS.

Structured Program for Economic Editing and Referrals (SPEER)

SPEER is intended primarily for continuous data under ratio edits for economic surveys conducted by the various divisions of the U. S. Bureau of the Census. SPEER applies the Fellegi and Holt methodology to ratio edits. Within that realm, SPEER performs edit generation and analysis, error localization. and imputation. Additivity edits can also be handled in SPEER. Other edits are handled either in satellite routines within SPEER or in a program outside of SPEER. Data are edited and imputed in batch mode first. On-line (interactive) review of referral records is an

essential part of SPEER. Records can be designated as referrals based

on criteria such as size of firm or on specific editing actions.

SPEER runs on mainframes as well as on microcomputer LANs. All of the SPEER modules are programmed in FORTRAN. A FORTRAN compiler is required to program new applications. The use of FORTRAN as the base language has the advantage of flexibility. The limits of SPEER regarding imputation routines, screen design, etc. are the same as those of FORTRAN (there are very few limits). When using the system, the services of a programmer are required to incorporate survey specific expert information.

In SPEER, both the machine and the human editor play major roles. Subject matter expertise is incorporated into SPEER through the programming of flexible modules. A hierarchy of imputation procedures for each variable is set; that is, imputation is on a field-by-field basis. The procedures are tried one at a time until a value within the feasible region is found. If desired, human editing actions can also be modeled in SPEER, through the use of IF-THEN statements. Since SPEER can handle most problems, the analyst is spared the task of reviewing minor problems and can concentrate on unusual or large

cases. When necessary, however, the analyst can review records interactively.

In the interactive review, the screen display includes reported data, corrected data, a status indicator. and the lower and upper limits of the feasible region for each variable. This allows the editor to see the effect of the editing actions vis-a-vis the SPEER limits. Also incorporated into SPEER is an audit trail, which keeps track of changes and reasons for them. The analyst requests a specific record and reviews the processing done by the automated system. The human expert can override the decision rules residing in the automated system and replace them based on alternative information about the case under review. The analyst typically has access to one or more of the following: the original response form, auxiliary information about the establishment under review, or the respondent by a telephone call. Based on this additional information and personal experience. an analyst may alter the decision rules built into the automated system. If there is reason to believe that the most appropriate imputation value lies outside the acceptable region, the analyst can select an imputed value outside the range. This system has also been

used as a data entry vehicle for late arrival forms. The late form are entered into the data file by subject matter specificants using SPEER and they are edited as they are being entered.

Brief Description of Other Systems or Programs

An Example of an Expert System Application.

An expert system application has been developed by the Quality Assurance Division of the Energy Information Administration. The program has been written for the Monthly Power Plant Survey (EIA-759). It was written to assist in the process of disposing of items that fail computer edit routines and to compensate for insufficient expertise and training of editors manually performing the process of disposing of edit failures. It was thought that the expert system could guide and assist the data editors through the more difficult dispositions of items that have failed edits thus

allowing the data to be edited according to the standard required. Though the system is ready for its first use, it has yet to be implemented operationally.

PEDRO, a System for the On-Site Entering and Editing of Data.

The Petroleum Electronic Data Reporting Option (PEDRO), developed at the Energy Information Administration (EIA), is an on-line system for data entry in which the respondents are involved in the data editing (Swann 1988). The respondents can use a personal computer for data entry or import data from the mainframe or another microcomputer system using a predefined format. The PEDRO system software then provides them with an image of a printed form which they proceed to "fill-out". The PEDRO data entry programs include a wide variety of edit checks to detect data errors at the time of entry. Users can enter and exit the PEDRO data entry function as often as they want while working to resolve any errors in the data. After data are entered, checked by PEDRO, and reviewed by the respondent, the data are transmitted to EIA.

Examples of the edits include a check to determine whether a total equals the sum of its parts and whether current month beginning stocks are equal to previous month ending stocks. Range edits that use historical data are included among the other system edits. Sometimes error messages will be generated for values that are actually correct. In that situation, the respondent is asked to provide an explanation for the anomalous value in the comments screen. This information is also transmitted to EIA making it unnecessary for an analyst to contact the respondent to explain the anomaly. Currently PEDRO is used by approximately 61 respondents to the "Monthly Refinery Report" and 10 respondents to the "Monthly Imports Report." Other offices in EIA are currently in the testing phase of using PEDRO for their surveys.

DIA, a System for the Automatic Editing of Qualitative Data

DIA is the name of a system developed by the National Statistical Institute of Spain (Garcia-Rubio and Villan, 1990). It applies the

Fellegi and Holt methodology to qualitative data. Only the minimum number of fields necessary are changed in order to satisfy the edits. The only specification necessary for imputation is that of the conflict (edit) rules. Each record is edited once and distributions are maintained. Random errors are distinguished from systematic errors, however, a rules analyzer ensures that both types of errors are treated consistently. Detailed information is provided by DIA on the whole editing and imputation process.

Micro-Macro Statistical Analysis System

The Micro-Macro Statistical Analysis System system is a graphics-on-screen, interactive, macro-editing system developed by the Bureau of Labor Statistics for use on the Current Employment Survey (CES). It is meant to replace the current batch system that generates thousands of computer printout pages. First, a table of industry identification codes for industries with suspicious estimates is presented. The analyst chooses one industry to work with. At this point. the analyst will try to find suspicious sample data which might have caused the problem. This can be done in either of two modes of operation: query

or graphics. In the query mode, tables of estimates for specific cells are displayed. The analyst can ask logical questions about a set of sample members

in order to select suspicious members. For a particular record, the analyst can reduce its weight so that it represents. orgy itself, can reject it, or can change entries. The effect of these micro changes can be seen at the cell (macro) level. In the graphics mode, current versus previous data points are displayed in a scatter plot for each variable. Outliers are easily seen and can be marked for further inspection in the query facility. Records that are changed in the query mode are marked when displayed in the graphics mode. A full audit trail is generated as changes are made in order to facilitate supervisory oversight of the process.

Other Software

Paul Cotton (1988), reviewed four systems in a paper entitled "A Comparison of Software for Editing Survey and Census Data". The paper is in two parts. A set of criteria for evaluating Editing Software is discussed followed by a review of the -four systems. In addition to the GEIS system, the paper describes three systems used primarily in the Third World. They are the Integrated System for Survey Analysis (ISSA), PCEDIT, and the Integrated Microcomputer Processing System (IMPS). ISSA was developed by the Institute for Resource Development Westinghouse, to process demographic and health surveys in Third World countries on IBM personal computers. It can perform data entry, data editing, and tabulation. ISSA is described in Cushing (1988). PCEDIT is available from the Department of Technical Co-operation for Development of the United Nations. It is meant to be used to process population (demographic) data. IMPS, developed by the International Statistical Programs Center of the U. S. Bureau of the Census, consists of six major components, one each for data definition, data entry, editing, tabulation, census management, and census planning. The name of the editing package is CONCOR. IMPS was developed to process census data in developing countries. The U.S.

Census Bureau is using CONCOR to edit and impute data for the 1990 Decennial Census for the U. S. Pacific Islands, (Guam, American Samoa, Northern Marianas, and Paluau). CONCOR is also being used to test edit specifications for population and housing characteristics for the basic 1990 United States Census long-form questionnaire. IMPS is described in Toro and Chamberlain (1988).

CHAPTER V

RESEARCH ON EDITING

A. INTRODUCTION

All survey or census data must go through some level of editing. In the absence of correction activities, errors could introduce serious distortions into the data and derived statistics. Surveys, survey staff, and processing capabilities all change over time, and procedures for editing change as well. Redesign or improvement for

edit systems can be minor to correct for slight problems, or there can be large research efforts to introduce major changes in methodology. These investigations can be carried out by specialists for a specific survey, programmers focusing on computer enhancements, or methodologists working on edit research.

Three related goals of the Research Subgroup of the Subcommittee have been to identify areas in which improvements to edit systems will prove most useful, describe recent and current research activities to enhance edit capabilities, and make recommendations for future research. The Edit System Questionnaire discussed in preceding chapters included questions about edit improvements. One question asked was "For future applications, what would you like your edit system to do that it doesn't do now?" Another source of information was discussions with those responsible for edit tasks within a number of Federal agencies. Two areas emerged as priorities: (1) on-line, human interaction with a computer edit system and (2) better ways to detect potentially erroneous survey responses.

Section B of this chapter provides examples of research in the two

areas mentioned above. Section C briefly describes editing research in other countries. Section D presents case studies of editing research in United States Federal Statistical Agencies. A summary is provided in Section E. In Appendix D an annotated bibliography describes research efforts over the past years and we discuss this bibliography in section F. The bibliography is particularly important because it is difficult to locate and identify research on edit development. Sometimes the research is part of a quality assurance project. Often, research findings are not written up as such, but they are implemented and evolve into practical and useful software. The chapter is limited to research on editing as opposed to imputation.

B. AREAS OF EDIT RESEARCH

One area of current research interest is that of on-line edit capabilities in which survey takers interact with editing software to edit responses at the time of data collection. This occurs in a CATI (Computer Assisted Telephone Interviewing) or CAPI (Computer Assisted

Personal Interviewing) setting. The BLAISE system discussed earlier is an example of edit software used in support of a CAPI and CATI program. Computer Assisted Self Interviews (CASI), is an innovative extension of these ideas which is to provide respondents with software to allow them to edit their own responses before transmission to the collecting agency. One software system and supporting

hardware for this purpose in use in Federal agencies is the PEDRO system which is described in Chapter IV. The topic of computer assisted data collection activities has been investigated in detail by the Computer Assisted Survey Information Collection (CASIC) Working Group.

Another use of on-line, interactive edit programs is in the review of edit referral documents. Most survey editing, especially of economic data, is a combination of automated batch computer runs and a follow-up review of selected cases by subject matter staff. The reason for targeting a record may be changes to a large case, large or unusual changes, or the need for an analyst to supply an imputation. An on-

line referral system should allow an analyst to make changes in a record, enter the change to the data file, and have the edit system validate the change or indicate that further. adjustments may be necessary. After an analyst completes the review of a record using an on-line system, the record should require no further action. This is in contrast to procedures currently in place in which an analyst will make "paper and pencil" changes to a referral document, changes will then be entered through some data entry process, the revised record will be run through an automated batch system, and the record may be targeted for further review. With an on-line, interactive referral system for analyst/clerical review of individual cases, the review process should be more efficient, less error prone, and less tenuous. Research into this area has a major system design orientation with the primary focus on software development rather than on new editing methodologies.

Several of the systems described in Chapter IV, EDITING SOFTWARE, incorporate interactive review. Blaise is a system in which interactive review is the primary method of data editing and which integrates editing with computer assisted data collection. SPEER is

a system where interactive review is tied in with Fellegi and Holt editing principles. In PEDRO, the respondent fills an electronic form that is edited at the same time. The Micro-Macro Statistical Analysis System incorporates interactive tabular and graphical review in order to perform macro-editing. The systems ISSA, PC EDIT, and CONCOR also have interactive capability.

A second area of active research is in the detection of potentially erroneous responses. The method for error detection most commonly used in Federal agencies is to employ explicit edit rules. For example, edit rules may require that:

(a) the ratio of two fields lie between prescribed bounds,

(b) the current response be within some range of a predicted value based on a time series or other models, or

(c) various linear inequalities and/or equalities hold.

Edit rules and parameters are highly survey-specific. A related editing research area is the design of edit rules and the development of methods for obtaining sensitive parameters. For some automated

edit systems the primary activity is to screen records which fail some combination of edit rules, after which data correction or verification is completed by subject specialists. This is especially true for questionnaires having, in part, a regulatory purpose or having only a small number of cases. For such edit systems, research will focus on selecting the appropriate edit rules. deriving sensitive bounds, and setting up flagging procedures. A related area of interest focuses on optimal methods to target cases for review as one does not want to burden the review process with an excessive number of referral cases nor does one wish to let many errors escape detection.

Several research studies are described in Section D in which the editing objective is to detect potentially erroneous responses. The first case study on methods to develop edit rules and tolerances was conducted at the Federal Reserve Board to derive set rules and

parameters for editing bank deposit data. One objective of this study was to determine procedures to group reporting units into clusters and form edit parameters by cluster. A related study at the Federal Reserve Board (FRB) to investigate the use of more model based range limits is described as well. Three case studies follow on the use of time series data on a firm's performance to predict current reporting and then edit actual reported values against those predicted. The first two studies describe research at Energy Information Administration (EIA) and are followed by a description of work at the National Agricultural Statistics Service (NASS). These studies illustrate the type of research being conducted at various Federal agencies and should prove useful as a source of ideas, directions, and considerations in edit system design.

In contrast to the rule-driven method for the detection of potentially erroneous response combinations within a record, one alternative procedure is to analyze the distribution of questionnaire responses. Records which do not conform to the observed distribution are then targeted as outliers and are selected for review and examined further for potential errors (Little and Smith, 1984 for example). Although there has been research interest in this topic, no application of

these multivariate methods was found. In addition, an investigation of the joint use of outlier detection procedures and rule-driven edits to detect potentially erroneous responses may prove valuable.

C. EDITING RESEARCH IN OTHER COUNTRIES

Much editing research has been conducted in national statistical offices around the world. It is these organizations, which conduct huge and complicated surveys, that have the most to be gained from developing new systems and techniques. They also have the resources upon which to draw for this development. The following are citations of people and organizations about which the members of this Subcommittee have knowledge.

Leopold Granquist of Statistics Sweden has presented papers on both the purposes of editing (Granquist 1984), and on macro-editing (Granquist, 1987). Granquist has also developed a typology of survey errors with which to judge the effectiveness of editing systems.

Members of the Australian Bureau of Statistics have given editing

papers at two recent Annual Research Conferences of the U. S. Bureau of the Census. The first by Linacre and Trewin (1989) addresses the optimal allocation of resources to various survey functions (including editing) in order to reduce non-sampling errors. The second by Hughes, McDermid, and Linacre (1990) concerns the use of graphical techniques to find outliers at both the micro and macro level.

The National Statistical Institute of Spain has developed a Fellegi and Holt system for edit and imputation of categorical data. In a recent paper, Garcia-Rubio and Villan (1990) discuss the applicability of the Fellegi and Holt methodology to randomly generated and systematic errors. They have made modifications in the methodology in order to better handle errors of the latter type. The Netherlands Central Bureau of Statistics (CBS) is the world leader in the use of microcomputer

Local Area Networks for the processing of survey data. Keller and

Bethlehem (1990) describe the systems, organizational issues and their resolution related to this new technology. Currently, the CBS has 2,000 microcomputers installed in 60 LANS. All the day-to-day processing of survey data is now carried out on these LANs using standardized software tools. The CBS has also carried out a "Data Editing Research Project" to determine the need for an interactive computer assisted procedure (Bethlehem, 1987). In Statistics Canada, Hidiroglou and Berthelot (1986) have developed a method of statistical editing for periodic business surveys.

An international group called the Data Editing Joint Group, has been meeting for a few years under the auspices of the Statistical Computing Project Phase 2 of the United Nations Development Program's Economic Commission for Europe. Countries represented include Sweden, Netherlands, Soviet Union, Yugoslavia, Hungary, Spain, France, Canada, and the United States. (The National Agricultural Statistics Service is the U. S. representative.) This group concentrates more on the systems aspects of editing and will be making recommendations about systems specifications both for their own use and for systems development in the third world. Phase 2 will be finished in the au

of 1990. The group intends to continue its work under the auspices of the European Statisticians Association with a focus on cooperating for their own benefit.

D. CASE STUDIES

Respondents to the questionnaire on editing practices expressed interest in deriving sensitive tolerance edits and using more sophisticated and extensive validation editing. They also mentioned that they would like to employ historic data to test the current data. An important aspect in the development of edits is determination of bounds or tolerance limits to use in identifying potentially erroneous data. Several recent research studies have focused on various ways of setting the bounds and on the limitations of the approaches.

Determining Optimal Clusters for Model Specification

If a large number of separate clusters or groupings are used to determine tolerances for edit rules, the procedure for providing ranges can become unwieldy. On the other hand, if too few groups are

used. erroneous items may not be flagged as ranges may become insensitive. Research to reduce the number of cells used to set tolerance limits has been carried out at the Federal Reserve Board (Pierce and Bauer, 1989).

To edit data that banks and other depository financial institutions submit to the Federal Reserve System, tolerance bands are constructed for groupings of institutions felt to be homogeneous by size. location, and type of institution. However, an objective measure of this homogeneity was not available. Since the edits were designed to flag or identify observations falling into the tails of the distribution of week-to-week changes. the measure proposed to assess the degree of homogeneity of different institutions was the variances of the changes for those institutions. As a result of performing an analysis of variance, with these variances as the cell observations, it was determined that an unnecessarily fine subdivision of groups was being used since the sample variances were not significantly different between many of the groups. Based on the results of multiple comparisons, new ways of combining the groups were suggested.

The Federal Reserve Board is also exploring an approach to editing referred to as "model-based edit design". The basic ideas behind this are that information in addition to the previous value of the item being edited is relevant in tolerance-band construction for edit checks (for example, last month's or year's values of the item, values of related variables, calendar or tax-date information), and such information is best incorporated into the editing procedure through a model which can then be used in determined the edit tolerances and executing the edits. Moreover, tolerance widths can be determined from the model's standard error estimate and given a probabilistic interpretation. The Federal Reserve Board will experiment with such models and model determined tolerances for pilot items and for selected sources of additional information and then move toward a more systematic development of this modeling approach. One topic to be investigated concerns the prospect for having common models for classes of similar financial institutions. This would avoid the

necessity for building an unwieldy number of models, while still having each model provide a sufficiently accurate description of the relevant behavior of the variable being edited for each institution. Intermediate possibilities include having a common model specification within a class but with different parameter values permitted for individual banks, or fixing those values but allowing different standard deviations (tolerance widths).

Use of Time Series Methods to Set Bounds

Another approach when historical data are available is to use past data values for a particular respondent to predict the current value and to then use the predicted value to construct tolerance limits for the new data. The Energy Information Administration (EIA) uses this approach for its weekly surveys on petroleum supply and its monthly surveys on petroleum marketing. Exponential smoothing is the particular technique used to obtain the predicted value (Burns, 1983). This technique has worked well during periods when the data are relatively stable; for example, on the weekly series on petroleum supply. However, it has not performed well when the data are erratic

such as when there are sharp price changes or seasonality.

To address the problem of this price change, the Petroleum Marketing Division of EIA has looked into the possibility of introducing a market shift into the edits that would account for real time market changes. The market shift is calculated from partial data from the current period. Before this was actually fully implemented, it was employed using an externally calculated market shift based on 'industry information. Later, these ideas were implemented by calculating the shift using as much current month data as available at the time of editing. This allowed not only full automation, but also targeted market shifts for varying populations and products as the data are received on a daily basis. To address the problem of seasonality in the monthly series on petroleum supply, the Petroleum Supply Division implemented tests on month-to-month differences rather than using exponential smoothing.

Research has also been conducted on the Kalman filter implementation of exponential smoothing (Kirkendall, 1988). The EIA used this procedure to obtain preliminary estimates of crude oil purchases and production. The procedure provided a method to both estimate and edit

state data. In some states the difference between the data on purchase volumes and production has remained relatively constant since 1983. In other states abrupt changes in the relationship or the presence of outliers were observed. Actually both transfer function models and ARIMA models were tried.

However, these procedures were not satisfactory in states in which large outliers or abrupt level shifts appeared.

Use of Robust Estimators to Set Bounds

The National Agricultural Statistics Service (NASS) of the Department of Agriculture has performed research on using Tukey's biweight to develop bounds for their statistical edit of data on livestock slaughter at the plant level (Mazur, 1990).

In searching for a statistical estimator to determine edit boundaries,

two desired properties are that tolerances quickly stabilize to new levels if true changes occur and that they return to old levels in the presence of outliers. Therefore. robust methods were considered because they are more resistant to outliers than the standard statistical methods and work well on many distributions as compared with the standard methods which work best when the distribution is normal. Initially, four measures of central tendency were considered: the mean, the median, the an (sum of the upper and lower quarides plus twice the median, entire quantity divided by four), and the 20 percent ed mean (the lowest n*0.20 values where n is the sample size and the highest n*0.20 values are dropped and the mean of the remaining values is computed). Four measures of spread were also considered the standard deviation, the inter-quartile range, median absolute difference, and the 20 percent trimmed standard deviation.

The mean and standard deviation were greatly affected by outliers. The other estimators seemed inadequate because they excluded good values. There was also the concern that they may underestimate the measure of spread. Because of these limitations, further research was conducted using the biweight. The biweight differs from other

estimators in that the weights are dependent on the data set. Therefore. it tends to include good values and excludes unreasonable ones. If the data are normal, the biweight is like the mean, but if the data are not normal, it is more like the median. The edit limits will be calculated for each plant, using the plant's 13 previous week's data. Research was also conducted on identifying inliers, that is, values that do not change much over time and are suspicious for that reason.

A key feature of the process is the use of stratification to provide edit limits for slaughter plants with insufficient data to edit small plants, or to impute for missing data. Also, a journal provides an audit trail. The analyst resolves error signals interactively on a microcomputer. Future research is being considered to extend the biweight approach to other data series which collect data from the same reporting units over time and to develop a system to plot historic data. Other possibilities include research to determine whether seasonality could be incorporated into the biweight (mainly for large plants) and whether a capability to identify plant holidays could be added.

E. SUMMARY

The survey on editing practices indicated that there was little analysis of the effect of editing on the estimates that were produced. Considering that the cost of editing is significant for most surveys, this is clearly an area in which more work is required. A related issue is to attempt to determine when to edit and not to edit. Clearly, all the errors are not going to be found and we should not

attempt to find them all at the risk of over-editing. An interesting task is in designing guidelines for determining what is an acceptable level of editing.

Another neglected research area in this country concerns the editing of data at the time they are keyed from maid responses. Data entry systems typically have some keying error detection capabilities of a univariate nature, typically range checks and checks to detect when an unacceptable character has been keyed. The primary focus of checks

at this stage is to detect data entry errors. This area is usually discussed in the setting of quality control; however, it is an area that can benefit from further research from the perspective of data editing. A number of surveys have reduced this sort of error through the use of double keying. In the Netherlands Central Bureau of Statistics, subject-matter specialists enter data and edit them interactively as they are entered.

The advancement of computer and peripheral technology is playing a dual role in affecting survey editing. On the one hand, some developments have helped to eliminate or reduce the need for some edits. Computer Assisted Data Collection systems (e.g., CATI, CAPI) not only reduce data entry errors but reduce other errors as well. The use of machine-readable forms and bar codes will eliminate keying errors. On the other hand. the increased speed, memory, and storage of computers, and networking have allowed statisticians to consider computationally-intensive techniques for editing that previously would have been possible, particularly considering survey deadlines, and to utilize other databases.

The questionnaire respondents expressed interest in the use of expert systems to improve survey specific sensitivity in editing. The term "expert system" is not really well defined and different analysts attach different meanings to it. With respect to data editing, it refers to the treatment of survey specific information in a structured way. In that regard, the computer is simulating, to some degree, the role of the subject matter specialist. Two systems already in use that have expert system components are SPEER and PEDRO. (These systems are described in Chapter IV.) In these systems, decisions that may have been made by subject matter specialists are now made by using rules that have been programmed 'into an automated system.

In this chapter, examples of research on methods to detect erroneous values were discussed. With improved technology, the techniques have become more sophisticated and undoubtedly will continue to become more so. The question then becomes how effective are the techniques in actually detecting errors. Two related areas for further research are monitoring the effectiveness of the edits and determining guidelines for when to use each technique. To address these issues, it is necessary to track the proportion of flagged items that are actually

errors (often referred to as the "hit" rate). This, of course, only gives one side of the picture; it does not address the issue of errors that are not detected by a specific procedure. Despite this limitation, tracking the "hit" rate is useful and ways of automatically alerting the analyst that it has gone out of control would be helpful.

As more techniques become computationally feasible, the analyst is confronted with more choices in designing an edit system. It would be useful to know when the techniques work well. For example, research has already indicated that exponential smoothing does not work well when the data are erratic. If findings could be made available about other techniques, time could be saved in developing new edits.

In conclusion, there are several recommendations for research in data editing that are contained in the preceding paragraphs. However, the

most important recommendation we can make is that agencies recognize the value of editing research and place high priority on devoting resources to their own research, to monitoring developments in data editing at other agencies and elsewhere, and to implementing improvements.

F. BIBLIOGRAPHY

It is quite difficult to provide a complete assessment of current research activities in the area of editing because so much of the research, progress, and innovations are described only in survey-specific documentation. The difficulty is even more fundamental. Innovations in editing methods made by survey staff are often viewed as enhancements to processing for that particular survey, and little thought is given to the broader applicability of methods developed. Accordingly, survey staff do not typically prepare a discussion of new methods for publication or for other forms of wide dissemination. A description of editing methods and system design might be found in survey processing specifications. instructions to programming staff, or in survey processing code. Innovations that are computer intensive

often are regarded not as method changes, but rather as computer enhancements. In other cases, edit activities may be included in the general area of "quality assurance" with little thought of the subject of editing per se.

For these reasons, any bibliography on editing will undoubtedly miss important areas of research and innovations. Fortunately, a number of researchers did see editing as distinct from other processing tasks and have taken the time to describe their experiences. Some of the papers in the bibliography can be viewed as case studies for a particular editing strategy employed on a particular survey. To some extent, authors of such papers wanted to record their activities, subject them to public scrutiny, and offer up their techniques to others who may be working under similar conditions and who may find their suggestions useful. It is often in such articles that methods which may be applicable to more than one survey are first introduced and described.

There are features of the editing process that cut across surveys, and this realization has encouraged the development of general

methodologies and multiuser systems. Much recent research in the area of editing has focused on the development of multipurpose edit systems, and a number of papers in this bibliography discuss multipurpose edit systems. Some of these systems have imputation components while others do not. nm preceding chapter on Editing Software described three multipurpose software packages: GEIS, BLAISE, and SPEER. In the respective specialized bibliographies, ([A], [B], and [C)), we include papers which describe underlying methods, the software, proposed uses, and possible advantages of the respective systems. The bibliographic citations provide the theoretical and research background for these systems and constitute a link between the software chapter and this research chapter.

APPENDIX A

RESULTS OF EDITING PRACTICES PROFILE FROM

QUESTIONNAIRE RESPONSES

	Frequency	Percent
1. What type of survey are you engaged in?		
a. Sample	90	77%
b. Census	27	23%
2. What is the purpose of the survey?		
a. Statistical	98	84%
b. Regulatory	0	0%
c. Both	19	16%
3. How would you classify your survey?		
a. Single-time survey	6	5%
b. Repeated survey (cross-sectional)	50	43%
c. Panel Survey (longitudinal)	39	34%

d. Rotating panel survey	11	10%
e. Split panel survey	6	5%
f. Other	3	3%
Not answered	2	

4. What is the frequency of your survey?

a. Weekly	4	4%
b. Monthly	23	20%
c. Quarterly	12	10%
d. Annual	39	33%
e. Other	39	33%

5. What is your sampling unit?

a. Individual	21	18%
b. Household	11	9%
c. Farm	6	5%
d. Economic Establishment or firm	58	50%
c. Institution	8	7%
f. Other	13	11%

	Frequency	Percent
6. How many units are included in your survey?		
22 through 1,000,000		
7. Is response to your survey mandatory?		
a. Yes	43	37%
b. No	74	63%
8. Averaging across all items, what level of item nonresponse does your survey experience?		

a. None	18	16%
b. Less than 5%	43	38%
c. 5% or greater, but less than 10%	20	18%
d. 10% or greater, but less than 20%	19	17%
e. 20% or greater	12	11%
Not answered	5	

9. What is your primary data collection method?

a. Computer-assisted telephone interview (CATI)	4	3%
b. Computer-assisted personal interview (CAPI)	0	0%
c. Telephone interview	9	8%
d. Personal interview	25	21%
e. Mailed questionnaire	49	42%
f. Administrative records	18	16%
g. Other (please specify)	12	10%

10. What secondary data collection method(s) do you use?

(Circle all that apply)

a. Computer-assisted telephone interview (CATI)	19
b. Computer-assisted personal interview (CAPI)	2
c. Telephone interview	62
d. Personal interview	24
e. Mailed questionnaire	24
f. Administrative records	16
g. Other (please specify)	5

	Frequency	Percent

11. What type of computer do you use for data processing?

(Circle all that apply)

a. Mainframe	107	
b. Minicomputer	20	
c. Microcomputer	40	
d. None	0	

12. What is your data processing environment?

a. Batch mode	49	42%
b. On-line	9	8%
c. Both	58	50%
Not answered	1	

13. If your survey is computerized. what sort of file structure do you use? (Circle all that apply)

a. Sequential	71
b. Database using ORACLE software	7
c. Database using ADABAS software	5
d. Database using DBASE software	10
e. Database using other software (please specify)	35

14. Are you limited in your ability to disseminate data by confidentiality (privacy) restrictions?

a. Yes	104	89%
b. No	13	11%

15. Do you release microdata (respondent-level data)?

a. Yes, and imputed data items are identified (flagged)	36	31%
b. Yes, and imputed data items are not identified	19	16%
c. No	62	53%

16. When you release aggregated data, do you provide information as to the percentage of a particular data item which has been imputed?

a. Yes	24	21%

b. No	89	79%
Not answered	4	

Frequency Percent

17. Are there minimum standards for reliability for the data you disseminate; e.g., do you require that an estimate have less than an established maximum variance or be based on more than an established number of observations before the estimate can be released?

a. Yes	79	71%
b. No	33	29%
Not answered	4	

18. What documentation exists for your survey?

a. All aspects of the survey are well documented	88	76%
b. The data editing system is well documented, but some of other aspects are not	10	9%
c. Some aspects are documented, but not the data editing system	1%	
d. Some documentation exists, but it is neither complete nor current throughout the system	16	14%
e. No documentation exists for this survey	0	0%
Not answered	2	

19. Is information available on the . cost of data editing in your survey?

a. Yes	44	38%
b. No	73	62%

20. Please estimate the percentage of the total survey cost spent on data editing.

(Please include all of the aspects of editing, such as any edits made at the time of data entry, clerical work, computer time, design, testing, monitoring, analyst review, call-backs, and review of summary tables.)

Range	5% through 90%
Mean	41.4%
Median	35%
Mode	10%
Standard Deviation	27.2%

21. Who designed your data editing system? (Circle all that apply)

a. Subject matter analysts	108
b. Methodologists or mathematical statisticians	70
c. Computer systems analysts	88

	Frequency	Percent
26. Are there any procedures in place to monitor actions of editors (clerks or analysts) to detect a pattern of undesired actions in order to remedy the cause(s)?		
a. Yes	76	66%
b. No	39	34%
Not answered	2	
27. Are there any procedures in place to monitor actions of the automated part of your editing procedure to detect a pattern of undesired actions in order to remedy the cause(s)?		

a. Yes	84	74%
b. No	29	26%
Not answered	4	

28. Is there an audit trail in your data editing system, i.e., is a record kept for all data editing transactions?

a. Yes	70	61%
b. No	45	39%
Not answered	2	

29. Do you regularly collect performance statistics in order to evaluate your data editing system?

a. Yes	69	61%
b. No	45	39%
Not answered	3	

30. Has any analysis been done on the effect of data editing practices on the estimates produced?

a. Yes	42	38%
b. No	69	62%
Not answered	6	

31. Do you release data editing information on your survey?

a. Yes, with the data	20	18%
b. Yes, in a different publication	12	10%
c. No	83	72%
Not answered	2	

Frequency Percent

32. Validation editing is defined as a procedure which detects inconsistent data within a record. One example might be a test which verifies that the sum of detail items equals a total field. Another example might be a test to detect an individual who reports that he is single, but also reports the name of a current spouse. Do you do validation editing?

a. Yes	115	98%
b. No	2	2%

33. Macro-editing is defined as a procedure which tests aggregated data to detect inconsistencies. An example might be the comparison of a data table from the current period's survey to one from a previous period to look for gross changes. Do you use macro-editing?

a. Yes	97	83%
b. No	20	17%

34. Do you do any data editing techniques other then validation or macro editing?

a. Yes	66	57%
b. No	50	43%
Not answered	1	

35. If you use other types of edits (for example, edits based on the statistical distribution of the data, using time-series techniques, or employing ratios between data items), briefly describe them below.

36. May a record which fails some edits be accepted as a part of the final file (to be used for tabulation and/or dissemination)?

a. Yes	102	87%
b. No	15	13%

	Frequency	Percent

37. What method do you use of impute for data edit failures?

(Circle all that apply)

a. Imputing zeroes	11	
b. Adjusting sample weights	24	
c. Cold-deck imputation procedure	17	
d. Hot-deck imputation procedure	29	
e. Model-based imputation (regression, etc.)	19	
f. Multiple imputation procedure	8	
g. Imputing last period's value	31	
h. Other (Please specify)	21	
i. None	37	

38. Does your data editing system perform those tasks which you intend it to?

a. Yes	112	96%
b. No	5	4%

39. If you are not completely satisfied with your data editing system, do you feel the changes that need to be made are considered minor, or would a complete overhaul of the system be needed?

a. Satisfied with our system at this point	53	47%
b. Complete overhaul of the system needed	6	5%
c. Minor changes needed	28	25%
d. Depending on system, some changes are minor, some major	23	20%
e. Not possible to determine at this point	4	3%
Not answered	3	

40. For future applications, what would you like your data editing system to do that it doesn't do now?

41. Please describe any research efforts which you are currently engaged in on the topic of data editing?

APPENDIX B

CASE STUDIES

I. ABSTRACTS

Bureau of Labor Statistics (BLS):

CPI: Commodities and Services

This monthly survey is an important statistical series which is a longitudinal sample survey with about 90,000 quotations per month. Data are continuous and are collected by BLS professional data collectors in 88 primary sampling units. Response rates are very

high. Historical data and trends are used. CPI uses mainframe and a micro-based PC network. Micro/PC is used for data capture and microdata editing processes. The mainframe is the official database for all historical data.

Data collection forms are pre-edited manually for completeness and then a machine edit is done. Some records require review by subject matter specialists who review about 30% of the price quotations for exception edits. Extensive use is made of these specialists. About 10% of the data are changed by specialists. Data are used to compute indexes, and are again reviewed in this form. CPI approach to data editing is "bottom up" in that subject matter specialists concentrate first on individual price quotations referred to them for review. A new approach to editing is being studied using computer assisted technologies, in which manual pre-editing is eliminated. Changes in the current process are recommended, so that the first review step is summary level data, along with detailed analyses of outliers.

Internal Revenue Service (IRS):

US Corporation Income Tax Returns

This survey collects information from a split-panel sample of approximately 85,OW corporation income tax returns to produce population estimates from financial aggregates and to provide a microdata file for revenue estimation. Returns representing a given fiscal year are selected over a two calendar-year period. Both continuous and categorical data are collected. Data are processed in a in environment in batch mode.

Skeleton records are drawn from IRS's revenue processing system. which processes some, but not all of the information on each return. This information has been tested in revenue processing. Additional information is manually abstracted. Some information for small firms is imputed rather than abstracted as a cost-saving measure. All data are subjected to an iterative process of machine testing and manual correction. After all microdata edits have been passed, returns are restratified based on edited information (about 3,750 returns were reclassified in TY 1986) and population totals are adjusted. Data are then tabulated and the tabulated data are examined for suspicious or inconsistent values by industry experts. New methods are being

studied to incorporate on-line data entry and editing, ratio edits and time-series edits.

National Center for Education Statistics (NCES):

National Education Longitudinal Study of 1988 (NELS:88)

This survey is the third in a series of longitudinal sample studies which take place about every 10 years. NELS:88 began in 1988 with 25,000 eighth grade students, who will be re-surveyed every two years. Both continuous and categorical data are collected in large group administration sessions by survey personnel in schools. Some data from parents and teachers are collected by mail. Microdata is released to the public. Item response rates are generally very high (over 95 percent in most cases). Estimated cost of editing is about

10-12% of entire cost of survey.

Some on-site edits are done for critical items. Computers are used to perform critical item edits, and missing data are retrieved by phone. Interview validation is carried out on 10% of instruments to detect discrepancies. The mainframe is used to put data through "forced cleaning" and item specific issues. Very little imputation is done for missing values. Range checks, inter-item consistency check and some statistical editing are carried out. In addition, subject matter specialists perform various types of analyses.

Federal Reserve Board (FRB):

Edited Deposits Data System (EDDS)

The EDDS data are the primary source of information for money supply estimates, and are collected daily and weekly by mail from 9900 depository financial institutions (DFIS) which meet minimum size criteria. DFIs complete a, standard form and data are continuous. There is no imputation done and no problem with non-response. Historical data are used to impute if a DFI is late in reporting.

Microdata are confidential and are not released.

When data are received at the 12 Federal Reserve Banks (FRBS) from DFIS, they are transcribed into machine readable form. They are edited for both validity and quality (statistical editing) by the FRBS. If edit failures occur, FRB personnel investigate. Data are transmitted to the Board and are again edited for validity and quality. Editing is conducted on mainframe processing system. Validity edits (for macro and micro data) and quality edits (for macro and micro data) are carried out, including use of stochastic edit checks. Historical data, including changes in observed figures, are widely used. If quality edits are violated, FRBs call back to institutions to verify. Quality edits are prioritized, based on the size of the DFI, total deposits, and degree of fluctuation from past numbers seen. Research is now underway to improve editing procedures by modelling and by incorporating expert system technology into the editing process.

II. IN-DEPTH CASE STUDIES

National Center for Health Statistics (NCHS):

National Health Interview Survey (NFHS)

Environment In Which Survey Takes Place

The National Health Interview Survey (NHIS) is a household sample survey which has been operated by the National Center for Health Statistics (NCHS), Centers for Disease Control, since 1957. Its purpose is to produce a wide range of general purpose statistics on the health of the U.S. population. It consists)f a continuing standard health and demographic questionnaire, questionnaires on special health topics, and ad hoc longitudinal follow-up studies (not discussed further in this document). The survey is designed, processed, and analyzed by the Division of Health Interview Statistics and other NCHS staff, numbering about 60 FTES. Under an interagency agreement with NCHS, field work for the survey is done by the Field Division, Bureau of the Census.

To avoid seasonal bias, data are collected continuously throughout the

year, each week of data collection comprising a nationally representative sample. Although reports can be produced for individual weeks or quarters of data collection (and have been), most reports aggregate data for a full calendar year. A standardized annual summary report is published each October for data collected in the previous calendar year.

Each year, interviews are completed in 47,000 households, yielding information on 123,000 sample persons. The amount of information collected on each sample person varies, but averages several hundred items.

Before 1988, all data were collected in face-to-face interviews in sample households using paper and pencil (PAPI) (although interviewers were allowed to administer questionnaires over the telephone after repeated call backs failed to produce a face-to-face interview). Beginning in 1988, computer assisted personal interviewing is being phased in: one special health topic questionnaire (on AIDS knowledge) is entirely on CAPI. and by 1990 all special topic questionnaires will be on CAPI.

For the basic health and demographic questionnaire, interviews are completed in about 95% of eligible households; for the special topic questionnaires, which usually require self-response by one adult subsampled in the household. interviews are completed with about 85-90% of eligible respondents, yielding an overall response rate of 81-90%. Item response rates are in the range of 97-99% for nearly all items, although nonresponse on income questions is about 15%.

Editing Practices

Data Processing Environment and Dispersion of the Work

CAPI and PAPI are done in the field by about 100 interviewers working in about 200 first-stage sampling areas in all parts of the country. Preliminary hand edits are done by field supervisors in 10 regional offices of the Census Bureau.

Editing, coding, and data entry (with electronic range edits) are done at the NCHS processing facility in Research Triangle Park, North Carolina. Machine edit specifications and edit programs for consistency edits and code changes are written at NCHS headquarters in Hyattsville, Maryland. These edit programs are the basic components of a well-defined processing system for each special health topic questionnaire. Data processing is run on a mainframe computer at the North Carolina facility, and edit failures are adjudicated between the analysis and programming staffs at Hyattsville and the coding staff at North Carolina.

Editing of questionnaires on special health topics, which change at least annually, uses the same general approach as the editing of the basic questionnaire, but because subject matter and personnel vary, there are more inconsistencies in methods within and between these questionnaires. Editing procedures for CAPI are different than those for PAPI. Editing occurs in the field by the interviewer, at regional

offices by clerks, during coding and keying at the processing facility, during machine processing, and at the data analysis and reporting stage.

Audit Trail

A complete audit trail for machine edits of the basic health and demographic questionnaire is produced, printed, and made available to in-house users for each data year. These documents are not available to public users.

Micro-, Macro-, and Statistical Editing

Extensive micro-computing is done by machine. Some macro editing is done by hand examination of the results of machine edit reports, such as ex g counts of the number of times certain types of errors occurred and examining distributions for outliers. Standard sets of estimates are produced quarterly and annually and visually compared with the tables for earlier years to search for anomalies. For in-house users, a complete audit trail is made available as part of file

documentation; public use data tape documentation summarizes edit and imputation information.

Prioritizing of Edits

Priority is given to identifiers needed to link data files. and to questionnaire items used to weight the data to national estimates (age, race. sex). Otherwise. no formal system is used to give higher priority to some edits rather than others. Informally, through the allocation of staff skills, staff time and staff interest, priority may be given to some questionnaire topics or questions.

Imputation Procedures

Unit nonresponse (missing sample cases) is imputed by inflating the sample case weight by the reciprocal of the response rate at the final stage of sample selection, or by a poststratification adjustment based on independent estimates of the population size in 60 age-race-sex categories.

Item non-response (missing question answers) is imputed, where possible, by inferring a certain or probable value from existing information for the respondent. Most imputation is done by machine, although some otherwise unresolved edit failures may be imputed by hand. No mathematical, hot deck. or cold deck imputation is done.

Editing and imputation Standards.

Error tolerance standards are established for interviewer performance, and enforced by editing and telephone reinterviews conducted by regional supervisors. Error tolerance standards are established for keying of data, and are enforced by re-keying of a sample of questionnaires. No formal standards are established for editing at other stages of the survey process.

Costs of Editing

The costs of editing include approximately 20 FTE's per year and about $ 1,000,PW per year.

Role of Subject Matter Specialists

The primary role of subject matter specialists is to write edit specifications, from which edit programs are prepared, to review results of edit runs and adjudicate failures in collaboration with programmers. Their secondary role is to compare standard sets of estimates with historical series to identify anomalies. In addition, they also consult with survey design staff on field edits.

Measures of Variation

Estimates of sampling errors are produced and published for each component survey. No estimates of nonsampling error are produced.

Current and Future Research

The Division of Health Interview Statistics recently implemented a program of methodological research which will include measurement of data quality as a function. In a current project, respondent reports of health conditions are being compared with medical records of the respondents.

National Agricultural Statistics Service (NASS):

Agricultural Survey Program: Quarterly Agricultural Surveys

Environment In Which The Survey Takes Place

The Quarterly Agricultural Surveys are the main vehicles for collecting current agricultural production data. One of the essential aspects of data quality is its timeliness. In every quarter, data are collected starting on the first of the month. Data must be collected and editing must be finished by about the 15th. Results of the

surveys are released starting at the end of the month and continuing into the first few weeks of the next month. That is, all data analysis is carried out and all summaries are executed within a month of the end of data collection. Estimates are released according to a very rigid schedule that is published before each calendar year. The estimates of current agricultural supply, when combined with estimates of demand by economics, form the basis for price forecasts in agricultural commodity markets.

Nationally, data arc collected on about 500 agricultural items. Each state collects data on about 70 variables. In order to customize questionnaires to each state's culture, about 35 versions of the master questionnaire must be produced in each quarter.

Any farm, ranch, individual or partnership that has land on which any crops have been grown, grains or oilseeds have been stored. or any livestock has been raised in the calendar year, in any of the 48 contiguous states is 'in the survey universe.

The sample size varies between quarters, from about 75,000 to 86,000.

In 1988, unit nonresponse was 12%. The number of forms to be processed in a quarter ranges from about 65,000 to 80,000. NASS also requires that certain refusals be hand-estimated by statisticians in the state offices. This is done for very large operations from the list frame and for all livestock data in the area frame. This hand estimation is considered to be part of the editing process. A rotating panel design is used for an annual cycle. A new panel, however, is created every June quarter. Large operations are interviewed every quarter, and others are rotated in and out during the year to reduce burden.

The primary mode of collection is by phone onto paper questionnaires. There are 14 states in which data are collected by Computer Assisted Telephone Interviewing (CATI). Data are also collected by face-to-face interview for difficult respondents or for those who do not own a phone. Mail is also used heavily in a few states, but this mode makes up only a small percent of the national total. NASS is procuring a microcomputer based LAN for each of its field offices. When all of these LANs are installed, it is expected that CATI will be the major mode of collection for this survey.

Editing Practices

Data Processing Environment and Dispersion of the Work

Data are collected through 42 field offices. There is one field office for each state, with the exception of New England (one office for 6 states) and one office for Maryland and Delaware. Hawaii and Alaska are excluded from the survey. Each field office is responsible for the collection, capture, editing and initial analysis of its data. The Agricultural Statistics Board for

NASS releases and publishes the official statistics along with the state level data. The processing is coordinated in headquarters in

Washington DC. All programs are written at headquarters. All data processing is carried out on one leased mainframe located in Florida. The exception is CATI which requires the use of minicomputer servers in the field offices for collection. Data are sent between each field office and mainframe through dedicated telecommunication lines. The results of each stage of processing are sent back to the field offices and are printed out. Editing occurs at data collection, especially for CATI, and prior to data entry and during the analysis stage.

Audit Trail

There is no automated audit trail except with the CASES software used in CATI. In theory, it is possible to trace the editing actions for an operation by working through a paper trail. For example, if the editor changes data during the hand edit, the changes are to be marked on the questionnaire in red pencil. If the editor makes a change in data on a computer printout (the change is then keyed into the data file), the editor is supposed to mark the change on the questionnaire in green pencil. In current procedures, there is no way to know the effect of editing on the final estimates. It is not possible to

summarize the editing actions for more than a few questionnaires. It would be a very tedious, hand intensive job.

Micro-, Macro-, and Statistical Editing

In the 14 states that employ CATI, much of the micro-editing is done at the time of data collection. The CATI instrument avoids some errors such as routing errors because the program controls the flow of the interview. Some errors, such as consistency errors, are caught by the program immediately. For example, if a farmer has more harvested acres of a crop than planted acres, the program will require that one cell or the other is changed. Noncritical, or suspicious errors, are also caught in CATI. If a farmer has an unusually high or low yield, he can be asked to confirm that he understood the question correctly. In a 1983 cattle survey study, CATI reduced the number of signals generated by 77% and 3% for critical and noncritical errors respectively. While the noncritical error signals were not greatly reduced, more of these were allowed to pass to summary as they had been verified with the farmer.

After CATI collection, data are transferred to the normal batch editing system. This system is called the Survey Processing System (SPS), which is written in SAS. When errors are flagged in this system the editors do not have a form to refer to when evaluating the error signals. With a special effort, the data may be displayed on the screen of the CATI computer. However, the CATI software is not used to do any post-codection editing.

In non-CATI data collection, micro-computing can be considered to be a two stage process. First is a hand edit. This may be done by clerks or by statisticians or both. Second is a machine edit. This is performed in batch using the Survey Processing System (SPS). The SPS was developed by and is used exclusively in NASS.

In the hand edit, the editor inspects the form for notes, performs required calculations, checks for reasonableness, and will rewrite unclear numbers. Changes may be made to the data at this stage. If so changes are to be noted on the form with a red pencil. Also at this stage, essential coding will

be done. This coding is used to trigger certain edits and informs the imputation routine how to handle the questionnaires.

After the hand edit is performed, the data are entered by high-speed data entry operators - The data are transferred to the leased mainframe computer and are edited in batch. Error signals and messages along with selected data are printed out. The statistician reviews the printouts and marks on both the printout and the form (in blue pencil) the adjustment to be made. These changes in data are then re-keyed and re-edited in batch.

After the data are declared clean in the machine edit, and after the data are run through the imputation routine, an "analysis package" is run for each state. This analysis package contains elements of both macro- and statistical editing. Any errors detected at this stage can still be updated in the data file.

Macro-editing, defined as applying edits to aggregated data, is carried out at the stratum level. Expansions for selected items are generated by stratum. These expansions are compared to expansions from the previous period. This kind of inspection helps to trace suspicious data that have an impact at an aggregated level.

The statistical editing for this survey consists of reviewing the highest 20 and lowest 20 values of an item or ratios of items before and after expansion factors are applied. This may be done across all records or within certain strata.

Imputation Procedures

Imputation occurs at two stages in the processing. Imputation may occur during the hand edit or an automated imputation routine which is run after data are declared clean in the batch edit.

Hand imputation is used when the data editor feels that there is enough information on the form to enable an accurate determination for

cells that are not filled in. The most common example of this is when an enumerator writes notes that enable a quantity to be calculated. For example, NASS asks for hay stocks in tons of hay. Many farmers can give a number of bales and the average weight per bale, but not an overall figure. The tons are derived by hand and the figure is imputed into the cell. For production items (e. g., number of bushels of corn produced) or for stock items (e. g., number of bushels of corn stored), the statistician may write a "missing value" code in the cell. This figure will invoke the automated imputation routine for that cell.

For list and area frame questionnaires, when crops or stocks sections are totally blank, the statistician is to fill in a "completion code" during the hand edit. This code informs the imputation routine that the operation has the items of interest does not have the items of interest or it is unknown whether the operation has the items of interest. The level of the amputations are then based on the value in the completion code. For example, if the farmer is known to have the items, then the imputations will be higher than if the operation's status is unknown. If livestock sections are blank for list frame

questionnaires. and the questionnaire is otherwise usable, the statistician is required to impute the livestock data by hand using whatever historical data are

available. For area frame questionnaires that have not been filled in, the statistician is required to make the form usable through hand imputation. Any auxiliary data that are available are used.

The automated imputation procedure attempts to make use of administrative data or data from previous surveys. If these data are unavailable, then the routine imputes appropriate averages for the blank cells. For a particular operation, averages are calculated from as fine a sample as possible. For example, averages may be based on farms reporting an item within the operation's stratum and district in the state. If there are not enough reports, then a higher level of aggregation is used. This customizing of imputation has the effect of making the imputed values very sensitive to the coding that is

done. In other words, if the coding is done incorrectly, then the level of the estimates may be affected.

Editing, and Imputation Standards

Quite detailed guidelines are published in a Supervisory and Editing Manual. Pre-survey practice, and editing sessions at National survey training schools try to ensure that editing is done consistently across the country.

Costs of Editing

Complete survey process management data are not currently generated by NASS which would summarize the costs of editing in detail. However, it is estimated that about 15 percent of the total survey cost can be attributed to editing.

Role of Subject Matter Specialists

Subject matter specialists are called agricultural statisticians in

NASS and their expertise is considerable. Agricultural and mathematical statisticians are responsible for reconciling problems at the state level. Few actions are entrusted to a machine. The agricultural statistician is held to be fully competent in making decisions. The edit programmers in DC usually have had experience as agricultural statisticians in state offices.

Measures of Variation

Standard errors are calculated along with the estimates but are not published. The standard errors do not take into consideration the effects of editing.

Current and Future Research

An interactive editing research project has just been carried out in Ohio and Wisconsin for the 1989 December Agricultural Survey using the Blaise system. This project investigated moving away from the centralized batch environment to a microcomputer-based interactive environment. It also investigated the possibility of capturing and

editing "raw" data (data that has not been hand edited prior to data entry). The new processes took 50-80% of the time necessary for the conventional process.

The Interactive Editing Working Group concluded that interactive microcomputer-based editing should become the standard process for NASS. As NASS is currently procuring microcomputers corrected in LAN's for each office, there is an opportunity to collect this survey in CATI and then to edit it interactively. The wider implementation of CATI would reduce considerably the need for editing, and the editing that remains would be done very efficiently.

More sophisticated macro- and statistical edits, possibly on an interactive basis, are beginning to be researched. These include the

editing of data after it has been entered (i.e. no previous hand edit) and outlier checks based on expanded, censored or robust techniques, statistical graphics, and multivariate relationship data have all been identified as ways editing could be improved. Also, research on more automated imputation (to replace the judgmentally based imputation) for the large operations and livestock data could be done.

Bureau of the Census: The Enterprise Summary Report

Environment in Which The Census Takes Place

The Enterprise Summary Report (ES-9100) and the Auxiliary Establishment Report (ES-9200) are part of the Census Bureau's Economic Censuses. The economic censuses are conducted at five year intervals in the years ending with a 2 or a 7. The ES-9100 and ES-9200

censuses target the following industries: Mineral, Construction, Manufacturing, Wholesale, Retail, Selected Services and Selected Transportation.

Respondents to the ES-9100 are all companies in the target population with 500 or more employees, and the universe for the ES-9200 consists of all auxiliary establishments of those firms. An auxiliary establishment is an establishment, typically a non-revenue producing establishment, that provides support activities to other establishments of a company. Examples of auxiliaries are research and development centers, warehouses and administrative offices.There were 8.811 companies in the ES-9100 universe and 39,461 auxiliary establishments in the ES-9200 universe for the year 1987. The primary mode of data collection was through questionnaire mail-out with a follow-up for large operations that did not respond using telephone call backs. A unit response rate of approximately 85% was achieved.

Editing Practices

Data Processing Environment and Dispersion of the Work

The batch version of the complex edit for these censuses was implemented on the SPERRY mainframe. It was then adapted to micro-computers without any difficulty. The micro-computer version is an interactive, on-line edit used by analysts in their review of referral cases. The microcomputers are connected through a LAN sharing a single database access. The programs have also been adapted and tested for the VAX/VMS system.

Editing and quality control activities take place throughout the survey process. Processing begins at the Census Bureau's data center in Jeffersonville, Indiana. It is here that data entry, following up with correspondence to establishment, microfilming forms, and clarifying types of business are done. The editing done at Jeffersonville is basically manual and is a standardized procedure across the economic areas. The clerks follow specific procedures documented by subject matter specialists at Census headquarters.

Audit Trail

A complete audit trail is kept for all actions of the Structured Programs for Economic Editing and Referrals (SPEER) complex edit. During the complex edit, referral flags as well as informative flags are set. A referral flag is a flag that targets a record for analyst review. This flag also has priorities attached to it. For example, a flag indicating a large change to a basic item will cause a record to be referred even though this may be the only flag set during the entire survey process. An informative flag is set simply to convey information. For example, each imputation option has a flag connected with it to let the analyst know how a specific field was imputed.

Micro-, Macro-, and Statistical Editing

For the most part, complex editing is very survey-specific. Most of

the more involved editing is done at Census headquarters in-

Washington. One of the most important aspects of the entire survey process is the Standard Industrial Classification (SIC) coding. Each establishment must be coded into a single category, and each category has its own set of ratio limits related to that type of 'industry activity.

The SPEER system was used for inter-item editing for the 1987 Enterprise Summary Report and the 1987 Auxiliary Establishment Report. Most of the editing that takes place in the complex edit deals with ratio edits. There is a front end program which takes the explicit ratios and calculates implicit ratios. This front end program is run only once unless the explicit ratios need to be changed or updated. Explicit ratios are user-supplied and each basic item must be contained in at least one ratio. The original ratios may be modified or the program may find them to be inconsistent. If the ratios are found to be inconsistent, the explicit ratios must be revised before editing begins. This process of correcting inconsistent explicit ratios typically takes only one iteration. These implicit ratios are then used as input for the SPEER program and are used for editing purposes as well as at the time of imputation.

Below is a brief description of the general flow for the batch portion of the SPEER system. Basic items are edited against each other simultaneously to check for inconsistencies (that is, ratios not within SIC-based limits). In the event of an edit failure or failures, SPEER determines which basic item(s) will be deleted and marked for imputation depending on how many times a field is involved in an edit failure and the reliability weight of that field. Fields not deleted are mutually consistent. An imputation range for the deleted field(s) is then calculated using all other basic items that are reported and unchanged or imputed earlier, and the implicit ratios. This imputation range will ensure a value will be imputed for the missing field that is consistent with all other basic items present. Methods of imputation are determined by the subject matter specialists.

Satellite items are related to one or two basic items, but not all basic items. Satellite items are edited against one other field not in the satellite group, typically a basic item, and always a field that has already been edited. The field which is used to edit the

satellite item is called a primary indicator. If a satellite item fails the ratio edit with its primary indicator, it is targeted for change and then imputed. Complex editing for satellite items follows along the same lines as complex editing for basic items, but in general is a little less complicated.

Micro-editing is not the only type of editing that goes into the survey process. Prior to release of data. tables go through an editing routine of their own which relates current summary values to those in the prior reporting period. After corrections are made to the data, each industry is then tabulated again and sent through a disclosure analysis system before data is released to the public.

Prioritizing of the Edits

The fields on the questionnaire are divided into basic items and non-basic items, also called satellite items. Basic items are typically a small number of fields that are fundamental to the establishment's operations and, for the purpose of editing, are related to one another. Items in each

satellite are related to one another and are grouped together and for that reason are typically not related to items in other satellites.

Imputation Procedures

Missing fields or inconsistent fields that have been deleted are imputed on an item-by-item basis. Any value imputed for a specific field must fall within the imputation range, as mentioned above. There are four major imputation methods used: 1) rounding of dollar values, 2) administrative data substitution, 3) sum of the details substitution and 4) industry average tolerances. Dollar values are typically reported in thousands of dollars, but at times a respondent will report in actual dollar figures. A first impute would then be to divide a particular dollar field by 1000. Administrative data is available for the major fields on the questionnaire, such as Number

of Employees (EMP), Annual Payroll (APR) and 1st Quarter Payroll (QPR). An example of sum of the details substitution would be substituting Rental Payments for Buildings (RPB) + Rental Payments for Machinery (RPM) for the missing field Total Rental Payments (RPT). Imputation through industry average tolerances is simply calculating an average value for a certain ratio. For example, setting EMP = APR * the industry average of the ratio EMP/APR.

Interactive, On-line Processing

All records are edited using the batch version of the SPEER program on the SPERRY mainframe. Referral records, approximately 20% - 30% of the universe, are then flagged for analyst review. Analysts access these referrals, one at a time using micro-computers, through an interactive, online version of the same batch program.

Using the previous method of reviewing referral records, it would literally take weeks to correct a record. Using the SPEER programs for the 1987 edit, analysts review and correct a referral record at one sitting. They can enter a value for one field, or an entire record, and are able to edit that record within a manner of minutes

and need no further editing.

The on-line version of the SPEER edit is a menu-driven program which starts out with a generic menu and is then tailored to a user's specific needs. Options a user may incorporate are those which make correcting referral records easier and quicker for the survey under review. For example, an option exists that enables the analyst to insert administrative data for those fields it is available for, and then impute an entire record from that data. is useful if an analyst determines a record to be completely unreliable.

Editing and Imputation Standards

The frequencies of failed edit ratios are tabulated to determine if the ranges for the ratios are either too wide or too tight. Analysts use this information to adjust the ratios, if needed. Role of Subject Matter Specialists

Subject matter specialists are essential throughout the editing process. They determine the applicable edit rules and the initial

explicit ratio parameters, which in turn, drive the entire SPEER edit. They determine the imputation methods needed for missing and deleted fields. The

interactive, on-line system was created with the idea that subject matter specialists would be the end users.

Current and Future Research

One area of research involves adapting the SPEER system for macro-editing, that is, editing summary data at various geographic levels. Another area of research focuses on methods to assist subject matter staff in setting ratio edit bounds. In addition, Census is considering ways to incorporate additional imputation options, such

as the hot-deck procedure, into SPEER.

Energy Information Administration

The Weekly and Monthly Refinery Reports

Environment in Which the Survey Takes Place

The Petroleum Supply Division (PSD) of the Energy Information Administration (EIA) operates an information collection and dissemination system, called the Petroleum Supply Reporting System (PSRS). The PSRS includes one annual, seven monthly, and five weekly surveys. These surveys track the supply and disposition of crude oil, petroleum products, and natural gas liquids in the United States. Two surveys, the Monthly Refinery Report (Form EIA-810) and the Weekly

Refinery Report (Form ElA-800), will be used as prototypes to describe editing practices.

The monthly survey is a complete census, while the weekly survey is a sample (151 sampling units) from the universe of 255 petroleum refineries and blending plants that report to the monthly survey. The refinery universe is relatively small, but it is ever-changing due to company births, deaths, mergers, and splits. In order to maintain a survey frame that is current and complete, the frame is updated continuously, and a comprehensive investigation of the adequacy of the frame, an exhaustive research activity to identify new eligible respondents, is conducted triennially.

Unlike the monthly survey which collects data on as many as 50 variables, the weekly survey collects data only on crude oil, motor gasoline, jet fuel, distillate fuel oil, and residual fuel oil. Most of the data reported in the weekly survey are estimated by the reporting companies, while data reported in the monthly survey are based on company accounting records. Inventory data are reponed as of the end of the reference period. Data on inputs, production, receipts, and shipments show the total volume of activity for the

reference period. All quantities are reported in thousand barrels (42 U.S. gallons per barrel). Zeros often dominate the response, i.e., not all of the units produce and/or store all products. The distribution of the petroleum supply variables is highly skewed, i.e., there are many small units and few large ones.

Response rates for both surveys are very high, often above 95 percent for the weekly survey and range from 99 to 100 percent for the monthly survey.

Editing Practices

Data Processing Environment and Dispersion of the Work

The reference period for the weekly survey extends from 7 a.m. Friday to 7 a.m. the following Friday. Weekly estimates are published on Thursday following the close of the reference week. They are also used to calculate early preliminary monthly estimates. The reference period for the monthly survey begins 12 a.m. of the first day of the month and ends midnight of the last day of the month. Monthly

aggregates are published in preliminary form 60 days after the close of the reference month. Final aggregates, reflecting any necessary corrections. are published six months after the close of the calendar year.

When the monthly survey forms are received at ElA. they are reviewed, primarily for identification information (ID number, company name, date, etc.). The data on the forms are then key entered.

As part of key entry, data are checked for errors. In 1989, the mainframe technology previously used to process survey data was augmented with personal computers configured in a local area network (LAN). The LAN/mainframe system facilitates the on-line processing of surveys and publication of survey statistics.

Weekly data are processed via a Computer Assisted Telephone Interviewing (CATI) system. This is a menu-driven system where data are entered, edited, and updated on-line. It also automatically schedules and logs telephone calls to respondents and collects performance statistics during data collection and processing activities. About one-half of the responses to the weekly surveys are received by telephone; the remaining responses are submitted by telefax or mail.

Audit Trail

For both surveys, a complete audit trail is kept for all machine edits. Reports which show in tabular forms different types of performance statistics, are available to in-house reviewers.

Micro-, Macro-, and Statistical Editing

Two types of automated edit checks are performed in the monthly survey:

1. Consistency checks: these are designed to detect arithmetic errors, e. g., row (column) total is not equal to sum of components. They are also designed to detect data indicating the occurrence of impossible events.

2. Range checks: these are designed to detect reported values that are theoretically possible but significantly different from the company's historical reporting pattern. Specifically, these detect if the current reported value is significantly different from (1) the average of the non-zero values in the past 12 months, (2) the previous month's value, and (3) the previous month's value adjusted for the difference between the same two months in the previous year.

There are three types of data edit checks performed in the weekly survey:

1. Consistency checks: these are designed to verify the internal consistency of data on a form.

2. Frequency checks: these are designed to flag items reported as

zero if the company has usually reported nonzero quantities and to flag nonzero items which have usually been reported as zero.

3. Outlier checks: These are designed to flag nonzero values that are unusually large or small for a given company.

Frequency and Outlier checks are developed based on the simple exponential smoothing technique.

For both surveys, each item failing a check is assigned a flag, which is coded to indicate the severity of the failure. These flags remain associated with the item until the questionable value has been verified or otherwise corrected. The most severe failures (as indicated by the flag codes) are

corrected first. Resolution of an edit flag takes one of two forms. The suspicious datum may be verified as correct, despite its differences with that company's reporting history and the edit flag is overridden, or it may be identified as erroneous, and replaced with the correct value. Verification can be accomplished by calling the reporting company, or by employing information from other sources that verifies the value as correct.

Monthly aggregates (weekly estimates) are visually compared with the previous month's aggregates (week's estimates) to search for anomalies. This kind of inspection helps to trace suspicious data that have an impact on the aggregates. Subject matter specialists review aggregate level data and determine where further checks are required.

Both monthly and weekly surveys are designed to measure the same phenomena, only for different time intervals. Therefore. an on-going comparison of data submitted by individual companies on the weekly and

monthly forms is routinely conducted. Historical reporting patterns for both monthly and weekly surveys are compared and facilities that systematically report different values are identified and contacted. In addition, to monitor values of key products reported in the weekly and monthly surveys. graphical comparisons are drawn between monthly aggregates for a particular variable and the monthly value derived from the weekly estimates for that product.

Every year, a "Quality Control Notebook" is prepared to document the steps being taken to improve data quality. evaluate current data quality, summarize current research, and establish an agenda for future enhancements of studies.

Imputation Procedures

Imputation in the weekly system takes place during estimation procedures. The estimation program imputes a value if it encounters datum with a flag indicating nonresponse or a critical error. The imputed value used is the exponentially smoothed mean weighted average if the frequency of a nonzero response (probability of nonzero

response) is equal to or greater than 0.5, and zero otherwise.

Imputation in the monthly system takes place as an independent processing step before summary tables are produced. The program computes only for nonrespondents. Imputed data are kept on the data files. There is a code associated with each data element to indicate whether the value is actually reported or imputed. The imputation procedure used is to insert the previous month's values (whether they are actually reported or imputed) for missing data.

Costs of Editing

The day to day operational cost of the edit system, including follow-up phone calls. is roughly estimated at 10 percent of the survey costs.

Current and Future Research

In 1988, PSD developed an electronic data communications software package, called the Petroleum Electronic Data Reporting Option (PEDRO), that allows respondents to transmit edited data to EIA's mainframe computer. PEDRO uses a personal computer to display the image of a printed survey form; the user (respondent) can then enter the data via the keyboard or import them from another computer system. PEDRO can perform the automated edit checks described above and flag errors. PEDRO also automatically dials EIA's central computer and uploads the data. Currently, PEDRO is used by 61 respondents to the Monthly Refinery Report. PEDRO has increased the timeliness and accuracy of the data submitted in addition to reducing respondent burden by eliminating paperwork, providing immediate on-site correction of data errors, and reducing the need for follow-up calls and data resubmissions.

Currently, the PSD is working on improving the efficiency of PEDRO and expanding it to cover more of the petroleum monthly and weekly

surveys.

APPENDIX C

CHECKLIST OF FUNCTIONS OF EDITING SOFTWARE SYSTEMS

A. USING THIS APPENDIX

This is a list of possible functions and attributes of editing software systems. It can be used as an aid in evaluating editing software. Explanations of some of the more technical terms are given in brief notes in the checklist and in the glossary. In addition, Chapter IV on editing software discusses the main areas of editing systems' development and the capabilities and limitations of types of

editing systems. While the evaluation of editing systems is a time consuming procedure, the effort should pay off by reducing costs in the future. An existing system developed by another organization may satisfy most requirements. The cost of adapting an existing system may be a fraction of the cost of developing of a new one. Even if existing systems cannot be used, an inspection of their capabilities would broaden the organization's perspective when consider development of its own system.

Evaluation of systems should be done by a team of people drawn from various parts of the organization. At a minimum, one person each from the data processing, methodology, research, and end-user departments should be represented. They should start with an evaluation of the checklist itself and determine which of the over 200 items apply to their organization. Some systems can be quickly eliminated based on the answers to the GENERAL FEATURES section. For example, an editing system may not work on all operating systems. (This may eliminate procurement of the system, but it may still be worthwhile to review it for its other features.) Evaluation may proceed by reading descriptions of the systems such as presented in Chapter IV of this

report, reviews of systems (e. g., Cotton, 1988), and systems documentation provided by the developers. Other good software review techniques include: attending a demonstration of a system, and running an application provided by the developers on diskette.

For the few systems that show promise, a complete evaluation will require that the software be acquired and that edits from a questionnaire (or section(s) of a questionnaire) be programmed. Data should be run through the system. The system should be evaluated on the items that have been deemed important. It may be necessary to make notes for most features. That is, it is often not enough to merely check yes or no for a function. The way the system carries out a function may be as important as whether it carries it out. If the system still looks promising, the evaluators should contact system users in the originating organization, not those who have developed it, and get the users' opinions about the system. Do they like it? If yes or no, why? Is it easy to use? How is the survey flow managed with the system? At this point, a trip to the developing organization may be necessary. It should be determined whether the system will be supported and if future updates are planned.

If you have evaluated a system by obtaining the software and trying it out, take the time to write to the developers with comments on the system, both good and bad. This will repay the developers for spending time on your requests. This feedback will generally be used to improve the systems.

Even if your organization rejects the system, keep an eye on its developments. Your comments may well have been taken into account by the developers and incorporated into the system in a subsequent release. Likewise, do not reject a system just because it is missing a few key items. First ask the developers ff they plan to add the necessary features. If not, try to determine what it would take to add them yourselves or to have them added. The cost of developing your own system may exceed one million dollars and take years to do.

1. GENERAL FEATURES

1.1 Can the system handle:

Categorical data? (A number represents a category in a group.

For example, 1 = female, 2 = male.)

Continuous data? (A number represents a quantity.)

Decimal values are allowed?

Character data? (Text is accepted as data. For example, a

comment may be allowed, or the words "female" or "male" may be

used directly.)

1.2 Can the system be used for Economic surveys?

Agricultural production surveys? Social surveys?

Demographic surveys?

1.3 Is the system a(n):

Editing system only?

Editing and imputation system?

CATI (Computer Assisted Telephone Interviewing) system?

CAPI (Computer Assisted Personal Interviewing) system?

Data entry system?

Tabulation system?

Data analysis system?

Summary system, (estimates, tables, matrix, manipulation, etc.)?

1.4 One part of a larger system?

1.5 Can the system be used on the following computers:

Mainframes?

Mini-computer?

Super-mini (e. g., Sun workstation)?

Stand-alone microcomputers?

Microcomputers in a Local Area Network (LAN)?

1.6 Can the system be used on the following operating systems:

MS DOS?

OS/2?

UNIX?

MVS?

CMS?

VMS?

Other (please specify)?

1.7 Does the system require other software for the developer or end user.

ORACLE?

Turbo Pascal?

SAS?

Fortran Compiler?

Expert System Shell?

Other software (please specify)?

1.8 For a microcomputer:

Does the system require a numeric co-processor?

Can the system use a numeric co-processor?

Does the system require more than 640 kilobytes of Random Access Memory (RAM)?

Are there other special requirements such as:

Extended memory?

Graphics monitor?

1.9 Are changes in the data made primarily by:

Human data editors?

Computer?

Both have important roles in correcting data?

1.10 Is the system available to others? (Is the organization willing

to sell it or give it to other organizations?)

1.11 Is the system in the public domain?

2. SURVEY MANAGEMENT

2.1 Can the system check-in records against a sample master?

2.2 Does the system have a "missing reports" capability?

2.3 Can the system provide data on the cost of the editing process?

2.4 Can the system indicate that records are in need of a recall or

re-check?

2.5 Can the system determine:

which individual records belong to a particular questionnaire?

That all records of a particular questionnaire are present?

That extra records are present?

2.6 Does the system have an audit trail? That is, can it track changes made (and possibly the reasons for them)?

2.7 Is information provided on the impact of the system on estimates?

2.8 Is information provided on contributions from nonresponse?

2.9 Regarding reports:

Can the system provide reports?

Can the specialist select which reports are to be generated?

Can the specialist select geographic regions or other breakdowns?

Can the specialist select records by values of variables?

Can the specialist request a dump (listing of values) of records?

Can the system determine the number of times each edit rule failed?

Per record?

Per file?

Can the system determine the edit rules that failed for each questionnaire? Can the system determine the frequency with which each field is involved in an edit failure? Can the system determine the frequency with which each field is identified for change? Can the system determine the number of times particular donor values were assigned to a variable? (This is important in hot- or cold-deck imputation where there are few donors and many recipients.)

Can the system provide information on enumerator performance?

3. SYSTEMS ITEMS

3.1 Are corrections made to the original file?

3.2 Are corrections made to a copy of the original file?

3.3 Are original data values part of the final record?

3.4 Is the audit trail part of the final record?

3.5 Can the following items be extracted from the file: Any combination of data from the original input file? Any generated variable?

3.6 Can the system generate setups in the following formats:

ASCII?

ORACLE?

DBase?

SAS?

SPSS?

Spreadsheet program (please specify)?

Other (please specify)?

3.7 Can the user program a customized format (e. g., to produce a data set for statistical or database software not specifically provided for by the system)?

3.8 Does the input file description include: Record layout?

Field specification?

Length, type, and other attributes of each variable?

Field that determines the record type?

Fields that identify the record?

Fields that identify geographic regions or other breakdowns?

3.9 Is there a coding module (e. g., the system has a catalogue of codes for occupations)?

Is there a step-wise coding module (the coder negotiates through a hierarchy of codes until the proper code is obtained)?

Is there a dictionary coding module (the open-ended answer is used to search for a similarly spelled coded answer)?

If both a step-wise coding module and a dictionary coding module are available, can they be used in combination?

3.10 Is the system a:

Compiled system? (The programming code is translated into low-level machine language for fast execution. The user does not have to have a copy of the language in order to use the editing system. Thus the execution of the program is faster and more flexible than in an interpreted system. However, changes to the code must be re-compiled in order to be put into effect.)

Interpreted system? (The translation of the programming code into a low-level machine language takes place on the fly, (as the code is used). Execution in this mode is slower than in a compiled language. In addition, the user must have a copy of the software in order to use the editing system. However, updates to the code do not have to be re-compiled to be put into effect.)

3.11 Is the system constructed in modules? (This allows parts of the system to be updated without affecting the other modules. This eases the development of updates and allows prototypes to be tried for specific functions.)

3.12 Does the system have a "specifications generator"? (This allows data, routes, edits, and other information to be specified once and used in other parts of the processing system.)

3.13 Does the system have a menu for:

End users?

Developers?

3.14 Is the system secure? (Does it require passwords or other measures in order to use it?)

3.15 Is the system embedded in a database?

Is it linked to data from a sampling frame?

Is it linked to historical data from a previous survey?

Is it linked to previous summary level data?

3.16 Can the system accept input data from the following sources:

CATI

CAPI

Data entry software

Data files generated from other software

ASCII formatted data

User specified format

Other (please specify)

3.17 Does the system generate a log file? (The system keeps track of every key-stroke.)

3.18 Does the system have specialized storage structures optimized for retrieval functional

3.19 Does the system promote statistical defensibility by:

Supplying the user with defensible modules and methods?

Relying on the integrity and expertise of the staff

Providing an audit trail?

3.20 Does the system allow "calls" to some user programmed sub-routine?

4. EDIT WRITING

4.1 Do specialists enter edits directly into system (perhaps using an integrated text editor or screens that have been provided)?

Do specialists specify edits that others enter into the system?

What programming language is used for edit specification:

Fortran

SAS

System specific language

Other (please specify)

4.2 Can edits be tested:

Interactively or on-line (i.e., easily and without waiting for paper output)? In batch (i.e., must wait for paper output)?

4.3 Are acceptable values specified (e. g., if 2 < X then OK)?

4.4 Are unacceptable values specified (e. g., if X ó 2 then fail)?

4.5 Does the specialist design the layout of a computer screen?

Can the specialist choose which variables to flag for each edit failure? Can historical variables appear on the screen?

Can calculated variables appear on the screen?

Can variables be protected from change?

Can the specialist control where the variables appear?

Can colors, fonts. and highlighting be specified?

4.6 Is a table format available?

Can the analyst move freely within the table?

Is it possible to specify one line of a table (edits, routes, etc.,) then specify the number of times the line is to be repeated (this feature will save much programming time)?

4.7 When messages appear on a computer Screen, are:

Error signals flagged at failing variables? (An error signal is an indication that an edit has failed. This may be a flag or a number next to the variable on the screen, or it may be a status code associated with the variable.)

Error signals given at the bottom (or top) of the screen?

Error messages viewed through pop-up windows?

Error messages viewed at the bottom (top) of the screen?

Can labels be programmed that make the screen more readable?

4.8 Does the specialist design the layout of a computer printout?

4.9 When messages appear on a computer printout:

Do messages appear with failing variables?

Can the specialist choose variables to be printed out for each error signal? For a variable involved in X error signals, will the value be printed out X times?

Once?

4.10 Are messages in coded (E.g.,error 333)?

4.11 Are messages in a spoken language (e. g., in English as opposed to a code)?

4.12 Can calculations of auxiliary variables appear in messages?

On screen?

In messages?

5. TYPES OF EDITS

5.1 Can the system handle edits with the following mathematical properties:

Linear edits, (Their graphical representations are straight lines. Linear edits are more amenable to edit analysis than nonlinear edits.)

Nonlinear edits

Conditional edits (They are not triggered until a condition is met. These are difficult to handle in edit analysis.)

Ratio edits (e. g.. 1 ó a / b ó 2)? (These can be restated linearly or ratio edits can be analyzed within themselves.)

Variables that can accept negative values?

Quantitative edits?

Qualitative edits?

5.2 Can the system handle edits with the following functions: Valid values checks. (univariate range checks)? Consistency checks, (multivariate. record-level checks):

Consistency "balance" checks (i.e., parts (or details) add up to total)?

Other consistency checks (e. g., marital status vs. age)?

Route checks and skip patterns (i.e., the proper path was followed through the form,

including skips based on certain values)?

Generated values against external standards (e. g., average wage ò minimum wage)?

5.3 Can the system accept logical operators, e. g., "and", "or", and "not".

5.4 Is there just one level of edit failure or priority?

5.5 Are there two or more levels of edit failure or priority (e. g., critical vs. noncritical)?

5.6 Can the system perform non-statistical cross record checks? For example, can it check values in subfiles against values in the master file?

5.7 Can states, provinces, other regions, or strata have customized error limits?

5.8 Can the system perform statistical editing: Based on historical data?

Based on cross-record checks:

univariate outlier detection?

multivariate outlier detection?

5.9 Can the system handle record-level edits:

Based on frame data?

Based on previous reports (nonstatistical historical editing)?

5.10 Can the system perform macro editing (editing at an aggregate level): At the summary level?

At the stratum level?

Using historical data?

And allow tracing back to individual records?

And perform on-line correction of records and recalculation of

aggregate level statistics?

5.11 Can the system perform graphical inspection of the data?

On paper printouts?

Interactively on computer screens?

At an aggregated (macro) level?

And allow tracing back to individual reports?

6. EDIT RULE ANALYSIS

6.1 Can the system check edits for redundancy? (A redundant edit does

not further restrict the set of acceptable values.)

6.2 Can the system check that edits are not contradictory? (If two edits are contradictory, then no record will pass the edits.)

6.3 Can the system generate implied edits? (Implied edits are derived by implication from explicitly written edits, e. g., if $1 < a/b < 2$ and $2 < b/c < 4$ are explicit edits, then $2 < a/c < 8$ is an implied edit. Implied edits allow the specialist to check that the data is not being restricted in an unintended way. They are also important for determining which variables to correct.)

For linear edits?

For ratio edits?

For other nonlinear edits (e. g., conditional edits)?

6.4 Can the system generate external records? (The vertices of the acceptable region are generated for inspection. For example, A = 0, B = 1000, and C = 2 may be the vertices of an acceptable region. This analysis will allow the specialist to examine the worst possible records that can pass through the edits without an error signal being generated.)

6.5 Can the system analyze groups of edits separately (some edits may be applied to only certain questionnaires depending on coded values or reported data).

7. DATA REVIEW AND COMMON

7.1 Does the end-user access the system through menus?

7.2 In the computer processing flow, does the system:

Require substantial cleaning up of the data before it is applied (either manually or by another system module)?

Require substantial cleaning up of the data after it is applied (either manually or ty another system module)?

Make (almost) all corrections automatically in batch?

Require manual correction of machine generated signals?

Have batch capability with manual resolution of records in error?

7.3 Does the system allow manual review and changes: Before data

entry?

During data entry?

After data entry?

7.4 Are manual changes made:,

Interactively through terminals or work stations?

Is there full screen mobility?

Is there full between-screen mobility?

Are changes made through prompts that appear at the bottom (top)

of the screen?

In a batch only environment using paper printouts?

Are error messages generated on paper but changes made on a

terminal?

7.5 Does the system determine which variables are to be changed? This

is called error localization.)

For all variables and edits?

For subsets of variables and edits?

7.6 Does the system allow automatic correction (imputation)?

Does it allow simultaneous multivariate imputation?

Does it allow several imputation options per variable?

Does it do deterministic imputation (where only one value will work)?

Does it do hot-deck donor imputation? (Values of records from the current survey are used to impute for missing values in recipient records.)

Does it do cold-deck donor imputation? (Values of records from a previous survey are used to impute for missing values in current recipient records.)

Does it perform administrative (frame) imputation?

7.7 Does the system have other integration features such as:

Consistent menu formats between modules?

Historical integration, (the ability to update questionnaires

easily from one period to the next)?

Can edit specification for different parts of the questionnaire

be done independently by two or more people?

Are associated software packages available that can be used with

the editing system for related functions such as tabulation or

survey management?

7.8 In the data entry program, can:

Data be entered and edited at the same time?

Data be entered without correction? (High speed, heads down.)

Full screen data entry be done?

Item code data entry be done?

Verification be done? (Data are entered twice.)

Implied decimals be typed? (e. g., type 23 for 2.3.)

For certain variables that have a predetermined length, can the

typist go on to the next value

without pressing < Enter >.

Are data entry statistics generated?

7.9 Can the system handle complex data structures such as:

Hierachical data (e. g., individual questions asked within a household)?

Complicated skip patterns?

Can the system handle subfiles (the variables are split up between different files, making further processing easier in some cases, allowing data to be stored more efficiently in other cases)?

Is cross record checking allowed (e. g., between lines of a subfile from a table of questions and the master record)?

7.10 Does the system allow respecification of edits without having to manually redo other parts of the processing software? (For example, can changes in edits be automatically reflected in data entry, data editing, and data collection modules without

respecifying changes in each module?)

7.11 Does the system allow respecification of data without having to manually redo other parts of the processing software? (Same note as in 7.10.)

8. SUPPORT, UPDATES, AND TRAINING

8.1 Is the system updated systematically (e. g., future updates planned)?

8.2 Is there a cost for the update?

8.3 Is the system supported with training?

8.4 Is there a cost for the training?

8.5 Is the development environment of the system available in:

English?

French?

Dutch?

Spanish?

Can it be convened to other languages?

8.6 Does the system developer offer support?

8.7 Is there a. cost for the support?

8.8 Is the system comprehensible to the end user? (That is, can the end user understand what the system is doing methodologically speaking? It is not just a black box.)

8.9 Could one person use all aspects of the system for a small one-time survey without much support (this is an indication of the ease of development and use)?

8.10 Is there on-line documentation?

A help screen of keys only?

Is the on-line documentation "context sensitive"?

8.11 Do the written documentation items include:

Installation instructions and "getting started" information? A

tutorial for novice applications programmers?

A user's guide?

A reference manual?

A manual of examples?

A discussion of the overall approach used by the system in

editing a survey?

8.12 What are there capacity and performance limits for The number of

edits?

The number of variables?

The number of questionnaire versions?

The number of executable statements?

9. References

A Comparison of Software for Editing Survey and Census Data. Paul Cotton. Presented October 1988, at the Statistics Canada Symposium, The Impact of High Technology on Survey Taking. (The topical division of this checklist and about half (perhaps more) of the items are taken from this report.)

An Evaluation of Edit and Imputation Procedures Used in the 1982 Economic Censuses in Business Division. Brian Greenberg and Thomas Petkunas, Statistical Research Division, Bureau of the Census, 1986.

A Review of the State of the Art in Automated Data Editing and Imputation. Mark Pierzchala, National Agricultural Statistics Service, United States Department of Agriculture, 1988.

CAI Software, An Evaluation of Software for Computer Assisted Interviewing. Steven E. de Bie, Ineke A. L. Stoop, and Katrinus L. M. de Vries. VOI, Association of Social Research Institutes, Data

Collection Group. Amsterdam, the Netherlands, March 1989.

The checklist was written in its present form by Mark Pierzchala based first on Cotton's work and then modified by reference to the other papers. It was revised with the help of the members of the Software Subgroup. Many other people have reviewed it including other members of the Subcommittee on Editing in Federal Statistical Agencies and members of the Data Editing Joint Group, of the Statistical Computing Project Phase 2, of the United Nations. A version of this checklist was provided to latter group. Various other reviewers of this report have also added some comments.

APPENDIX D

ANNOTATED BIBLIOGRAPHY ON EDITING

A. LISTING OF PAPERS BY TOPIC FOR:

[A] GEIS

[G] BLAISE

[C] SPEER

[D] TIME SERIES AND OUTLIER DETECTION APPLICATIONS

[E] ERROR LOCALIZATION METHODS

In the annotated bibliography each paper belonging to a special topic [A]-[E] will be so indicated. The annotations are brief and are only intended to give a very general idea of paper content. If the content is clear from the paper title. no annotation is provided.

Specialized Bibliography

[A] GEIS PAPERS

Fellegi and Holt (1976); Giles (1986,1987,1988); Giles and Patrick (1986); Greenberg (1987b); Hidiroglou and Berthelot (1986); Kovar (1991); Kovar, MacMillian and Whitridge (1988); Nesich (1980); Sande, G. (1976,1978,1981), Sande, I. (1988); Schiopu-Kratina and Kovar

(1988).

[B] BLAISE PAPERS

Bethlehem (1987); Bethlehem, Denteneer, Hundepool. and Keller (1987); Bethlehem, Hundepool, Schuerhoff and Vermeuler (1987); Denteneer, Bethlehem, Hundepool, and Schuerhoff (1987a,1987b); Greenberg (1987b); Keller and Bethlehem (1990).

[C] SPEER PAPERS

Greenberg (1981,1982,1987a,1987b); Greenberg and Petkunas (1987); Greenberg and Surdi (1984).

[D] TIME SERIES AND OUTLIER DETECTION APPLICATIONS

Bell (1983); Burns (1980,1983); Cherriick (1982,1983); Chemick, Downing and Pike (1982); Chemick and Murthy (1983); Dinh (1987); Hill and French (1981); Kirkendall (1988); Little and Smith (1983,1987); Mazur (1990); Miller, Meroney and Titus (1987); Passe, Carpenter and Passe (1987); Pierce and Bauer (1989).

[E] ERROR LOCALIZATION METHODOLOGIES

Fellegi and Holt (1976); Garcia-Rubio and Villan (1990); Garfinkel

(1979); Garfinkel, Kuruiathur, and Liepins (1986); Greenberg (1981); Liepins, Garfinkel and Kunnathur (1982); Liepins and Pack (1980,1981); Little and Smith (1983,1987); McKeown (1984); Sande (1978); Schiopu-Kratina and Kovar (1988).

B. ANNOTATED BIBLIOGRAPHY

ABRAHAM, B. and YATAWARA N. (1988), "A Score Test for Detection of Time Series Outliers," Journal of Time Series Analysis, 9, 109-119. [D] Outlier detection in time series; not typically employed in survey editing.

ASHRAF, A. and MACREDIE, I. (1978), "Edit and Imputation in the Labour Force Survey," in Proceedings of the Section on Survey Research Methods, American Statistical Association, 425-430. Describes edit and imputation system used in Canadian Labor Force Survey.

BAIR, R. B. (1981), "CONCOR: An Edit and Automatic Correction Package," incomputer Science and Statistics: Proceedings of the 13th symposium on the Interface, 340-343. Automated System developed at the Census Bureau primarily focusing on validity edits and their correction.

BANISTER, J. (1980), "Use and Abuse of Census Editing and Imputation," Asian and Pacific Census Forum, 6. Raises concerns about excessive editing and imputation in censuses.

BELL, W.R. (1983), "A Computer Program for Detecting Outliers in Time Series," in Proceedings of the Section of Business and Economic Statistics, American Statistical Association, 634-639. [D] Categorizes three major causes of outliers in a time series framework and presents tests sensitive to each.

BETHLEHEM, J. G. (1987), "The Data Editing Research Project of the Netherlands Central Bureau of Statistics," in Proceedings of the Third Annual Research Conference of the Bureau of the Census, Washington,

D.C.: U. S. Bureau of the Census, 194-203. [B] Discusses problems of editing at CBS and need for Blaise.

BETHLEHEM, J. G., DENTENEER, D., HUNDEPOOL, A. J., and KELLER, W. J. (1987). "The Blaise System for Computer-Assisted Survey Processing." in Proceedings of the Third Annual Research Conference of the Bureau of the Census, Washington, D.C.: U. S. Bureau of the Census, 194-203. [B] Report on Blaise.

BETHLEHEM, J. G., HUNDEPOOL, A. J., SCHUERHOFF, M. H. and VERMEULER, L. F. M. (1989). "Blaise 2.0 An Introduction." System Documentation, Netherlands Central Bureau of Statistics. [B] Introduction to the Blaise system for Computer Assisted Editing, Data Collection. and Data Entry.

BIEMER, P. P. (1980), "A Survey Error Model Which Includes Edit and Imputation Error," in Proceedings of the Section on Survey Research Methods. American Statistical Association, 610-615. Presents a survey error model which includes edit error, imputation error, response error and sampling error.

BILOCQ, F. (1989), "Analysis on Grouping of Variables and or Detection of Questionable Units." Business Surveys Methods Division, Statistics Canada. Groups of variables are identified to optimize editing. Questionable units identified by statistical techniques for change or follow-up.

BILOCQ, F., and BERTHELOT, J. M. (1989) "An Edit Scheme Based on Multivariate Data Analysis," Business Surveys Methods Division, Statistics Canada. Multivariate data analysis techniques for grouping items for editing.

BRANT, J. D. and CHALK, S. M. (1985), "The Use of Automatic Editing in the 1981 Census," Journal of the Royal Statistical Society (Series A), 148, 126-146. Describes the automatic editing and imputation

system used in the. 1981 census in England and Wales.

BURNS. E.M. (1980). "Procedures for the Detection of Outliers in Weekly Time Series," in Proceedings of the Section of Business and Economic Statistics, American Statistical Association, 560-563. [D] Compares five outlier detection procedures for repeated surveys. Each uses a different set of features from historical data to test incoming data.

BURNS, E.M. (1983), "Editing and Imputation for the EIA Weekly-Petroleum Surveys," in Proceedings of the Section on Survey Research Methods, American Statistical Association, 539-543. [D] Discusses use of exponential smoothing in editing EIA's weekly petroleum surveys.

CHERNICK, M. R. (1982), "The Influence Function and its Application to Data Validation," American Journal of Mathematical and Management Sciences, 263-288. [D] Hampel's Influence Function used to detect bivariate observations, which have unusual influence on estimates of correlation.

CHERNICK, M. R. (1983), "Influence Functions, Outlier Detection, and Data Editing," in Statistical Methods and the Improvement of Data Quality, ed. T. Wright, New York: Academic Press, 167-176. [D] Review article on influence functions to detect outliers.

CHERNICK, M. R., DOWNING. D. J., and PIKE, D. H. (1982), "Detecting Outliers in Time Series Data," Journal of the American Statistical Association, 77,743-747. [D] Effect of outliers on time series by considering the influence function for the auto correlations of a stationary time series.

CHERNICK. M. R., and MURTHY, V. K. (1983), "The Use of Influence Functions for Outlier Detection and Data Editing," American Journal of Mathematical and Management Sciences. 3,47-61. [D].

COTTON, P. (1988), "A Comparison of Software for Editing Survey and Census Data," in Proceedings of Symposium 88, The Impact of high Technology on Survey Taking, Ottawa, Ontario, Canada: Statistics Canada, 211-241. Criteria to evaluate editing software. Four systems that operate on MS-DOS microcomputers are reviewed.

CUSHING, J. (1988), "A Report on Recent Survey Processing in Developing Countries: The Demographic and Health Surveys Microcomputer Approach." in Proceedings of Symposium 88, The Impact of High Technology on Survey Taking, Ottawa, Ontario, Canada: Statistics Canada: 201-210. Describes the Integrated System for Survey Analysis (ISSA), a microcomputer survey processing system.

DENTENEER, D., BETHLEHEM, J. G., HUNDEPOOL, A. J., and SCHUERHOFF, M. S. (1987a), "Blaise, A New Approach to Computer Assisted Survey Processing," Staff Report, Netherlands Central Bureau of Statistics. [B] Discusses Blaise and contains Blaise related bibliography.

DENTENEER, D., BETHLEHEM, J. G., HUNDEPOOL, A. J., and SCHUERHOFF, M.

S. (1987b), "The Blaise System for Computer-Assisted Survey Processing," in Proceedings of the Third Annual Research Conference of the Bureau of the Census, Washington, D.C.: U. S. Bureau of the Census, 112-127. [B] Discusses Blaise.

DINH, K. (1987), "Application of Spectral Analysis to Editing a Large Data Base,"Journal of Office Statistics, 3, 431- 483. [D] Simple statistical procedure to minimize reporting and data entry errors in repeated surveys.

FELLEGI, I. P. and HOLT, D. (1976), "A Systematic Approach to Automatic Edit and Imputation," Journal of the American Statistical Association, 71, 17-35. [A],[E] Introduces "Fellegi-Holt" procedure for automated data edit and imputation; corner-stone paper in this area.

FREUND, R. J. and HARTLEY. H. O. (1967), "A Procedure for Automatic Data Editing," A procedure for the American Statistical Association, 62,341-353. Proposes an automated edit scheme for a large variety of surveys.

GARCIA-RUBIO, E., and VILLAN. I. (1990), "DlA System: Software for the Automatic Editing of Qualitative Data," in Proceedings of the Sixth Annual Research Conference of the Bureau of the of Census, Washington, D.C.: U. S. Bureau of the Census, (to appear). [E] Describes the DIA system for data editing and imputation based in part on Fellegi-Holt methodology.

GARFINKEL. R. S. (1979), "An Algorithm for Optimal Imputation of Erroneous Data," College of Business Administration Working Paper Series, The University of Tennessee, Knoxville. [E] Algorithm based on Fellegi-Holt for error localization.

GARFINKEL R. S., KUNNATHLTR, A. S., and LIEPINS, G. E. (1986), "Optimal Imputation of Erroneous Data: Categorical Data, General Edits," Operations Research. 34, 744-751. [E] Algorithm for error localization.

GILES, P. (1986). "Methodological Specifications for a Generalized Edit and Imputation System." Business Survey Methods Division. Statistics Canada. [A] Specifications for development of GEIS, with

proposed requirements and priority for development.

GILES. P. (1987). "Towards the Development of a Generalized Edit and Imputation System." in Proceedings of the Census Bureau Third Annual Research Conference of the Bureau of the Census Washington, D.C.: 185-193. [A] Methodological issues related to development and implementation of GEIS.

GILES, P. (1988), "A Model for Generalized Edit and Imputation of Survey Data." The Canadian Journal of Statistics, 16, Supplement, 57-73. [A] Focuses on methodology of GEIS.

GRANQLIST, L. (1982), "On Generalized Editing Programs and the Solution of the Data Quality Problems," UNDP Statistical Computing

Project, Data Editing Joint Group. Discusses impact of editing, need to concentrate on systematic errors and resource allocation.

GRANQLTIST, L. (1994), "On the Role of Editing," Statistisk Tidskrift, 2, 105-118. Guidelines on editing surveys of official statistical agencies.

GRANQUIST L. (1987). "A Report of the Main Features of a Macro-editing Procedure which is used in Statistics Sweden for Detecting Errors in Individual Observations." Report presented at the Data Editing Joint Group, Statistics Sweden.

GRAVES, R. B. (1976), "A Generalized Edit and Imputation System in a Data Base Environment," presented to the Electronic Data Processing Working Party of the Conference of European Statisticians, Geneva. Describes CAN-EDIT, an implementation of Fellegi-Holt methodology for categorical data.

GREENBERG, B. (1981), "Developing an Edit System for Industry Statistics," Computer Science and Statistics: in Proceedings of the

13th Symposium on the Interface New York: Springer-Verlag, 11-16. [C] First paper describing what has evolved into the SPEER system.

GREENBERG, B. (1982), "Using an Edit System to Develop Editing Specifications," in Proceedings of the Section on Survey Research Methods, American Statistical Association, 366371. [C],[E] Discussion of how edit system diagnostics used to develop edit criteria.

GREENBERG, B. (1987a), "Edit and Imputation as an Expert System," in Statistical Policy Working Paper Number 14: Statistical Uses of Microcomputers in Federal Agencies, Washington, D.C.: Office of Management and Budget, 85-92. [C] Describes the Expert System features in SPEER.

GREENBERG, B. (1987b), Discussion, "Session on Designing Automated Data Editing Systems," in Proceedings of the Third Annual Research Conference of the Bureau of the Census, Washington, D.C.: U. S. Bureau of the Census, 204-212. [A],[B],[C] Discusses paper by Giles on GEIS and by Bethlehem on Blaise.

GREENBERG, B., and PETKUNAS, T. (1987), "An Evaluation of Edit and

Imputation Procedures Used in the 1982 Economic Censuses in Business Division," in 1982 Economic Censuses and Census of Governments Evaluation Studies, U. S. , Department of Commerce. Bureau of the Census. [C] Work completed as part of Evaluation Studies Task Force for 1982 Economic Censuses.

GREENBERG, B., and SURDI, R. (1984). "A Flexible and Interactive Edit and Imputation System for Ratio Edits," in Proceedings of the Section on Survey Research Methods. American Statistical Association, 421-426. [C] Discusses actual uses of SPEER system and illustrates options based on user needs.

HIDIROGLOU, M. A. and BERTHELOT, J. M. (1986), "Statistical Editing and Imputation for Periodic Business Surveys," Survey Methodology,

12, 73-83. [A] Methods for editing and imputing data for.units contacted on a periodic basis.

HILL, C.J. (1978), "A Report on the Application of a Systematic Method of Automatic Edit and Imputation to the 1976, Canadian Census," in Proceedings of the Section on Survey Research Methods, American Statistical Association, 474-479. Rationale of edit and imputation system for the Canadian Census and a non-technical description of methodology.

HILL, E., Jr. and FRENCH, C. (1981), "Editing Very Large Data Bases," in Proceedings of Conference on Information, Science, and Systems, 70-78. [D] Algorithm which employs spectral analysis to classify historical data series and to edit data according to classifications.

HODGES, B.S. M (1983), "Using the Computer to Edit and Code Survey Data," in Proceedings of the Section on Statistical Computing, American Statistical Association, 238-240. Computer edits and codes single word or short phrase responses to open-ended questions.

HUGHES, P. J., McDERMID, I., and LINACRE, S. J. (1990), "The Use of Graphical Methods in Editing," in Proceedings of the Sixth Annual Research Conference of the Bureau of the Census, Washington, D.C.: U.S. Bureau of the Census (to appear). [D] Examines the potential for graphical methods in the different stages of editing, such as setting edit bounds and in output (aggregate) editing.

KELLER, W.J. and BETHLEHEM, J.G. (1990), 'The Impact of Microcomputers on Survey Processing at the Netherlands Central Bureau of Statistics," in Proceedings of the Sixth Annual Research Conference of the Bureau of the Census, Washington, D.C.: U. S. Bureau of the Census (to appear). [B] Discussion of Blaise implementation activities.

KIRKENDALL, NJ. (1988), "An Application of a Kalman Filter for Estimation and Edit". in Proceedings of the Business and Economic Statistics Section, American Statistical Association, 510-515. [D) Application of exponential smoothing, and Kalman filter models; for series with shifts in levels.

KOVAR, J. (1981), "Edit and Imputation Package for the Integrated

Agriculture Survey," Statistics Canada. [A] Summarizes IAS Edit/Imputation package and flexibility it may provide.

KOVAR. J. G. (1990), Discussion, "Session on Editing," in Proceedings of sixth Annual Research Conference of the Bureau of the Census, Washington, D.C.: U. S. Bureau of the Census (to appear). Discusses papers by Garcia-Rubio and Villan on DIA and by Hughes, McDermid and Linacre on graphical methods.

KOVAR, J. G., MacMILLAN J. H., and GE, P. (1988). "Overview and Strategy for the Generalized Edit and Imputation System." Working Paper No. BSMD-88-007E. Methodology Branch, Statistics Canada. [A] Overview of GEIS.

LIEPINS, G.E. (1983), "Can Automatic Data Editing Be Justified? One Person's Opinion," in Statistical Methods and the Improvement of Data Quality, ed. T. Wright, New York: Academic Press, 205-213. Issues concerning application of automatic data editing.

LIEPINS, G. E., GARFINKEL, R. S. and KUNNATHUR, A. S. (1982), "Error Localization for Erroneous Data: A Survey," TIMS Studies in Management Sciences, 19, 205-219.[E]Survey of Error Localization Techniques.

LIEPINS, G.E. and PACK, DJ. (1980), "An Integrated Approach to Data Editing," in Proceedings of the Section on Survey Research Methods, American Statistical Association, 777-781. [E] Simulation study on error localization for categorical data from large data sets.

LIEPINS, G.E. and PACK, DJ. (1981), "Maximal Posterior Probability Error Localization," in Proceedings of the Section on Survey Research Methods, American Statistical Association, 192-195. [E] Simulations suggest that MPPEL is not superior to minimal fields to impute.

LINACRE, S. J. and TREWIN, D. J. (1989), "Evaluation of Errors and Appropriate Resource Allocation in Economic Collections." in

Proceedings of the Fifth Annual Research Conference of the Bureau of the Census, Washington, D. C.: U. S. Bureau of the Census, 197-209. A study by Australian Bureau of Statistics of a near optimal allocation of resources to reduce nonsampling errors.

LITTLE, R.J. A., and SMITH, P.J. (1983). "Multivariate Edit and Imputation for Economic Data," in Proceedings of the Section on Survey Research Methods, American Statistical Association, 518-522. [D],[E] Edit and Imputation procedure based on outlying cases.

LITTLE, R J. A., and SMITH, P. J. (1987), "Editing and Imputation for Quantitative Survey Data," Journal of the American Statistical Association, 82, 58-67. [D],[E] Methodological tools include distance measures, graphical procedures, maximum likelihood and robust estimation for incomplete multivariate data.

MADOW, W.G., NISSELSON, H., and OLKIN, I. (eds.) (1983). Incomplete Data in Sample Surveys, Volume 1, Report and Case Studies. New York: Academic Press. (See below).

MADOW, W.G., and OLKIN, I. (eds.) (1983). Incomplete Data in Sample Surveys, Volume 3, Proceedings of the Symposium. New York: Academic Press. (See below).

MADOW, W.G., OLKIN, I., and RUBIN, D.B. (eds.) (1983). Incomplete Data in Sample Survives, Volume 2. Theory and Bibliographies. New York: Academic Press. Broad treatment of incomplete data with extensive, partially annotated, bibliography in Volume 2.

MAGNAS, H.L. (1989), "An Expert System to Assist in the Disposition of Computer Edit Error Flags." Presented at American Statistical Association Committee on Energy Statistics spring meeting. An expert system designed to standardize the follow-up procedure for flagged items.

MAZUR. C. (1990). "Statistical. Edit System for Livestock Slaughter Data," NASS Staff Report.' SRB-90-01, National Agricultural Statistics Service., U. S. Department of Agriculture. [D] Statistical edit in which a plant's historical data define edit limits. Classical and simple robust estimates and Tukey's biweight estimator used to

calculate edit limits.

MCKEOWN, P.G. (1984), "A Mathematical Programming Approach to Editing of Continuous Survey Data," SIAM Journal. of Scientific and Statistical Computing, 5, 784-797. [E] Mathematical programming approach to error locations

MILLER, R., MERONEY, W. and TITUS, E. (1987), "Identification of Anomalous Values,in Energy Data," Proceedings of the Business and Economic Statistics Section, American Statistical Association, 241-246. [D] Presents results of using Box-Whisker techniques to identify outliers in refinery production data.

NAUS, J. I. (1975), Data Quality Control and Editing, New York: Marcel Dekker. Variety of techniques for probabilistic edits and quality

control.

NAUS, J.I. (1982), "Editing Statistical Data," in Encyclopedia of statistical Sciences (Vol. 2), eds. S. Kotz, N. L. Johnson, and C. B. Read, New York: Wiley, 455-461. Review article.

NAUS, J.I., JOHNSON. T.G. and MONTALVO. R. (1972), "A Probabilistic Model for Identifying Errors in Data Editing," Journal of the American Statistical Association, 67, 943-950. Procedures for assigning likelihood of error given simultaneous failure of deterministic edits.

NESICH, R. (1980), "General Methodological Approach to Edit and Imputation in the 1981 Census of Agriculture." Statistics Canada. [A] Approach to be used in editing and imputing the 1981 Census of Agriculture data.

NORDBOTTON, S. (1965), "The Efficiency of Automatic Detection and Correction of Errors in Individual Observations as Compared with other Means for Improving the Quality of Statistics," in Bulletin of the International Statistical Institute, Proceedings of the 35th Session,

Belgrade 41, 417-441. Compares effect on survey results of editing to other methods to improve the quality.

PASSE, R. M., CARPENTER, M. J., and PASSE, H. A. (1987), "Operational Outlier Detection, Communications in Statistics (Theory & Methods) 16, 3379-91. [D] Outlier detection techniques.

PIERCE, D. A. and BAUER, L. L. (1989), "Tolerance-Width Groupings for Editing Banking Deposits Data: An Analysis of Variance of Variances," Finance and Economics Discussion Series, No. 72. [D] Determination of optimal clusters for setting edit tolerances.

PIERZCHALA, M. (1988), "A Review of the State of the Art in Automated Data Editing and Imputation," NASS Staff Report, SRB-88-10, National Agricultural Statistics Service, U. S. Department of Agriculture. Explores approaches to automation of data editing and imputation with focus on state of the art. Includes bibliography.

PODEHL, W. M. (1974), "Introduction to the Generalized Editing and Imputation System using Hot-Deck-Approach." General Survey Statistics

Division, Statistics Canda. Application of Fellegi-Holt methods to 1971 Census of Population and Housing.

PRITZKER, L., OGUS, J., and HANSEN, M. H. (1965), "Computer Editing Methods-Some Applications and Results," in Bulletin of the International Statistical Institute, Proceedings of the 35th Session, Belgrade, 41, 442-465. Summarizes philosophy, problems, and practice in editing and imputation at the Census Bureau.

PULLUM, T. W., HARPHAM, T., and OZSEVER, N. (1986), "The Machine Editing of Large-Sample Surveys: The Experience of the World Fertility Survey," International Statistical Review, 54, 311-326. Addresses trade-offs between improvements in data quality and costs of editing.

SANDE, G. (1976), "Numerical Edit and Imputation," invited paper

presented at the International Association for Statistical Computing, 42nd session of the International Statistical Institute, Manila, Philippines. [A] Precursor to GEIS.

SANDE, G. (1978), "An algorithm for the fields to impute problems of numerical and coded data." Statistics Canada Report. [A],[E] Introduces linear programming methods for error localization.

SANDE, G. (1981), "Descriptive Statistics Used in Monitoring Edit and Imputation Processes," presented at Workshop on Automated Edit and Imputation, Computer Science and Statistics; Proceedings of the 13th Symposium on the Interface, Springer-Verlag, New York. [A) Discusses descriptive statistics in precursor to GEIS.

SANDE, I. G. (1982), "Imputation in Surveys: Coping with Reality," The American Statistician, 36, 145-152. A variety of practical and methodological considerations in edit and imputation.

SANDE, I. G. (1988). "A Statistics Canada Perspective on Numerical Edit and Imputation in Business Surveys," presented at the Conference of European Statisticians in Geneva. [A] Discusses some ideas behind

GEIS.

SCHIOPU-KRATINA, I. and KOVAR, J. G. (1988) "Use of Chernikova's Algorithm in the Generalized Edit and Imputation System," Working Paper 89-001, Methodology Branch, Statistics Canada. [A],[E] Presents methodology in GEIS which uses linear programming to find fields to impute.

SIMINOFF, J. S. (1984), "The Calculation of Outlier Detection Statistics," Communications in Statistics, Simulation & Computation, 13, 275-285. Provides a program to calculate four outlier detection statistics.

STUART, WJ. (1966). "Computer Editing Survey Data - Five Years of Experience in BLS Manpower Surveys," Journal of the American Statistical Association, 375-383. A review of steps at BLS to computerize the screening process.

SZAMEITAT, K. and ZINDLER. H. J. (1965), "'The Reduction of Errors in Statistics by Automatic Corrections," in Bulletin of the

International Statistical Institute, Proceedings of the 35th Session, Belgrade, 41, 395-417. Discusses practical and theoretical considerations of reducing errors in statistical processing.

TORO, V. and CHAMBERLAIN, K. (1988). "The Use of Microcomputers for Census Processing in Developing Countries: An Update,"' in Proceedings of Symposium 88, The Impact of High Technology on Survey Training, Ottawa, Ontario, Canada: Statistics Canada, 181-199. Describes the Integrated Microcomputer Processing System (IMPS).

APPENDIX E

GLOSSARY OF TERMS

Acceptance region - The set of acceptable values defined by the edits

for each record. For categorical data the acceptance region can be represented as a set of lattice points in N-Space. For numerical data it is a set of convex regions in N-space. Also called a feasible region.

Audit trail - An accounting of changes to values in a field and the reasons for the changes.

Balance edit - An edit which checks that a total equals the sum of its parts. Also called an accounting edit. Example: Closing inventory = Opening Inventory + Purchases - Sales. (Example from Kovar, 1988)

Complete set of edits - The union of explicit edits and implied edits. Sufficient for the generation of feasible (acceptance) regions for imputation (that is if the amputations are to satisfy the edits).

Conditional edit - An edit where the value of one field determines the editing relationship between other fields and possibly itself. For example, suppose there are three fields A, B, and C. A conditional edit would exist ff the relationship between fields B and C as

expressed through the edits depended on the value in A.

Consistency edit - A check for determinant relationships, such as parts adding to a total or harvested acres are less than or equal to planted acres.

Consistent edits - A set of edits which do not contradict each other is considered to be consistent. If edits are not consistent then no record can pass the edits without invoking an error signal.

Deterministic edit - An edit, which if violated, points to an error in the data with a probability of one. Example: Age 5 and Status = mother. Contrast with stochastic edit.

Deterministic imputation - The situation when only one value of a field will cause the record to satisfy all of the edits. Occurs in some situations (such as the parts of a total not adding to the total). This is the first solution to be checked for in the automated editing and imputation of survey data.

Error localization - The automatic identification of the fields to

impute. That is, the determination of the al set of fields to impute for.

Explicit edit - An edit explicitly written by a subject matter specialist. (Contrast explicit edits with implied edits.)

Expert system - computer system that solves complex problems in a given field using knowledge and inference procedures, similar to a human with specialized knowledge in that field.

Failed edit graph - As used by the U.S. Bureau of the Census, a graph containing nodes (corresponding to fields) which are connected by arcs (an arc between two nodes indicates that the two fields are involved in an edit failure.) Deleting a node is equivalent to choosing that

field to be imputed. A minimal set of deleted nodes is equivalent to a minimal set as defined by Fellegi and Holt.

Hot-deck imputation - A method of imputation whereby values of variables for good records in the current (hot) survey file are used to impute for blank values of incomplete records.

Implied edit - A unstated edit derived logically from explicit edits that were written by a subject matter specialist.

Imputation - A procedure for entering a value for a specific data item where the response is missing or unusable.

Integrated Survey Processing - The concept that all parts of the survey process be integrated in a coherent manner, the results of one part of the process automatically giving information, to the next part of the process. The Blaise system is an example of integrated software in which the specification of the Blaise Questionnaire gives rise to a data entry module as well as CATI and CAPI instruments. The goals of Integrated Survey Processing include the one-time

specification of the data, which in turn would reduce duplication of effort and reduce the numbers of errors introduced into the system due to multiple specifications.

Linear edits - Edits arising from linear constraints. For example:

a. a<= F <=b.

b. a+b=c+d.

Local area network (LAN) - A group of microcomputers hooked together and which share memory and processing resources. Important to editing in that a LAN may be able to handle some editing tasks that might overwhelm one microcomputer while at the same time avoiding expensive processing on a mainframe.

Macro-edit - Detection of individual errors by: 1) checks on aggregated data, or 2) checks applied to the whole body of records. The checks are based on the impact on the estimates, (Granquist, 1987)

Micro-edit - editing done at the record, or questionnaire level.

Minimal set of fields to impute - The smallest set of fields requiring imputation that will guarantee that all edits are passed. See also "Weighted minimal set".

Multivariate edit - A type of statistical edit where multivariate distributions are used to evaluate the data and to find outliers.

Nonlinear edits - Edits from nonlinear constraints. For example:

a. Ratio edits

b. Conditional edits

The importance of nonlinear edits is that they occur often but are not amendable to theory in the determination of a al set. Some nonlinear edits, such as ratio edits, can be cast in a linear form.

Ratio edit - An edit in which the value of a ratio of two fields lies between specified bounds. The U.S. Bureau of the Census has implemented an automated editing and imputation system in the special case where all edits are ratio checks.

Repeatability - The concept that survey procedures should be repeatable from survey to survey and from location to location; the same data processed twice should yield the same results. (Also called reproducibility.)

Specifications generator - A module in an editing system from which files for paper questionnaires, data entry modules, editing software, CATI, CAPI, and summary software are generated. The specifications generator is the unifying feature in Integrated Survey Processing software. In the Blaise system, the Blaise Questionnaire can be considered to be a specifications generator. The specifications generator contains information relating to the data to be collected as well as to the edits and routes to applied to the data.

Statistical edit - A set of checks based on statistical analysis of respondent data, e. g., the ratio of two fields lies between limits determined by a statistical analysis of that ratio for presumed valid reporters (Greenberg and Surdi, 1984). A statistical edit may incorporate cross-record checks, e. g., the comparison of the value of an item in one record against a frequency distribution for that item for all records. A statistical edit may also use historical data on a firm by firm basis in a time series modeling procedure.

Stochastic edit - An edit which ff violated points to an error in the data with probability less than one. Example: 80 < yield < 120. Contrast with deterministic edit.

Subject-based edit - Checks incorporating real-world structures. which are neither statistical nor structural. Example: wages paid / hours worked > minimum wage.

Validation edits - Edits checks which are made between fields in a particular record. This includes the checking of every field of every record to ascertain whether it contains a valid entry and the checking

that entries are consistent with each other.

Weighted minimal set - A set in which fields are weighted according to reliability in generating amputations. All other things being equal, a choice of two or more minimal sets with the same number of elements is made by choosing the minimal set with the higher weight.

Reports Available in the Statistical Policy working Paper series

1. Report on Statistics for Allocation of Funds (Available through NTIS Document Sales, PB86-211521/AS)
2. Report on Statistical Disclosure and Disclosure-Avoidance Techniques (NTIS Document Sales, PB86-211539/AS)
3. An Error Profile: Employment as Measured by the Current Population Survey (NTIS Document Sales PB86-214269/AS)
4. Glossary of Nonsampling Error Terms: An Illustration of a

Semantic Problem in Statistics (NTIS Document Sales, PB86-211547/AS)

5. Report on Exact and Statistical Matching Techniques (NTIS Document Sales, PB86-215829/AS)

6. Report on Statistical Uses of Administrative Records (NTIS Document Sales, PB86-214285/AS)

7. An Interagency Review of Time-Series Revision Policies (NTIS Document Sales, PB86-232451/AS)

8. Statistical Interagency Agreements (NTIS Document Sales, PB86-230570/AS)

9. Contracting for Surveys (NTIS Document Sales, PB83-233148)

10. Approaches to Developing Questionnaires (NTIS Document Sales, PB84-105055/AS)

11. A Review of Industry Coding Systems (NTIS Document Sales, PB84-135276)

12. The Role of Telephone Data Collection in Federal Statistics (NTIS Document Sales, PB85-105971)

13. Federal Longitudinal Surveys (NTIS Document Sales, PB86-139730)

14. Workshop on Statistical Uses of Microcomputers in Federal Agencies (NTIS Document Sales, PB87-166393)

15. Quality in Establishment Surveys (NTIS Document Sales, PB88-232921)

16. A Comparative Study of Reporting Units in Selected Employer Data Systems (NTIS Document Sales, PB90-205238)

17. Survey Coverage (NTIS Document Sales, PB90-205246)

18. Data Editing in Federal Statistical Agencies (NTIS Document Sales, PB90-205253)

19. Computer Assisted Survey Information Collection (NTIS Document Sales, PB90-205261)

Copies of these working papers may be ordered from NTIS Document Sales, 5285 Port Royal Road, Springfield, VA 22161 (703) 487-4650