STATISTICAL POLICY
WORKING PAPER 25

Data Editing Workshop and Exposition

Prepared by the
Organizing Committee for the
Data Editing Workshop and Exposition
Federal Committee on Statistical Methodology

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

DECEMBER 1996

Members of the Federal Committee on Statistical Methodology

(March 1996)

Maria Elena Gonzalez, Chair
Office of Management and Budget

M. Denice McCormick Myers, Secretary
National Agricultural Statistics Service

Susan Ahmed
National Center for Education Statistics

Yvonne M. Bishop
Energy Information Administration

Cynthia Z. F. Clark
Bureau of the Census

Steven Cohen
Agency for Health Care Policy and Research

Lawrence H. Cox
Environmental Protection Agency

Zahava D. Doering
Smithsonian Institution

Daniel Kasprzyk
National Center for Education Statistics

Nancy Kirkendall
Energy Information Administration

Daniel Melnick
National Science Foundation

Robert P. Parker
Bureau of Economic Analysis

Charles P. Pautler, Jr.
Bureau of the Census

David A. Pierce
Federal Reserve Board

Thomas J. Plewes
Bureau of Labor Statistics

Wesley L. Schaible
Bureau of Labor Statistics

Rolf Schmitt
Bureau of Transportation Statistics

Monroe Sirken
National Center for Health Statistics

Alan R. Tupek
National Science Foundation

Denton Vaughan
Social Security Administration

Robert Warren
Immigration and Naturalization Service

G. David Williamson
Centers for Disease Control and Prevention

Foreword

This volume, No. 25 in the Federal Committee on Statistical Methodology (FCSM)'s Working Paper series, is the written record of the Data Editing Workshop and Exposition, held on March 22, 1996, at the Bureau of Labor Statistics (BLS) Conference and Training Center. This conference was over a year in planning, by an Organizing Committee that was an outgrowth of the FCSM's Subcommittee on Data Editing. From an initial plan of ten or 20 papers and computer software demonstrations, and perhaps 100 attendees, the registrations and submissions kept growing until the final program consisted of 44 oral presentations and 19 software demonstrations on data editing, with over 500 conference attendees. This success is probably due to several causes, not the least of which were the many outstanding contributions by the authors whose work appears in this volume. Perhaps the high participation level also suggests that data editing has been an overlooked area in the work of Federal, state, and international statistical agencies, as well as private-sector organizations.

From the start it was our intention to plan and produce this conference on as close to a zero budget as possible. Our holding this goal seemed to foster an atmosphere of cooperation in which contributions and offers of assistance came forth from numerous sources and at the times they were most needed. From the early publicity provided by several agencies and collaborating organizations to the preparation by IRS of the works published in this volume, donations of time and effort were most generous. The BLS staff was truly outstanding in anticipating and handling the many physical arrangements for the Workshop. And the agencies listed as affiliations of the Organizing Committee members all contributed varying degrees of staff time towards ensuring the success of this conference.

A Dedication in Memory of Maria Elena Gonzalez

We began the planning of the Data Editing Workshop and Exposition under the guidance of the FCSM and its founding chairperson, Maria Elena Gonzalez. After an illness, Maria passed away earlier this year and, while the FCSM continued its sponsorship of the conference and these *Proceedings*, Maria Gonzalez' departure is a deep personal and professional loss to all of us. Her career as a Federal government statistician spanned a quarter century, during which she made many contributions to improving the quality of Federal (and international) statistics. She did this both directly and as an outstanding leader in bringing forth and leveraging the talents of others for the many valuable statistical projects and conferences that she initiated. The editors would like to dedicate this *Proceedings* volume to Maria Gonzalez' memory, as was done by the Organizing Committee for the conference itself.

The next few pages contain the table of contents, followed by the conference contributions themselves. The conference program, including the list of sponsors and additional acknowledgments, is reproduced in an appendix.

David Pierce and Mark Pierzchala, Chairs
Organizing Committee for the Data
Editing Workshop and Exposition

Wendy Alvey and Bettye Jamerson, Editors
Proceedings of the Data Editing Workshop and Exposition

DECEMBER 1996

Data Editing Workshop and Exposition: Organizing Committee

David A. Pierce, Chair
Federal Reserve Board

Mark Pierzchala, Co-Chair
National Agricultural Statistics Service

Yahia Ahmed
U. S. Internal Revenue Service

Frances Chevarley
National Center for Health Statistics

Charles Day
National Agricultural Statistics Service

Rich Esposito
U. S. Bureau of Labor Statistics

Sylvia Kay Fisher
U. S. Bureau of Labor Statistics

Laura Bauer Gillis
Federal Reserve Board

Maria Elena Gonzalez
U. S. Office of Management and Budget

Robert Groves
Joint Program in Survey Methodology

Ken Harris
National Center for Health Statistics

David McDonell
National Agricultural Statistics Service

Renee Miller
U. S. Energy Information Administration

M. Denice McCormick Myers
National Agricultural Statistics Service

Jeff Owings
National Center for Education Statistics

Thomas B. Petska
U. S. Internal Revenue Service

Linda Stinson
U. S. Bureau of Labor Statistics

Paula Weir
U. S. Energy Information Administration

William E. Winkler
U. S. Bureau of the Census

TABLE OF CONTENTS

Page

Chapter 1

Overviews

Chair: Fred Vogel, National Agricultural Statistics Service

Linda M. Ball

Leopold Granquist

Kenneth W. Harris

A Paradigm Shift for Data Editing

Linda M. Ball, U.S. Bureau of the Census

Chapter 1

Abstract

Viewed through the current paradigm, the survey process consists of collecting, editing, and summarizing survey data. We think of survey data as the "stuff" that interviewers collect, the basic units of which are individual questionnaire items. On this view, pieces of data are either erroneous or not erroneous, you can correct erroneous data, and data editing is a manageable process for most surveys.

The author proposes that we instead view the survey process as engineering and managing socio-economic information systems. In this paradigm, a survey is an expression of a mental model about society. The basic units that make up the mental model are objects or concepts in the real world about which we wish to collect information. Our mental models fail to capture fully the complexity of those objects and concepts, and a questionnaire fails to capture fully the complexity of our mental model. It is no surprise, then, that surveys yield unexpected results, which may or may not be erroneous.

When an edit detects an "error," it often can't tell whether that "error" was simply an unexpected result or one of the host of errors in administering the questionnaire and in data processing that occur regularly in the administration of surveys. If we write "brute force edits" that ensure many errors are corrected, we may miss getting feedback on the problems with the mental model underlying the survey. If we take a more hands off approach, users complain that the data set has errors and is difficult to summarize and analyze. Is it, then,any surprise that we are usually not satisfied with the results we get from edits?

Abstract (cont'd)

We get a glimpse of the true complexity of the subject matter of a survey when we study the edits of a survey that has been around for a long time.

The longer a survey has been around, the more its edits evolve to reflect the complexity of the real world. For the same reason, questionnaires tend to become more complex over time. CATI/CAPI allowed us to climb to a new level of possible questionnaire complexity, and we immediately took advantage of it because we always knew that a paper questionnaire could not be designed to handle the complexity of the subject matter of most surveys.

One way to address unexpected results is to prepare some edits in advance and use an interactive data analysis and editing process after data collection to examine unexpected results. But there is a limit to the desirability of this because of the volume of labor intensive analysis that must be done, which interferes with the timeliness of data delivery that is so valuable to many data users and increases costs.

A better alternative may be to identify or develop a methodology for approximating the mental model that underlies the survey using information engineering techniques.

Information engineering is a family of modeling techniques specifically developed for information systems. First, one approximates the mental model using information engineering techniques. Then, he or she documents the linkage between the information model and the questionnaire. Everyone who works on or sponsors the survey helps to document the model and can propose changes to it.

The information system model improves considerably over the questionnaire and procedural edits. It provides a language for representing information and relationships (for example, entity relationship diagrams or object models), allows better economy of expression, is more stable over time, is more manageable and maintainable, serves as survey documentation for data users, and serves as a basis for database design. Data relationships would replace the data edits of the current paradigm.

By adopting an information engineering paradigm, we have at our disposal many well-established, tried and tested methods for managing what we usually call survey data (what the author would call socio-economic information systems). We can take advantage of existing training, professional expertise, and software, and we can integrate the practice of survey statistics with other information technologies.

A Paradigm Shift for Data Editing

Linda M. Ball, U.S. Bureau of the Census

Introduction

A paradigm is "a set of all inflected forms based on a single stem or theme" according to the Random House Dictionary. This paper proposes a new paradigm for data editing based on the central theme:

data edit = data relationship.

Examples are provided that illustrate how the information that is normally described in terms of "IF/ THEN/ELSE" procedural logic, can also be represented in the form of a logical data model (an entity-relationship diagram). (See Allen, C. Paul, 1991, for a definition..)

The implications that this paper describes as following from this central theme are the opinion of the author, and the reader is encouraged to come to her or his own conclusions about the implications. Although the implications may be a matter of opinion, the basic premise that the data relationships can be derived from current procedural edits and can be expressed as a logical data model in the form of an entity-relationship diagram is demonstrated in this paper.

For each example the following pieces of information are provided:

❒ Current Paradigm

- **List of Data Items**: including a short variable name for the item, a longer more descriptive variable name, a textual description of the data item, the actual questionnaire text of the question it represents (if applicable), and the possible values the data item can have.

- **Flowchart or Pseudocode**: depicting procedural edit logic.

❒ New Paradigm

- **Entity-Relationship Diagram**: depicting a logical information structure that is more informative than in the current paradigm.

Example 1

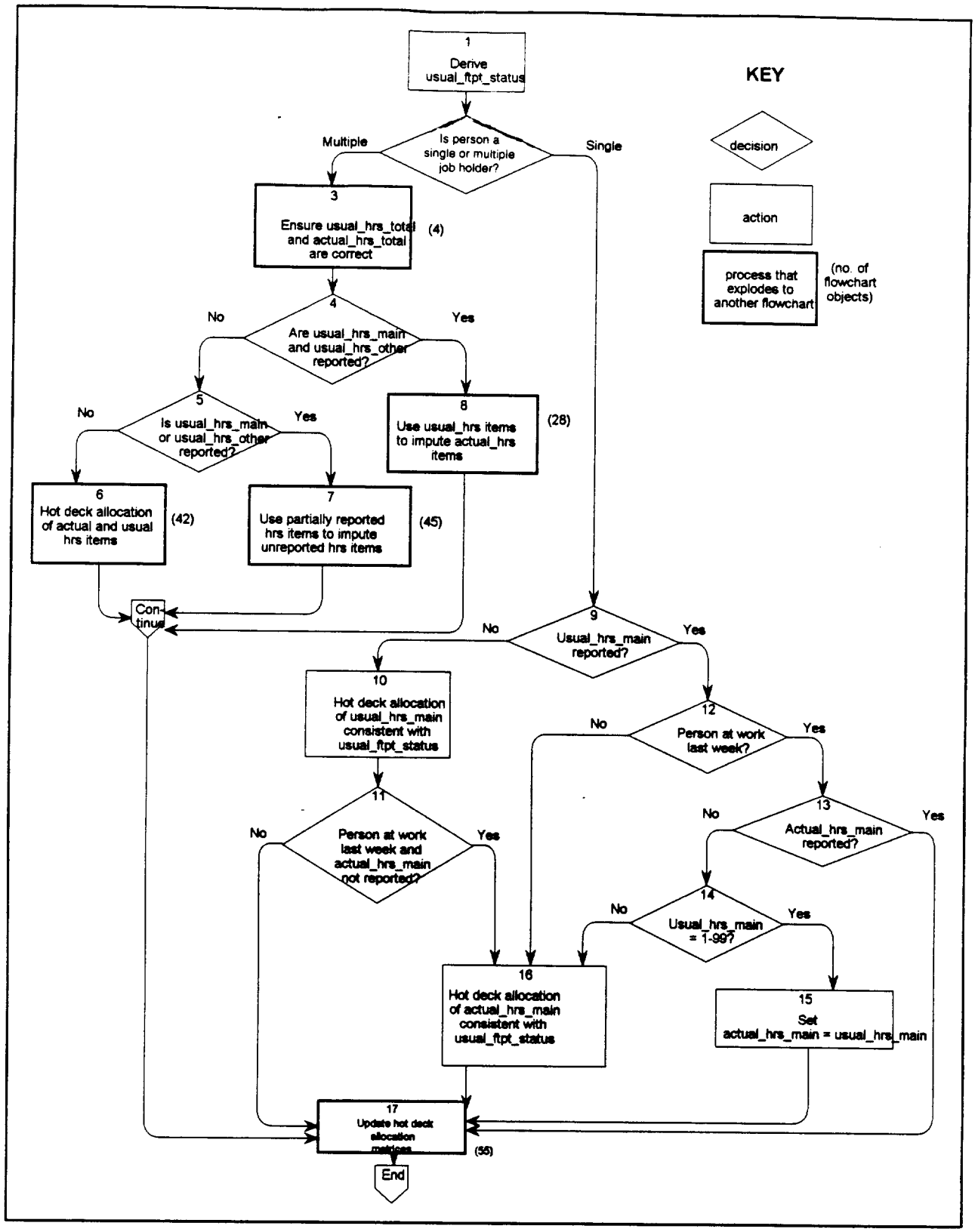
The first example is taken from the labor force section o f the *Current Population Survey*. The data items from the survey that are used in this example are shown in Table 1.

Under the current paradigm the data structure has minimal complexity and the edits have a high degree of complexity. The edits of the hours-worked items from the labor force section of the CPS questionnaire, as illustrated in Table 1 and Figure 1, take over 300 lines of pseudocode, which means some multiple of that number in FORTRAN code. This is a significantly large set of logic to document and maintain.

Table 1.--Current Paradigm: List of CPS Labor Force Data Items

Data Item		Description	Question Text (if applicable)	Values
Short Name	Long Name			
QSTNUM	QSTNUM	Unique identifier for a a questionnaire	Not applicable	1-*n* where n = the number of questionnaires in the survey
OCCURNUM	OCCURNUM	Unique identifier for a person about which the interviewer collects information	Not applicable	1-*n* where n = the number of persons interviewed at a particular address (usually 16 or less)
MJNUM	number_of_jobs	Number of jobs held last week	Altogether, how many jobs did you have?	2=2 jobs 3=3 jobs 4=4 or more jobs
HRUSL1	usual_hrs_main	Usual hours per week at main job	How many hours per week do you USUALLY work at your [main job? By main job we mean the one at which you usually work the most most hours./job?]	0-99=Number of hours v=Hours vary
HRUSL2	usual_hrs_other	Usual hours at other jobs	How many hours per week do you USUALLY work at your other (jobs/job)?	0-99=number of hours v=Hours
HRUSLT	usual_hrs_total	Sum of HRUSL1 and HRUSL2. If only one of them has a value, that value is stored in HRUSLT.	Not applicable	0-198=Number of hours v=Hours vary
HRACT1	actual_hrs_main	Actual hours at main job last week did	(So for ?) LAST WEEK, how many hours did you ACTUALLY work at your (MAIN/) job?	0-99=Number of hours
HRACT2	actual_hrs_other	Actual hours at other jobs last week	LAST WEEK, how many hours did you ACTUALLY work at your other (jobs/job)?	0-99=Number of hours
HRACTT	actual_hrs_total	Sum of HRACT1 and HRACT2. If only one of these has a value, that value is stored in HRACTT.	Not applicable	0-198=Number of hours
USFTPT	usual_ftpt_status	Usual full-time/part-time status. (derived)	Not applicable	1=Usually full time 2=Usually part time 3=Status unknown
ABSRSN	reason	Reason for absence from work last week.	What was the main reason you were absent from work LAST WEEK?	1=On layoff 2=Slackwork/business conditions 3=Waiting for a new job to begin 4=Vacation/personal days 5=Own illness/injury/medical problems 6=Child care problems 7=Other/family/personal obligation 8=Maternity/paternity leave 9=Labor dispute 10=Weather affected job 11=School/training 12=Civic/military duty 13=Does not work in the business 14=Other (specify)

Figure 1.--Current Paradigm: Flowchart Depicting Procedural Edit Logic



In contrast, the diagram in Figure 2, along with supporting information in Table 2, conveys most, if not all, of the information that is represented by the flowchart in Figure 1, but in a more concise manner. From the entity-relationship diagram in Figure 2, one can get the following information about CPS data items:

- An EMPLOYED ADULT is uniquely identified by their qstnum and occurnum.
- Number_of_jobs is an attribute of EMPLOYED ADULT.
- A single EMPLOYED ADULT has one and only one MAIN EMPLOYMENT and a single instance of MAIN EMPLOYMENT is had by one and only employed adult.
- An EMPLOYED ADULT has zero or one OTHER EMPLOYMENT and a single occurrence of OTHER EMPLOYMENT is had by one and only one EMPLOYED ADULT.
- A signle instance of MAIN EMPLOYMENT is uniquely identified by qstnum and occurnum.
- Usual_hrs_main, usual_hrs_total, and usual_ftpt_status are attributes of MAIN EMPLOYMENT.
- AT WORK LAST WEEK is a subtype of MAIN EMPLOYMENT.
- AT WORK LAST WEEK [main employment] is uniquely identified by qstnum and occurnum.
- Actual_hrs_main and actual_hrs_total are attributes of AT WORK LAST WEEK [main employment].
- ABSENT LAST WEEK is a subtype of MAIN EMPLOYMENT.
- ABSENT LAST WEEK [main employment] is uniquely identified by qstnum and occurnum.
- Reason is an attribute of ABSENT LAST WEEK [main employment].

Example 2

Example 2 is taken from the Demographics section of the Current Population Survey Questionnaire. Because of the complexity and length of the CPS demographic edit only an excerpt is shown below in Figure 3. The excerpt shown performs only one of many functions within the complete edit, but the example provides a feel for what kind of logic is necessary to edit the data under the current paradigm. What makes this section of the survey worth including as an example is the many relationships that exist among data items. They are more complex than those that inherently exist in the CPS Labor Force data shown in EXAMPLE 1.

Figure 2.--New Paradigm: Entity-Relationship Diagram of CPS Labor Force Information

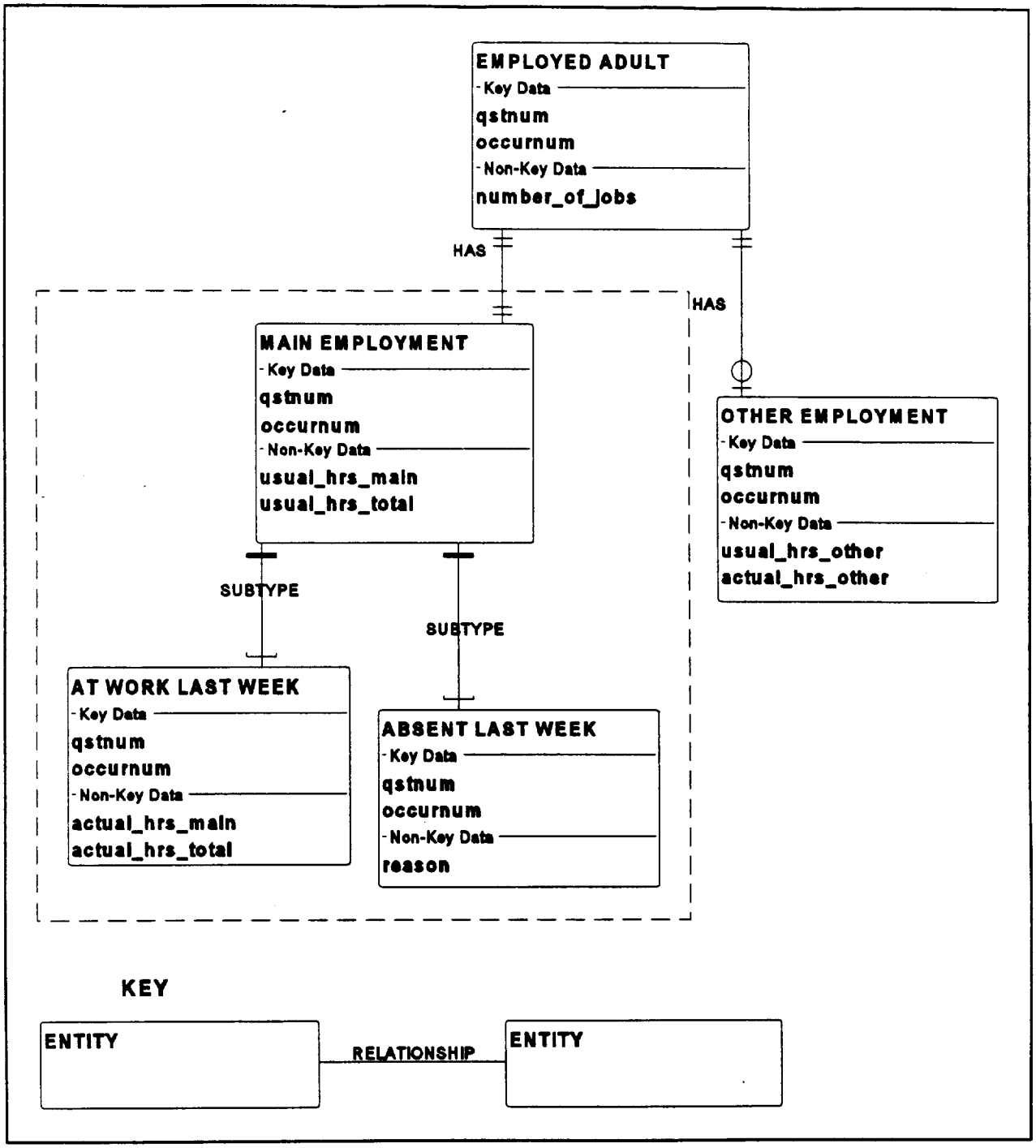


Table 2.--Current Paradigm: List of CPS Demographic Data Items

Data Item	Description	Question Text (if applicable)	Values
QSTNUM	Unique identifier for a questionnaire	Not applicable	1-*n* where n = the number of questionnaires in the survey.
OCCURNUM	Unique identifier for a person about which the interviewer collects information	Not applicable	1-*n* where n = the number of persons interviewed at a particular address (usually 16 or less)
AGE	Derived from date of birth	Not applicable	1-99
RRP	Relationship to Reference Person: Relationship to the first household member mentioned by the respondent, who is the owner or renter of the sample unit	How are you related to (reference person)?	1=Reference Person With Other Relatives in Household 2=Reference Person With No Other Relatives in Household 3=Spouse 4=Child 5=Grandchild 6=Parent 7=Brother/Sister 8=Other Relative 9=Foster Child 10=Nonrelative of Reference Person With Own Relatives in Household 11=Partner/Roommate 12=Nonrelative of Reference Person with No Own Relatives in Household 13=Nonrelative of Reference Person-Unknown Own Relatives
SPOUSE	Spouse Line Number: Line number of the person's spouse for household members whose spouse is a household member	Enter line number of spouse of [fill name] ASK IF NECESSARY	1-99=Line number 0=No one in household --
PARENT	Parent Line Number: Line number of the person's parent for household members whose parent is a household member	Enter line number of parent of [fill name] -- ASK IF NECESSARY	1-99=Line number 0=No one in household
MARITL	Marital Status	Are you now married, widowed, divorced, separated or never married?	1=Married, spouse present 2=Married, spouse absent 3=Widowed 4=Divorced 5=Separated 6=Never married
FAMNUM	Family Number: Each family unit within the household is assigned a sequential number	Not applicable	1-99
FAMREL	Family Relationship: Each family unit within the household has a reference person. Others in the family unit are assigned a code indicating their relationship to the family reference person	Not applicable	0=Not a family member 1=Reference person 2=Spouse 3=Child 4=Other relative

Figure 3.--Current Paradigm: Edit Pseudocode Excerpt

Description: If the reference person is married with their spouse present in the household, then this should be reflected consistently in the following items: RRP (Relationship to reference person); SPOUSE (Line number of spouse); MARITL (Marital status)

Pseudocode:

Do for each household:
If (person with RRP = "reference person with relatives") has SPOUSE > 0
Then Do:

- If ((person with SPOUSE = LINENO of (person with RRP = "reference person with relatives")) has RRP = "spouse of reference person"

Then:

- Set SPOUSE of (person with RRP = "spouse of reference person") = LINENO of (person with RRP = "reference person with relatives")

Else:
If SPOUSE of (person with LINENO = SPOUSE of (person with RRP = "reference person with relatives")) then

- Set RRP of (person with LINENO = SPOUSE of (person with RRP = "reference person with relatives) = "spouse of reference person"

Else:

- Set SPOUSE of (person with RRP = "reference person with relatives") = blank
- If MARITL of (person with RRP = "reference person with relatives) = "married, spouse present"

Then:

- allocate a value for MARITL that is one of the "unmarried" categories.

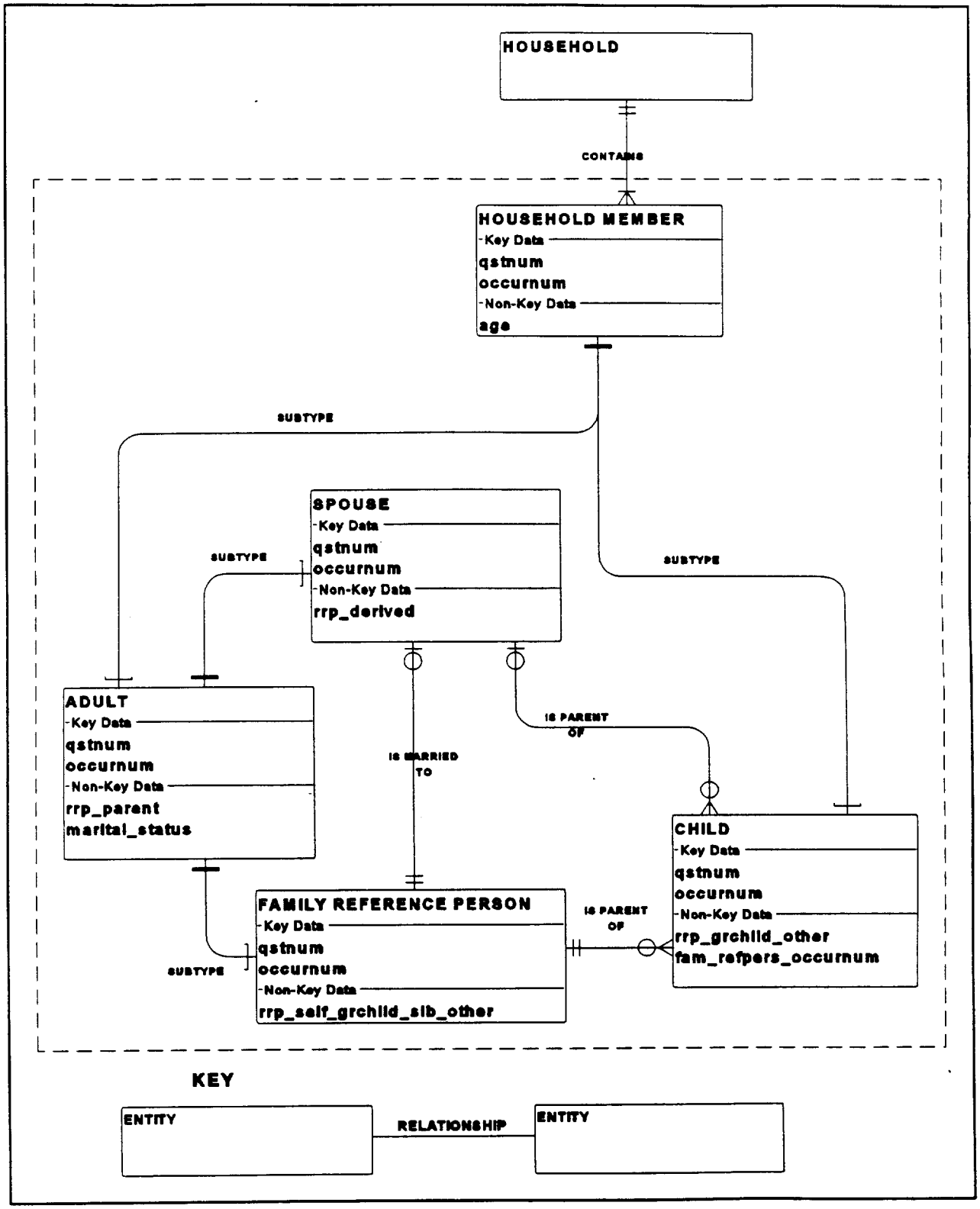
Endif

- Set MARITL of (person with RRP = "spouse of reference person") = "married, spouse present"
- Set MARITL of (person with RRP = "reference person with relatives") = "married, spouse present"

Again, as in EXAMPLE 1, under the current paradigm the data structure has minimal complexity and the edits have a high degree of complexity, but under the new paradigm, the logical data structure is more complex and informative. The following information is expressed in the entitity-relationship diagram in Figure 4:

- A HOUSEHOLD is uniquely identified by qstnum.
- A HOUSEHOLD contains one or many HOUSEHOLD MEMBERS, and a HOUSEHOLD MEMBER is contained by one and only one HOUSEHOLD.
- A HOUSEHOLD MEMBER is uniquely identified by qstnum and occurnum.
- Age is an attribute of HOUSEHOLD MEMBER.
- ADULT and CHILD are subtypes of HOUSEHOLD MEMBER.
- SPOUSE and FAMILY REFERENCE PERSON are subtypes of ADULT.
- ADULT, CHILD, SPOUSE, AND FAMILY REFERENCE PERSON are each uniquely identified by qstnum and occurnum.
- Rrp_parent and marital_status are attributes of ADULT.
- Rrp_grchild/other is an attribute of CHILD.
- Rrp_derived is an attribute of SPOUSE.
- Rrp_self/grchild/sib/other is an attribute of FAMILY REFERENCE PERSON.
- A SPOUSE is married to one and only one FAMILY REFERENCE PERSON, which is uniquely identified by qstnum, occurnum, and fam_refpers_occurnum where SPOUSE.fam_refpers_occurnum = FAMILY REFERENCE PERSON.occurnum; and a FAMILY REFERERNCE PERSON is married to zero or one SPOUSE.

Figure 4.--New Paradigm: Entity-Relationship Diagram of CPS Demographic



- A CHILD is the child of one and only one FAMILY REFERENCE PERSON which is uniquely identified by qstnum, occurnum, and fam_refpers_occurnum where CHILD.fam_refpers_occurnum = FAMILY REFERENCE PERSON.occurnum; and a FAMILY REFERENCE PERSON is parent of zero, one, or many CHILDREN.
- A CHILD is the child of zero or one SPOUSE which is uniquely identified by qstnum, occurnum, and parent2s_occurnum where CHILD.parent2s_occurnum = SPOUSE.occurnum; and a SPOUSE is the parent of zero, one, or many CHILDREN.

Problems With the Current Paradigm

If the central theme of the new paradigm is expressed as data edit = data relationship, then the central theme of the current paradigm would have to be expressed as:

data edit = logical process for changing a data item.

From an operational perspective, there are two distinct types of edits.

Type 1 allocates or imputes missing data and outliers and inconsistencies that the edit authors know about prior to data collection because of their knowledge of what needed to be edited in previous survey iterations, i.e. previous iterations of either the survey in question or another similar survey.

Type 2 allocates or imputes new values for unexpected results. These are outliers or inconsistencies that the edit authors do not know about until they or someone else examines the edited data from the survey. The real world socio-economic concepts about which we collect information, are more complex than the assumptions conveyed by a questionnaire (especially in the paper questionnaire environment). It is no surprise then that we get unexpected results. For example, until recently CPS did not allow same-sex married couples in its data. Our assumptions told us that if we found this in the dataset it was probably a data collection or keying error. The edit checked for this and edited the data to disallow it. Recently it was decided to allow same-sex married couples in CPS data. Our mental model changed based on our information about society, and this change was reflected eventually in the CPS questionnaire and edits. A difficulty with post-data-collection edits is that we don't always know whether we are changing a true outlier or correcting an error that occurred in any of the survey's operational processes leading up to the edits.

It is the Type 2 edits that consume the most resources during the time when survey operations staff are trying to meet deadlines for delivery of data to the survey sponsor. These unexpected results can originate from any point in any of the processes from questionnaire design through reformatting of data.

If we write edits that ensure all outliers are eliminated, we may miss getting feedback on the problems with the survey design or various operational procedures. If we take a more hands off approach, users complain that the data set has errors and is difficult to summarize and analyze.

The current paradigm leads to excess complexity in the edit process. To get a glimpse of the true complexity of the subject matter of a survey, one should study the edits of a survey that has been in operation for a long time. Because the longer a survey has been in operation, the more of the true complexity of the real world has been incorporated into the edits. For the same reason, questionnaires tend to become more complex over time. CATI/CAPI has allowed us to climb to a new level of possible questionnaire complexity, and we immediately took advantage of it because we always knew that a paper questionnaire could not be designed to handle the true complexity of the subject matter of most surveys.

To illustrate this tendency toward complexity, compare the CPS paper questionnaire that was in use prior to 1994 to the electronic version used since 1994. The paper questionnaire plus control card filled about 15-20 pages of condensed print. The electronic version fills hundreds of pages and has many more logical paths than the paper questionnaire.

Data-edits-as-procedures work well when they are attached to a data collection process that has been repeated a number of times without change. However, if a survey is annual or every 5 years and the survey design changes significantly each time, software maintenance side effects dominate and the edits can be unmanageable, or at least very expensive to manage (Pressman, 1992).

In other words there are two competing forces acting upon the manageability of edits. One tends to increase manageability over time, the other tends to decrease manageability over time. In the long run, even in surveys that don't change for a while, software maintenance side effects eventually take over because in the long run changes to a survey design always become necessary. These side effects are:

- Edits are often complex and difficult to understand and document.
- Any change to the questionnaire causes changes in the edits because of the very high degree of dependence between them.
- Changes to the edits are often poorly documented because they are developed in a hurry after data collection has ended.
- The accumulation of additions to the original edit design over time causes added complexity in the logic.
- People who understand the edits leave and newly hired people have a long learning curve and poor documentation to follow.
- Changes to one part of the edits may cause another part to work incorrectly.

Under the current paradigm, there is a tradeoff between usability of data and timeliness of data. If you accept the premise that a survey will yield unexpected results every time you implement a new or revised questionnaire or operational procedure, then you must conclude that data must be inspected and adjusted after data collection in order to provide the data usability that sponsors and end users require. Therefore you might conclude that you can have some edits prepared in advance but need an interactive data analysis and editing process after data collection to deal with unexpected results. Currently, there is a great desire for increased timeliness, but there is also a minimum standard for usability of data files that cannot be sacrificed.

A New Paradigm: Implications of the Central Theme

- ***Decreased Time Between Data Collection and Data Delivery***.--This could be achieved by

  - capturing data relationships as data is entered during the interview, or immediately after the interview, while the interviewer still has access to it; and

  - giving other participants in the survey process, no matter where they are physically located, immediate access (with appropriate confidentiality constraints) to that data and stored data relationships (Hammer and Champy, 1993) .

- ***Increased Maintainability***.--Separate technology maintenance from data administration. It is common wisdom within the software engineering field that the information in a system tends to be more stable over time than the processes in a system. Processes tend to be more technology dependent than information. Shifting some of the complexity of a survey from its component processes to its component information structures should make the operation as a whole inherently more maintainable.

- ***More Accurate and Up-to-date Documentation for Users***.--A logical data model, once developed could be used not only by the survey developers, but also by data users.

Difficulties

In addition to the beneficial implications listed above, it must be stated that there would most likely be difficulties in implementing a survey based on this new paradigm. Firstly, it would most certainly take more lead development time the first time it is tried for any given survey. Survey content experts would have to come together on a logical model of the data collected by the survey.

Secondly, there may be organizational problems involving the role of interviewers, the technology skills of survey staff, and the necessity of many organizational units coming together on a strategy for survey operations that emphasizes the joint development and sharing of complex datasets.

References

Allen, C. Paul, (1991). *Effective Structured Techniques from Strategy to Case*, New York, NY: Prentice Hall.

Hammer, M. and Champy, James (1993). Chapter 5 of *Reengineering the Corporation: The Enabling Role of Information Technology*, New York, NY: Harper Collins.

Pressman, Roger S. (1992) *Software Engineering: A Practitioner's Approach*, New York, NY: McGraw-Hill, 1992

Bibliography

Bureau of the Census, Demographic Surveys Division, *Edit Specifications for the 1994 Current Population Survey CATI/CAPI Overlap Test.*

Bureau of the Census, Demographic Surveys Division, *Data Documentation for the 1994 Current Population Survey CATI/CAPI Overlap Test.* ■

Chapter 1

The New View on Editing

Leopold Granquist, Statistics Sweden

Abstract

An international new view on editing has grown during the last five-ten years out of the results of research on editing carried out by eminent statistical agencies. The research shows that editing is expensive (20-40 percent of the budget cost), inefficient (the impact on quality negligible), hiding data collection problems, but that new strategies and methods may lower the cost substantially for the producer as well as for the respondent and in the long run increase the quality of data. The main point is gradually moving from cleaning up the data to identifying and collecting data on error sources, problem areas and error causes to get a basis for measures to prevent errors to arise in the data collection and processing. Thus, the editing process should produce data on the collection and processing, so called paradata, for a continuous improvement of the whole survey vehicle. This Total Quality Management (TQM) view on editing should imply lower cost and increased quality when checks are co-ordinated with the response and collection process and adapted to the respondent ability to provide data.

High quality cannot be accomplished by introducing as many and tight checks as possible and augmenting the number of follow ups with the repondents, but through careful design and testing of the set of edits, and fitting the checks currently to the data to be scrutinised. New types of edits and strategies should be used to focus the editing to those serious errors, which can be identified by editing. An important feature is to classify edits into critical and query edits. The critical edits shall be used to detect and remove fatal errors, that is those errors which the editing process has to remove from data. The query edits should be concentrated on those suspicious data, which when containing errors, may have a substantial impact on the estimates. The re-contacts to respondents have to be limited as much as possible, but when considered necessary, the contact should be used not only to find better data but to get intelligence of causes of errors, error sources and respondent problems of providing accurate data. The new technology with more and more powerful personal computers plays an important role for using new more efficient editing methods, as graphical editing; new strategies, as data entry editing and moving the editing closer to the data source in CAI and CASI modes of data collection. It is stressed that it is not an issue of translating old methods to a new technology, but to re-engineer the whole editing process under a new view on editing.

The New View on Editing

Leopold Granquist, Statistics Sweden

Introduction

It may be claimed that a new international common view on editing has grown the last five - ten years and become established through papers presented at the ISI conferences in Cairo 1991 (Linacre, 1991) and Florence 1993 (Lepp and Linacre, 1993; Silva and Bianchini, 1993; among others), the International Conference on Establishment Surveys at Buffalo 1993 (Granquist, 1995 and Pierzchala, 1995), the International Conference on Survey Measurement and Process Quality at Bristol 1995 (Granquist and Kovar, 1996), and at the annual Work Sessions on Statistical Data Editing organized by United Nations Statistical Commission and Economic Commission for Europe (ECE, 1994 and ECE, 1996).

The emphasis of the editing task is moving from just cleaning up the data, though still a necessary operation, to identifying and collecting data on errors, problem areas, and error causes to provide a basis for a continuous improvement of the whole survey vehicle. This Total Quality Management (TQM) view on editing implies lower cost for the producer, less respondent burden, and increased quality as checks are integrated and co-ordinated with the response and collection process and adapted to the respondent ability to provide accurate data.

This change in the editing process has been embraced by some statistical agencies when it was recognized that the heavy cost of editing cannot be justified by quality improvements as measured by numerous evaluation and other studies on editing. Some facts:

- Editing accounts for a substantial part of the total survey budgets. The monetary costs (hardware, software, salary and field costs) amount to 20-40 percent of the survey budget (Granquist,1984; Federal Committee on Statistical Methodology, 1990; and Gagnon et al, 1994, among others). Furthermore, there are costs related to lost opportunities, response burden and associated bad will, costs related to losses in timeliness (e.g., the machine editing of the World Fertility Survey delayed the publication of the results by about one year, Pullum et al., 1986), and indirect costs related to undue confidence in data quality and respondent reporting capacity and to using employees in inappropriate tasks (Granquist and Kovar, 1996).

- The ever ongoing rationalization of the editing process has not yet caused the process to be less costly or more efficient. The gains have been invested in attempts to raise the quality by applying more checks and to selecting more forms for manual review and follow-up (Pierzchala, 1995), resulting in only marginal improvement or more likely in overediting. There are even some examples that editing can be counter productive (Linacre and Trewin, 1989; Corby, 1984; Corby, 1986). Furthermore, some important error types cannot even be touched by editing (Pullum et al., 1986; Christianson and Tortora, 1995) or are impossible to identify by traditional error checks (Werking et al., 1988). Thus, editing is far less important for quality than survey managers believe, and high quality cannot be guaranteed by just adding more checks and augmenting the number of recontacts. On the contrary, such an approach may be counter productive and, worse of all, impose an undue confidence in the quality of the survey in the survey managers, in particular if editing is hiding problem areas instead of highlighting them.

- During the last five years new kinds of editing methods have been developed. They are termed macro-editing, selective editing or significant editing. Numerous studies and experiences of these methods indicate that the manual verifying work can be reduced by 50 percent or more without affecting the estimates (Granquist and Kovar, 1996). By removing unnecessary edits, relaxing the bounds of the remaining query edits by applying a method suggested in Hidiroglou-Berthelot (1986), and using a score function for identifying records for manual review developed by Latouche and Berthelot (1992), Engström (1995) succeeded in decreasing the manual review of a well-edited survey by 86 percent without significant consequences on the quality. The success of this research illustrates how important it is to design the set of query edits meticulously.

- The technological development has made it possible to move the editing closer to the data source, preferable while the respondent is still available. It opens the possibilities of getting more accurate data from the respondent, intelligence of what data are received and of the problems the respondent experiencies in delivering the requested data, comments concerning answers, and opportunities of conducting experiments. In general, CAI and CASI modes of data collection offer excellent possibilitties of getting data on error sources, problem areas and the reporting capacity among the respondents.

The Role of Editing

Definitions of data editing vary widely. Here editing is defined as the procedure for identifying, by means of edit rules, and for adjusting, manually or automatically, errors resulting from data collection or data processing (Granquist and Kovar, 1996).

An absolute and fundamental requirement on editing has always been, and should be, to identify outliers and those errors in individual data, which are recognizable as such to a user with access to individual data records, but without knowledge of the particular unit. Edits aimed at identifying such data , that is data which certainly are erroneous we call fatal edits and the process to ensure validity and consistency of individual data records is termed micro-editing (Granquist and Kovar, 1996). However, in surveys with quantitative data there might be errors, although not fatal, which significantly affect the estimates. Hence, query edits, that is edits pointing to suspicious data items, have to be added to the editing process.

As early as in the sixties, it was recognized that removing errors was not the most important aspect of editing, Pritzker et al. (1965) claim that it is more important to identify error sources or problem areas of the survey. Granquist (1984) agrees and says that the goals of editing should be threefold: To provide information about the quality of the data, to provide the basics for the (future) improvement of the survey, and to tidy up the data.

Editing -- A Historic Review

The low efficiency of editing processes in general is basically due to the survey managers ambition to allocate almost all resources to the third objective, cleaning up the data, especially to identifying errors by query edits. An explanation may be found in the following brief historical background.

Before the advent of computers, editing of individual forms was performed by large groups of clerks, often without any secondary school education. Though ingenious practices sometimes were developed, only simple checks could be undertaken. Editing was inconsistent not only between individual clerks but also over time for the same person. Only a small fraction of all errors could be detected. The advent of

computers was recognized by survey designers and managers as a means of reviewing all records by consistently applying even sophisticated checks requiring computational power to detect most of the errors in data that could not be found by means of manual review. The focus of both the methodological work and in particular the applications was on the possibilities of enhancing the checks and of applying automated imputation rules in order to rationalize the process, Naus (1975). Nordbotten (1963) was the theoretical basis for the checks and imputation rules used in the cumbersome main frame automated systems developed in the late sixties and the seventies. Implicitly the process was governed by the paradigm: The more checks and recontacts with the respondents the better the resulting quality. The early systems produced thousands of error messages that had to be manuall examined, by referring back to the original forms and in more complicated cases to the respondents themselves. Changes were entered in batch and edited once again by the computer. Many records passed the computer three or four times and sometimes records were reviewed by different persons each time they were flagged. Occasionally cycles of 18 could occur (Boucher, 1991).

The research and development work became focused on rationalizing the EDP departments work of providing the survey managers with application programs and on developing generalized software. A methodology for generalized editing and imputation systems was developed by Fellegi and Holt (1976), implemented for editing of categorical data (e.g., CAN-EDIT, AERO, DIA), and for quantitative data (e.g., SPEER and GEIS) ECE (1994). These systems are well suited for removing fatal errors, defined as data that do not meet certain requirements.

The great break in rationalizing the process came as a direct consequence of the PC revolution in the eighties, when editing could be performed on-line on personal computers, at the data entry stage -- data entry heads-up -- during the interview, and by the respondent in CASI modes of data collection. Bethlehem et al. (1989) describe in detail the gains in on-line data entry editing on micros or minis as compared to main frame processing systems, and on the basis of their findings Statistics Netherlands developed the world-renowned system BLAISE. The process was substantially streamlined, but the gains were often used to allow the editors to process more records and to make more contacts with respondents to resolve encountered problems (Pierzchala,1995). The paradigm -- the more checks and contacts the better the quality -- was still considered valid. However, some evaluations carried out around 1990 said something else, that later on was corroborated in numerous evaluation and other studies on editing methods: 10 to 15 percent of the biggest changes made in the editing of economic items contribute to around 90 percent of the total change; 5 to 10 percent of the biggest changes bring the estimate within 1 percent of the final global estimate; only 20 to 30 percent of the recontacts result in changed values; the quality improvements are marginal, none or even negative; many types of serious systematic errors cannot be identified by editing (Granquist and Kovar, 1996; Werking and Clayton, 1988; Christianson and Tortora, 1995; Linacre and Trewin, 1989; among others).

Why Over-Editing Occurred

The main reason why editing processes remained inefficient is that processes seldom were evaluated, nor were indicators or other performance measures produced. Initially the approach seemed to be successful. Errors, even serious errors were detected. Ambitious survey managers and designers thus continued to see editing as an outstanding tool to achieve high-quality data, and relied upon that editing to fix any mistakes committed in earlier phases of the data collection and processing (including the survey and in particular the questionnaire design). Results from evaluations of editing processes contradicting this view were not considered applicable to their surveys. They still believed that investing in more checks and recontacts would be beneficial for the quality. But editing can be harmful to the quality, e.g., delaying

the publishing of the results; causing bias when only certain types of errors are detected, for example those which move the estimates in a specific direction; inserting errors, when the reviewer manipulate data to pass the edits, so called creative editing Granquist (1995), which -also may give a wrong impression about the reporting capacity of the respondents; introducing errors by mistakes occurring in the recontact process.

Corner Stone of the New View

We have already established, that the primary and basic requirement on editing is to identify outliers and to remove fatal errors from individual data for which traditional editing is well suited. However, it is the second type of edits, the query edits, that are responsible for the high costs of editing, and accordingly the subject of this paper.

Careful Design and Evaluation of the Set of Query Edits

The design of the entire set of query edits is of particular importance in getting an acceptable cost/benefit outcome of editing (Granquist, 1995). The edits have to be co-ordinated for related items and adapted to the data to be edited. Probable measures for improving current processes are relaxing bounds by replacing subjectively set limits by bounds based on statistics from the data to be edited, and removing edits which in general only produce unnecessary flags. It should be noted that there is a substantial dependence between edits and errors for related items. Furthermore, the edits have be targeted on the specific error types of the survey, not on possible errors.

The properties of the edits have to be controlled continuously for example by examining data of the outcome of edits. It is an absolute requirement on any editing system to produce statistics on error flags, changes related to edits and reasons for imputation, etc.

Focus on Influential Item Values and Records

The edits and/or the system should have built-in prioritizing rules to focus the review on the suspicious data items or records that have most influence on the estimates A leading principle in most of the methods suggested in Granquist (1991) is: begin with the most deviating values and stop verifying when estimates no longer are changed.

Another way of prioritizing the recontacts is to utilize score functions, as suggested by Latouche and Berthelot (1992) or Lawrence and McDavitt (1994). The idea is: to run the edits as decided by the subject matter experts, assign a score to every record with at least one flagged item according to the weight of the record, the potential impact of the suspicious item value, and the importance of the flagged item; and review only those records, which get a score exceeding a threshold value, determined in advance on basis of historical experience. When fatal edits are included, then records containing fatal errors have to be handled automatically in cases where the score of that edit does not guarantee that the threshold value is exceeded.

Numerous studies indicate that applying any of these methods will yield a reduction of the manual review work by 50 percent or more without any significant impact on quality (Granquist and Kovar, 1996). In an experimental study, Engström (1995) uses all the methods and eliminates 84 percent of the current review work.

Focus on Response Problems and Error Causes

The main aim of recontacting respondents should be to collect intelligence of respondent problems, error causes and reporting capacity. It is essential to look upstream to reduce errors in survey data, rather

than to attempt to clean up single cases at the end. It is fundamental for the data quality that respondent data are of high quality. It means that the respondent in business surveys has to: understand exactly the question and underlying definitions; have data available in his information system; understand differences in definitions between the survey and his information system, and in case of large differences be able to give an acceptable estimate. Accordingly, to provide accurate data the survey design has to be adapted to the respondent's possibilities and conditions. Editing has to highlight respondent problems and reporting capacity, not hide them. This can be accomplished by changing the focus of the recontact process from ascertaining whether a suspicious value is wrong and finding a more accurate value, to acquiring knowledge of respondent problems and causes of errors. Thus, editing can be used to advantage in sharpening survey concepts and definitions and in improving the survey vehicle design.

CASI modes of data collection offer an excellent tool for furnishing intelligence about the response process and the accuracy of delivered data, provided that: edits and error messages are designed to help the respondent in understanding what data we want from him or her; possible error causes, definitions and other information needed for answering each particular item are prompted; the respondent can easily give comments to answers or explain why he or she cannot answer the question; warnings like did you include or exclude this and that component. Furthermore, experiments can be built in to get statistics on respondent behaviour, Weeks (1992) gives a detailed description of the possibilities.

The New View on Editing

Editing should be integrated with, but subordinated to, collection, processing and estimation. A main task is to provide a basis for designing measures to prevent errors.

Editing should be considered a part of the total quality improvement process, not the whole quality process. Editing alone cannot detect all errors, and definitely not correct all mistakes committed in survey design, data collection, and processing.

The paradigm -- the more (and tighter) checks and recontacts, the better the quality -- is not valid.

The entire set of the query edits should be designed meticulously, be focused on errors influencing the estimates, and be targeted on existing error types which can be identified by edits. The effects of the edits should be continuously evaluated by analysis of performance measures and other diagnostics, which the process shoud be designed to produce.

Editing has the following roles in priority order:

- Identify and collect data on problem areas, and error causes in data collection and processing, producing the basics for the (future) improvement of the survey vehicle
- Provide information about the quality of the data
- Identify and handle concrete important errors and outliers in individual data.

Besides its basic role to eliminate fatal errors in data, editing should highlight, not conceal, serious problems in the survey vehicle. The focus should be on the cause of an error, not on the particular error per se.

References

Bethlehem, J.G.; A.J., Hundepool; M.H., Schuerhoff; and L.F.M., Vermeulen (1989). BLAISE 2.0 An Introduction, Vorburg, The Netherlands: Central Bureau of Statistics.

Boucher, L. (1991). Micro-Editing for the Annual Survey of Manufactures: What is the Value Added? *Proceedings of the Annual Research Conference*, Washington, DC: U.S. Bureau of the Census, pp. 765-781.

Corby, C. (1984). *Content Evaluation of the 1977 Economic Censuses*, Statistical Research Division Report Series No. CENSUS/SRD/RR-84-29, Washington, DC: U.S. Bureau of the Census.

Corby, C. (1986). Content Evaluation of the 1982 Economic Censuses: Petroleum Distributors, *1982 Economic Censuses and Census of Governments, Evaluation Studies*, Washington DC: U. S. Department of Commerce, pp. 27-60.

Christianson, A. and R. Tortora (1995). Issues in Surveying Business: An International Survey, in B.G. Cox; D.A. Binder; N. Chinnappa; A. Christianson; M.J. Colledge; and P.S. Kott (eds.), *Business Survey Methods*, New York: Wiley, pp. 237 - 256.

Economic Commission for Europe, (1994). Statistical Data Editing: Methods and Techniques, Volume No. 1, United Nations New York and Geneva.

Economic Commission for Europe, (1996). Statistical Data Editing: Methods and Techniques, Volume No. 2, United Nations New York and Geneva (to appear).

Engström, P. (1995). A Study on Using Selective Editing in the Swedish Survey on Wages and Employment in Industry, Room paper No. 11, presented at the Conference of European Statisticians, Work Session on Statistical Data Editing, Athens, Greece, November 6-9, 1995.

Federal Committee on Statistical Methodology (1990). *Data Editing in Federal Statistical Agencies*, Statistical Policy Office, Working Paper 18, Washington, DC: U.S. Office of Management and Budget.

Fellegi, I. P. and Holt, D. (1976). A Systematic Approach to Automatic Edit and Imputation, *Journal of the American Statistical Association*, 71, pp. 17-35.

Gagnon, F.; Gough, H.; and Yeo, D. (1994). Survey of Editing Practices in Statistics Canada, unpublished report, Ottawa: Statistics Canada.

Granquist, L. (1984). On the Role of Editing, *Statistisk Tidskrift*, 2, pp. 105-118.

Granquist, L. (1991). A Review of Some Macroediting Methods for Rationalizing the Editing Process, *Proceedings of Symposium 90: Measurement and Improvement of Data Quality*, Ottawa: Statistics Canada, pp. 225-234

Granquist, L. (1995). Improving the Traditional Editing Process, in B.G.Cox; D.A. Binder; N.Chinnappa; A. Christianson; M.J. Colledge; and P.S.Kott (eds.) *Business Survey Methods*, New York: Wiley, pp. 385-401.

Granquist, L. and Kovar, J.G. (1996). Editing of Survey Data: How Much is Enough? in L. Lyberg et al. (eds.), *Survey Measurement and Process Quality*, New York: Wiley (to appear).

Hidiroglou, M. A. and Berthelot, J.-M. (1986). Statistical Editing and Imputation for Periodic Business Surveys, *Survey Methodology*, 12, pp. 73-84.

Latouche, M. and Berthelot, J.-M. (1992). Use of a Score Function to Prioritize and Limit Recontacts in Editing Business Surveys, *Journal of Official Statistics*, 8, pp. 389-440.

Lawrence, D. and McDavitt, C. (1994). Significance Editing in the Australian Survey of Average Weekly Earnings, *Journal of Official Statistics*, 10, pp. 437-447

Lepp, H, and Linacre, S. (1993). Improving the Efficiency and Effectiveness of Editing in a Statistical Agency, *Bulletin of the International Statistical Institute: Proceedings of the 49th Session*, Florence, Italy, Contributed Papers Book 2, pp. 111-112.

Linacre, S. J. (1991). Approaches to Quality Assurance in the Australian Bureau of Statistics Business Surveys, *Bulletin of the International Statistical Institute: Proceedings of the 48th Session*, Cairo, Egypt, Book 2, pp. 297-321.

Linacre, S. J. and Trewin, D. J. (1989). Evaluation of Errors and Appropriate Resource Allocation in Economic Collections, *Proceedings of the Annual Research Conference*, Washington, DC: U.S. Bureau of the Census, pp. 197-209.

Naus, J. I. (1975). *Data Quality Control and Editing*, New York: Marcel Dekker, Inc.

Nordbotten, S. (1963). Automatic Editing of Individual Statistical Observations, Conference of European Statisticians, Statistical Standards and Studies, No. 2, New York: United Nations.

Pritzker, L, J. Ogus and Hansen, M.H. (1965). Computer Editing Methods -- Some Applications and Results, *Proceedings of the International Statistical Institute Meetings* in Belgrade Yugoslavia, 1965, pp. 442-465.

Pierzchala, M. (1995). Editing Systems and Software, in B.G. Cox; D.A. Binder; N. Chinnappa; A. Christianson; M.J. Colledge; and P.S. Kott (eds.) *Business Survey Methods*, New York: Wiley, pp. 425-441.

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, pp. 311-326.

Silva, P. L. and Bianchini , Z. (1993). Data Editing Issues and Strategies at the Brazilian Central Statistical Office, *Bulletin of the International Statistical Institute: Proceedings of the 49th Session*, Florence, Italy, Contributed Papers Book 1, pp. 377-378.

Weeks, M.F. (1992). Computer-Assisted Survey Information Collection: A Review of CASIC Methods and Their Implications for Survey Operations, *Journal of Official Statistics*, 8, pp. 445-465.

Werking, G.; Tupek, A.; and Clayton, R. (1988). CATI and Touchtone Self-Response Applications for Establishment Surveys, *Journal of Official Statistics*, 4, pp. 349-362. ■

Chapter 1

Data Editing at the National Center for Health Statistics

Kenneth W. Harris, National Center for Health Statistics

Abstract

Data editing can be defined as the procedures designed and used for detecting erroneous and/or questionable survey data with the goal of correcting as much of the erroneous data as possible, usually prior to data imputation and summary procedures. This paper describes many of the data editing procedures used for selected data systems at the National Center for Health Statistics (NCHS), the Federal agency responsible for the collection and dissemination of the nation's vital and health statistics.

Data Editing at the National Center for Health Statistics

Kenneth W. Harris, National Center for Health Statistics

Background

The National Center for Health Statistics (NCHS) is the Federal agency responsible for the collection and dissemination of the nation's vital and health statistics. To carry out its mission, NCHS conducts a wide range of annual, periodic, and longitudinal sample surveys and administers the national vital statistics registration systems. These sample surveys and registration systems form four families of data systems: vital event registration systems, population based surveys, provider based surveys, and followup/followback surveys.

Much of what happens to the data covered by these data systems, from collection through publication, depends on the family to which they belong. At most steps along the way, various activities and operations are implemented with the goal of making the data as accurate as possible. These activities and operations are generally categorized under the rubric, "data editing." In the 1990 Statistical Policy Working Paper 18: *Data Editing in Federal Statistical Agencies*, [1] data editing is defined as:

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. [However, this report includes data imputation procedures.]

As will be shown in this report, data editing procedures vary greatly between NCHS data systems.

Twenty-four data systems are included in this report (see Table 1). For each data system, summary descriptions of NCHS data editing practices are provided in the following 11 areas:

Environment in Which Survey Takes Place
Data Processing Environment and Dispersion of the Work
Audit Trail
Micro-, Macro-, and Statistical Editing
Prioritizing of Edits
Imputation Procedures
Editing and Imputation Standards
Costs of Editing
Role of Subject Matter Specialists
Measures of Variation
Current and Future Research.

Table 1.--NCHS Data Systems Included in this Report

Registration Systems (8)

Mortality (MRS)
Fetal Mortality (FMRS)
Abortion (ARS)
Natality (NRS)
Marriage (MRG)
Divorce (DRS)
Current Mortality Sample (CMS)
Linked Birth and Infant Death Data Set (LBIDDS)

Population Based Surveys (3)

National Health Interview Survey (NHIS)
National Health and Nutrition Examination Survey (NHANES)
National Survey of Family Growth (NSFG)

Provider Based Surveys (7)

National Hospital Discharge Survey (NHDS)
National Survey of Ambulatory Surgery (NSAS)
National Ambulatory Medical Care Survey (NAMCS)
National Hospital Ambulatory Medical Care Survey (NHAMCS)
National Nursing Home Survey (NNHS)
National Home and Hospice Care Survey (NHHCS)
National Health Provider Inventory (NHPI)

Followup/Followback Surveys (6)

National Maternal and Infant Health Survey (NMIHS)
1991 Longitudinal Followup (LF) to the NMIHS
National Mortality Followback Survey (NMFS)
National Health and Nutrition Examination Survey (Cycle I) Epidemiologic Followup Study (NHEFS)
Longitudinal Study of Aging (LSOA)
National Nursing Home Survey Followup (NNHSF)

Within each of these areas, data editing practices are grouped according to the type of data system, i.e., vital event registration systems, population based surveys, provider based surveys, and followup/ followback surveys.

Environment in Which Survey Takes Place

Registration Systems

The vital event registration systems cover six vital events: mortality, fetal mortality, induced termination of pregnancy (abortion), natality, marriage and divorce. For each of these systems, data are obtained from certificates and reports filed in state registration offices and registration offices of selected cities and other areas. Coverage for each registration system is limited to its prescribed registration area (RA). The oldest registration areas, mortality, fetal mortality, and natality, have been complete since 1933. These three are national data systems; i.e., they cover the entire United States. The marriage RA started in 1957 with 30 states and reached its current coverage of 42 states plus selected areas in 1986. The Divorce RA started in 1958 with 14 states and by 1986 had expanded to 31 states plus selected areas [2,3]. The Abortion RA started in 1977 with five states and reached its current coverage of 14 states in 1987.

Mortality (approximately 2,000,000 annual events), fetal mortality (60,000) and natality (4,000,000) registration are required by all states; registration completeness for the mortality and natality systems exceeds 99 percent. The Abortion RA collects information on approximately 300,000 abortions per year, about 22 percent of the annual U.S. total. (Because of budgetary constraints, NCHS has not processed abortion data since 1993). The Marriage RA, excluding Puerto Rico and the Virgin Islands, covers approximately 81 percent (785,000) of U.S. marriages. The Divorce RA, excluding the Virgin Islands, accounts for 49 percent (280,000) of the annual U.S. divorce count.

In addition to these six registration systems, two other data systems, the Current Mortality Sample (CMS) and the Linked Birth and Infant Death Data Set, are based on data obtained from the Mortality and Natality Registration Systems. The CMS is a 10 percent systematic sample taken from the regular mortality file on a monthly (month of death) basis. The CMS covers the 50 states, the District of Columbia and New York City; it includes 17,000-20,000 deaths per month. The Linked Birth and Infant Death Data Set, which also covers the 50 states, the District of Columbia and New York City, links the more detailed information from the birth certificate with the information from the death certificate for each of the approximately 40,000 infants who dies before his/her first birthday.

Population Based Surveys

Three of the Center's data systems are classified as population based surveys. They are the National Health Interview Survey, National Health and Nutrition Examination Survey, and the National Survey of Family Growth. The designs of these surveys are based on stratified multistage samples of households, where the household is defined as the basic sample unit. Based on established criteria, a person (one or more) in the sample household is selected as the ultimate sample unit, i.e., the unit of analysis.

National Health Interview Survey (NHIS)

The ***NHIS*** is a continuing nationwide sample survey in which data are collected on the incidence of acute illness and injuries, the prevalence of chronic conditions and impairments, the extent of disability, the utilization of health care services, and other health related topics. Generally, personal interviews are completed in 47,000 households for about 123,000 sample persons.

National Health and Nutrition Examination Survey (NHANES)

The ***NHANES*** obtains nationally representative information on the health and nutritional status of the American population through a combination of personal interviews (mostly in the respondent's home) and detailed physical examinations. These examinations are conducted in specially equipped mobile examination centers (MEC) that travel around the country. The last survey, NHANES III, the sixth in the cycle of health examination surveys conducted since 1960 [4], collected data on topics such as high blood pressure, blood cholesterol, infectious diseases, diabetes, HIV infection, blood lead levels, allergies, osteoporosis, and other nutritional status measures.

The NHANES III [5], conducted over two 3 year phases, 1988-91 and 1991-94, covered the U.S. civilian, noninstitutional population aged 2 months and older. Each phase constituted a national sample of about 20,000 persons, with an expected interview completion rate of 85-90 percent and a response rate of about 75-80 percent for the medical examination. More than 78 percent of the persons selected for the 1988-91 phase participated in the medical examination. Selected subpopulations, children (< 5 years), older persons (60+), Black Americans and Mexican Americans, were oversampled.

National Survey of Family Growth (NSFG)

The Center's third population based survey, the ***NSFG***, is a periodic nationally representative household survey of women of reproductive age (15-44 years). The survey, first conducted in 1973 [6], collects data on fertility and infertility, family planning, and related aspects of maternal and infant health. The 1988 survey, the fourth in the cycle [7], selected 10,000 eligible sample households from the frame of households that participated in the NHIS between 1985 through 1987. A total of 8,450 women were interviewed in person, in their own homes, by trained female interviewers.

Provider Based Surveys

Seven NCHS data systems form the family of provider based surveys, collectively called the National Health Care Survey (NHCS). Included here are the National Hospital Discharge Survey(NHDS), National Survey of Ambulatory Surgery (NSAS), National Ambulatory Medical Care Survey (NAMCS), National Hospital Ambulatory Medical Care Survey (NHAMCS), National Nursing Home Survey (NNHS), National Home and Hospice Care Survey (NHHCS), and the National Health Provider Inventory (NHPI). Whereas population based surveys use the household as the basic sample unit, provider based surveys use the medical provider (physician, hospital, nursing home, etc.) as the basic sample unit. The provider furnishes information on samples of provider/patient contacts, e.g., office visits, hospital stays, nursing home stays, etc.

Samples for these surveys range in size from the approximately 475 emergency rooms in the NHAMCS to the 87,000 facilities covered by the NHPI.

Followup/Followback Surveys

Six of the NCHS data systems included in this report are classified as Followup/Followback surveys. They are:

National Maternal and Infant Health Survey (NMIHS)
1991 Longitudinal Followup (LF) to the National Maternal and Infant Health Survey
National Mortality Followback Survey (NMFS)
National Health and Nutrition Examination Survey Epidemiologic Followup Study (NHEFS)
Longitudinal Study of Aging (LSOA)
National Nursing Home Survey Followup (NNHSF).

Sample sizes range from 7,500 to 26,000 persons.

Data Processing Environment and Dispersion of the Work

There are many similarities in the data processing activities employed by NCHS offices for their respective data systems. This is especially true for data systems within the same "family of surveys." For example, registration areas provide NCHS with coded and edited computer tapes or microfilm copies of vital event certificates which are converted to uniform codes and subjected to machine edits. The other surveys use CAPI (Computer Assisted Personal Interview), preliminary hand edits, machine edits, etc. There are, however, a number of procedures that cross "family survey" lines that are gaining greater usage with the rapid advances made in survey technology. Of particular interest to NCHS is "source point data editing" (SPDE). This refers to editing survey data by any means of access to either the interviewer (or other data collector), the respondent, or records within a limited time following the original interview or data collection. The time limit reflects the period within which the persons involved can reasonably be expected to remember details of the specific interview or, in the case of data collected from records, a time within which there is reasonable expectation that there has been no change to the records which would affect the data collected. Thus, data completion and accuracy are much more likely to result when source point data editing is used.

Audit Trail - This term refers to a process of maintaining, either by paper or electronically, an accounting of all changes of sample or survey data item values and the reasons for those changes. The level of effort varies by data systems; some are manual, while others are automated.

Micro-, Macro-, and Statistical Editing

This section describes three types of editing processes. The following definitions are used in this section.

- **Micro-editing**.--Editing done at the record or questionnaire level.
- **Macro-editing**.--Editing to detect individual errors by checking on aggregated data or by applying checks to the complete set of records.
- **Statistical editing**.--Editing based on statistical analysis of respondent data. It may incorporate cross-record checks, as well as historical data.

Micro-, macro, and statistical editing for the eight registration systems are all very similar. Automated edits are designed to (1) assure code validity for each variable and (2) verify codes or code combinations which are considered either impossible or unlikely occurrences.

For each of the other three types of data systems, most or all of the following procedures are used:

- ❒ Extensive machine micro-editing.
- ❒ Where appropriate, comparison of current estimates with previous years.
- ❒ Assuring reasonableness of record counts, sampling rates, etc.
- ❒ Checking ranges, skip patterns, consistency of data from different sources.
- ❒ Checking medical data for compatibility with age and/or sex.

Priority of Edits

None of the registration systems gives special priority to any item in the editing procedures. The other data systems prioritize their edits based on:

- ❒ Identifiers needed to link data files.
- ❒ Questionnaire items used to weight sample data to national estimates.
- ❒ Medical data incompatible with demographic data.

Imputation Procedures

Imputation is defined as a process for entering a value for a specific data item where the response is missing or unusable.

Registration Systems

Except for Abortion Registration, which does not impute for missing items, imputation procedures among registration systems apply primarily to demographic items. In Mortality registration, imputation procedures are done by machine, which checks for invalid codes. The following variables are subject to imputation procedures: age, sex, date of death, marital status of decedent, race of decedent, and education of decedent.

Missing natality data that are imputed include child's race, sex, date of birth, and plurality. Data imputed for the mother include race, age, marital status and residence. Imputation is done by machine, either on the basis of a previous record with similar information for other items on the record (e.g., mother's age imputed on the basis of a previous record with the same race and total-birth order), or on the basis of other information on the certificate (e.g., marital status on the basis of mother's and father's

names, or lack of name). The tape documentation includes flags to indicate when imputation was performed.

Finally, marriage and divorce data imputation are limited to month of marriage (or divorce) and age of bride and/or groom (marriage only). Hot deck and cold deck imputation procedures are used. In hot deck imputation, a missing data item is assigned the value from a preceding record in the same survey having similar (defined) characteristics. In cold deck imputation, a missing data item is assigned the value from a similar record in a previous similar survey.

Population Based Surveys

Imputation procedures for the Center's other surveys differ from those used by the Registration Systems. In the case of the NHIS, 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. For example, in the NHIS, a missing "husband's age" (or "date of birth") is assigned the value of "wife's age + 2 years."

In the NHANES, the calculation of sample weights addresses the unit nonresponse aspects of the survey except for special cases.

In the NSFG, the sample weights adjust for unit nonresponse. Imputation of missing items in the NSFG was carried out by the contractor. For the most part, a hot-deck procedure was used to impute missing values.

Provider Based Surveys

The provider based surveys have established imputation procedures for three types of nonresponse: unit nonresponse, record nonresponse, and item nonresponse. Unit nonresponse is imputed by inflating the sample weight of similar responding units. Record nonresponse is imputed by inflating the sample weight of similar responding cases to account for the missing cases. Item nonresponse is imputed by inferring a certain or probable value from existing respondent information.

Followup/Followback Surveys

Four of the followback surveys, the LSOA, the NMFS, NHEFS, and the NNHSF, did not impute any data, although missing data items were filled in by using logical relationships as described in the above example. Unknown or inconsistent data were coded as "unknown." The other two used "hot deck" procedures.

Editing and Imputation Standards

For each of its registration systems, NCHS monitors the quality of demographic and medical data on tapes received from the states by independent verification of a sample of records of data entry errors. In addition, there is verification of coding at the state level before NCHS receives the data. All other

systems employ error tolerance standards established for interviewer performance (if applicable), and enforced by editing and telephone reinterviews. Error tolerance standards are also established for coding and keying of data, and are enforced by sample verification.

Costs of Editing

The costs of editing are very difficult to determine, though some surveys and data systems appear to have a better handle on this than others.

None of the eight registration systems could provide an estimate of their editing costs. All other systems estimated their data editing costs between 10-30 percent of total survey costs.

Role of Subject Matter Specialists

For all surveys and data systems, 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 to 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

- No sampling error for 100 percent registration systems; however estimates of variation are computed for vital events <20.
- Marriage and divorce data tables list sampling errors by area expressed as a percent of the area total.
- Other surveys produce estimates of sampling (but not non-sampling) errors.
  - Selected surveys present estimates based on assumptions regarding the probability distribution of the sampling error.

Current and Future Research

There are several ongoing research activities and a number of others are being considered for the future. Resource constraints, both money and personnel, are the major limiting factors. The following represent programmatic changes in data collection, data processing/editing, data analysis, etc., that will occur or be investigated in future years. Aside from these specifics, however, perhaps the biggest change, one which is well underway now, and cuts across all surveys, is the shift from paper and pencil data collection to computerized data collection. This shift makes it more difficult to omit data items, to enter inconsistent or impossible data, etc.

- Implementation of electronic birth and death certificates by the states.

- Implementation of the Super ***MICAR*** (Mortality Medical Indexing, Classification, and Retrieval) system by all states. The intent of Super MICAR is to allow data entry operators to enter cause-of-death information as it is literally reported on the death certificate. Under the current MICAR system, cause of death information must be entered using abbreviations or standardized nonmeclature (Harris et al., 1993). Implementation of Super MICAR is essential to a successful electronic death certificate system.

- Determination of optimum imputation techniques (single and multiple procedures) and their applicability to NHANES.

- Evaluation of automated data collection methodology for the NHDS.

- Feasibility of developing an automated system for coding and classifying medical entities using the ICD-10.

- Feasibility of developing an automated system for data correction and creation of an audit trail (Followback surveys).

Summary

Data editing practices at NCHS are quite extensive. Unfortunately, detailed descriptions of these practices for this report were precluded because of space limitations. However, some summary findings on NCHS data editing practices are provided below and in Table 2.

- Among NCHS data systems, the cost of data editing is the least documented variable. Only five data systems provided dollar and other resource costs of their data editing procedures. Most of the other data systems offered "guestimates" of 10-20 percent of total survey costs, with a few "guestimating" as high as 30 percent of total survey costs.

- About sixty percent of the Center's data systems collect data on an on-going basis throughout the year and publish data on an annual basis.

- Two-thirds of the Center's data systems report item non-response rates under 5 percent.

- One third of the Center's data systems use Computer-assisted telephone interviewing (CATI) as their primary or secondary data collection method.

- One-half of the Center's data systems release micro-data with identifiable imputed data items; another one-quarter release micro-data without identifying imputed data items.

- Virtually all NCHS data systems have rules establishing minimum standards of reliability that must be met in order to disseminate data.

- Slightly more than one-third of the Center's data systems monitor analysts/clerks in their data editing procedures; three-fourths monitor their automated editing procedures. However, only three data systems formally evaluate their data editing systems.

Table 2.--Frequency of Selected Data Editing Practices Among NCHS Data Systems

	Yes	No	NA/DK
1. Data dissemination limited by confidentiality (privacy) restrictions?	24		
2. Does data system release microdata (respondent level data)?	19	5	
3. Are imputed items identified?	13	11	
4. For aggregated data, is information provided on percentage of a particular item which has been imputed?	5	16	3
5. Are there minimum standards for reliability of disseminated data?	21	1	2
6. Is information available on the cost of data editing?	5	19	
7. Are there procedures for monitoring editors, clerks, analysts, etc.?	9	14	1
8. Are there procedures for monitoring automated editing procedures?	18	4	2
9. Is there an audit trail (i.e., a record kept) for some or all data editing transaction?	21	2	1
10. Are performance statistics maintained in order to evaluate the data editing system?	3	21	
11. Has any analysis been done on the effect of data editing on estimates produced?	5	19	
12. Is survey data editing information released?	15	9	
13. Is validation editing performed?	22	2	
14. Is macro-editing used?	17	7	
15. Are any other data editing techniques performed?	6	17	1

Reference

Harris, Kenneth W.; Rosenberg, Harry, et al. (1993). Evaluation of an Automated Multiple Cause of Death Coding System, *Proceedings of the American Statistical Association*, Social Statistics Section, Washington, DC.

Footnotes

[1] Statistical Policy Working Paper 18: *Data Editing in Federal Statistical Agencies*, Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget, Washington, DC, May 1990.

[2] National Center for Health Statistics: Data Systems of the National Center for Health Statistics, *Vital and Health Statistics*, Series 1-No. 16, Hyattsville, MD, December 1981.

[3] National Center for Health Statistics: Data Systems of the National Center for Health Statistics, *Vital and Health Statistics*, Series 1-No. 23, Hyattsville, MD, March 1989.

[4] National Center for Health Statistics: Cycle I of the Health Examination Survey, Sample and Response, United States, 1960-62, *Vital and Health Statistics*, Series 11-No. 1, Rockville, MD, May 1965.

[5] National Center for Health Statistics: Sample Design: Third National Health and Nutrition Examination Survey, *Vital and Health Statistics*, Series 2-No. 113, Hyattsville, MD, September 1992.

[6] National Center for Health Statistics: National Survey of Family Growth, Cycle I, *Vital and Health Statistics*, Series 2-No. 76, Rockville, MD, June 1977.

[7] National Center for Health Statistics: National Survey of Family Growth, Cycle IV, Evaluation of Linked Design,*Vital and Health Statistics*, Series 2-No. 117, Hyattsville, MD, July 1993. ■

Chapter 2

Fellegi-Holt Systems

Chair: John Kovar, Statistics Canada

William E. Winkler ◆ Thomas F. Petkunas

Joel Bissonnette

William E. Winkler ◆ Lisa R. Draper

Chapter 2

DISCRETE: A Fellegi-Holt Edit System for Demographic Data

William E. Winkler and Thomas F. Petkunas,
U. S. Bureau of the Census

Abstract

This document provides background on the workings and an application of the DISCRETE edit system. The system is a prototype whose purpose is to demonstrate the viability of new Operations Research (OR) algorithms for edit generation and error localization. While the OR algorithms are written in a general fashion that could be used in a variety of systems, the i/o, data structure, and imputation sections of the code are written in a survey-specific fashion. The source code cannot easily be ported to a variety of computer systems and is not easy to maintain. The first two sections consist of a description of the basic edit system and an example showing specific details of the input and output files used by the software. The final section is a summary.

DISCRETE: A Fellegi-Holt Edit System for Demographic Data

William E. Winkler and Thomas F. Petkunas,
U. S. Bureau of the Census

Description of the DISCRETE Edit System

The following subsections describe aspects of the DISCRETE edit system.

Purpose, Model, and History

The DISCRETE edit system is designed for general edits of discrete data. The system utilizes the Fellegi-Holt (FH) model of editing. Source code for DISCRETE was written by the author (Winkler, 1995a) and is based on theory and computational algorithms from Fellegi and Holt (1976) and Winkler (1995b).

Software and Computer Systems

The software consists of two programs, gened.for and edit.for. The software is written in FORTRAN and is not easily portable. With some work, the software runs on IBM-PCs under DOS and UNIX workstations. The programs run in batch mode and the interface is character-based.

The first program, gened.for, generates the class of implicit edits that are necessary for the error localization problem. The error localization problem consists of determining the minimum number of fields to impute so that an edit-failing record will satisfy all edits. It uses as input a file of explicit edits that have been defined by an analyst. As output, it produces the file of implicit edits that are logically derived from the explicit edits and also checks the logical consistency of the entire set of edits. The class of implicit edits that are generated are so-called maximal implicit edits. The class of originally defined explicit edits plus the class of maximal implicit edits is known to be sufficient for solving the error localization problem (Garfinkel, Kunnathur, and Liepins, 1986, hereafter, GKL) and known to be a subset of the class originally defined by Fellegi and Holt. The method of generating the maximal implicit edits is due to Winkler (1995b) and replaces an earlier method of GKL. The GKL edit-generation algorithm has a driver algorithm for traversing nodes in a tree and an algorithm for generating new implicit edits at each node in the tree. The nodes are the locations at which new implicit edits can be generated. The Winkler algorithm has a different driver algorithm for traversing the nodes in the trees, an in-between algorithm that determines the subset of edits that are sent to the implicit-edit-generation algorithm, and an edit-generation algorithm similar to the one of GKL.

The second program, edit.for, performs error localization (i.e., determines the minimal number of fields to impute for a record failing edits) and then does imputation. The input files consist of the set of implicit edits produced by gened.for and the data file being edited. The error localization algorithm

(Winkler, 1995b) is significantly faster than an error localization due to GKL because it first uses a greedy algorithm and then, if necessary, uses a cutting plane algorithm. Error-localization by GKL is via a pure cutting plane argument which is orders of magnitude slower than the greedy algorithm even with moderate size problems. While greedy algorithms can yield suboptimal solutions with general problems, greedy algorithms typically yield optimal solutions with edit problems. Cutting-plane arguments are generally known to be the most effective for solving integer programming problems (Nemhauser and Wolsey, 1988). Another difference between Winkler (1995a) and GKL is that the number of edits passed to the error localization stage grows at a somewhat slower exponential rate under Winkler (1995b) than under GKL. The slower exponential growth is due to a more precise characterization of the implicit edits needed for error localization (Winkler, 1995b). As computation in integer programming is known to grow faster than the product of the exponential of the number of edits and the exponential of the number of variables associated with the edits, the new error localization procedure should be much faster in practice. The imputation module of edit.for currently delineates the set of values for the minimal set of variables needing to be changed so that all edits are satisfied. In applications of the DISCRETE edit system, the imputation methodology currently consists of analyst-defined if-then-else rules of substitution. The substitutions for edit-failing data satisfy the edit rules and are very survey specific. Although general substitution rules within the restraints imposed by the Fellegi-Holt theory could be developed, they often would not be as acceptable to subject-matter specialists as the survey-specific rules. The advantage of the general substitution rules is that they would greatly speed the implementation on new surveys because analysts would not have to spend as much time defining edit rules and substitution rules.

The outputs from the second program consist of summary statistics, the file of edited (i.e., containing imputes) data, and a file giving details of each record that was changed. The details consist of the failed edits, the minimum fields to impute, and other information related to specific data records.

Documentation

The only documentation associated with the DISCRETE edit system is Winkler (1995a). The documentation is minimal and only describes how to compile and run the software on the example included with it.

Limitations

As computation grows exponentially as the number of variables and the number of value-states of variables increase, large systems of edits may be slow. At present, we do not know the the larges size the system will handle. The system, which has i/o modules based on an earlier system that utilized algorithms of GKL, does not easily recompile and run. A large number of include files must be modified and initial values of some data structures that describe the data are hard-coded.

As the software is an early prototype version, insufficient time has been spent on debugging source code. While the OR portions of the source code run perfectly on a variety of test decks, it may fail in certain data situations that have yet to be encountered. Because the i/o portions of the code are survey-specific, they are very difficult to port to new surveys because the size and initial values of several of the data structures need to be hardcoded in the include files.

Strengths

The DISCRETE system deals with completely general edits of discrete data. If the FORTRAN include files (see above) can be properly changed, then the software is straightforward to apply in all situations. Checking the logical consistency of the set of edits (via gened.for) does not require test data. Error localization (via edit.for) should be far faster than under previously written FH systems for discrete data.

Maintenance of DISCRETE Code

As it is presently written, DISCRETE code is not sufficiently well organized and documented so that it can be maintained. Hundreds of lines of code associated with i/o and data structures are survey-specific.

Future Work on DISCRETE

The DISCRETE system will be improved with general i/o modules, more efficient algorithms for determining the acceptable value-states of the set of error-localized variables, and an indexing method for tracking the set of imputes for each set of edit failures. The optimization loops of the error-localization code may also be improved. The advantage of the indexing method is that it will make the code more easily useable on large surveys such as censuses because many of the optimization loops associated with error localization will only be used once. A loop in the future code will produce a string based on the set of failing edits, perform a binary tree search on previously computed strings associated with edit failures, find the index and set of error-localized fields if the index exists, and, if the index does not exist in the existing trable, perform optimization and add the appropriate error-localized fields for the new index. The main overhead of the indexing method is a sorting algorithm that periodically rebalances the binary tree after a certain number of updates.

Example

The example basically shows what the inputs and outputs from running the two programs of the DISCRETE system look like. The first program generates all the implicit edits that are needed for error localization and checks the logical consistency of the entire edit system. An edit system is inconsistent when no data records can satisfy all edits. The second program uses the entire set of implicit edits that are produced by the first program and edits data records. For each edit-failing record, it determines the minimum number of fields (variable values) to change to make the record consistent.

Implicit Edit Generation

The first program, gened.for, takes a set of explicit edits and generates a set of logically derived edits. The edits are generated by the procedure of FH and consist of the smallest set needed for error localization. Two tasks must be performed. The first is to create an input file of explicit edits. The edits are generally created by subject-matter analysts who are familiar with the survey. An example is given in Table 1. There are 5 edits involving 6 fields (variables). The kth variables takes values 1, ..., nk, where the number of values nk must be coded in a parameter file. A record fails the first edit if variable 1 takes values 1 or 2, variable 4 takes values 1 or 2, and variable 5 takes value 1. Variables 2 and 3 may take any values in edit 1.

Table 1.--Example of Explicit Edit Input File

```
Explicit edit # 1:  3 entering field(s)
     VAR1           2 response(s): 1 2
     VAR4           2 response(s): 1 2
     VAR5           1 response(s): 1

Explicit edit # 2:  4 entering field(s)
     VAR2           2 response(s): 3 4
     VAR3           1 response(s): 2
     VAR5           1 response(s): 2
     VAR6           2 response(s): 1 2

Explicit edit # 3:  3 entering field(s)
     VAR3           1 response(s): 1
     VAR4           2 response(s): 2 3
     VAR6           3 response(s): 2 3 4

Explicit edit # 4:  2 entering field(s)
     VAR2           2 response(s): 1 2
     VAR4           2 response(s): 1 3

Explicit edit # 5:  3 entering field(s)
     VAR1           2 response(s): 2 3
     VAR3           1 response(s): 2
     VAR6           1 response(s): 1
```

Table 2.--Example of Selected Implicit Edits from Output File

6	VAR3	VAR4	VAR5	VAR6
	1	0	0	0
		1		
7	VAR3	VAR4	VAR5	VAR6
	1	0	1	0
		2		1
8	VAR4	VAR5	VAR6	
	2	1	1	
9	VAR3	VAR4	VAR6	
	1	0	0	
10	VAR2	VAR4	VAR5	VAR6
	2	1	1	1
	3	2		
11	VAR2	VAR3	VAR6	
	0	0	1	
	1		2	
			3	

The second task is to change a parameter statement at the beginning of the program and recompile the program. The statement has the form

PARAMETER (MXEDS=20,MXSIZE=8,NDATPT=8,NEXP=5,NFLDS=6).

MXEDS is the upper bound on the storage for the number of edits. MXSIZE is the maximum number of values that any variable can assume. NDATPT is the sum of the number of values that all the variables assume. NEXP is the number of explicit edits. NFLDS is the number of variables (Table 2).

The example of this section is a modified version of the example of GKL. The modification consisting of permuting the variables as follows: 1 -> 3, 2 -> 4, 3 -> 5, 4 -> 6, 5 -> 1, and 6 -> 2. The DISCRETE software generates all 13 implicit edits whereas the GKL software generate 12 of the 13 implicit edits. With an example using actual survey data and 24 explicit edits, the DISCRETE software generates all 7 implicit edits whereas the GKL software generates 6 of 7. The reason that the GKL software does not generate all implicit is due to the manner in which the tree of nodes is traversed. The GKL software traverses the tree of nodes according to their theory.

Error Localization

The main edit program, edit.for, takes two inputs. The first is the set of implicit edits produced by gened.for. The second input is the file being edited. A FORTRAN FORMAT statement that describes the locations of the input variables in the second file must be modified. A large parameter statement that controls the amount of storage needed by the program is not described because of its length. Eventually, the parameter statement will have to be described in comments.

Two output files are produced. The first consists of summary statistics. The second (see Tables 3 and 4) contains details of the edits, blank fields, and possible imputations for each edit-failing record. The edit code presently only delineates acceptable values for the fields designated during error localization. The actual imputed values could be determined via statistical modelling by analysts. The imputation could be written into a subroutine that would be inserted at the end of error localization.

In a typical application, the revised values (Tables 3 and 4) would not be left blank but would be imputed according to rules developed by analysts familiar with the specific set of survey data.

Application

A prototype application of the DISCRETE edit was developed for the New York City Housing and Vacancy Survey (NYC-HVS). This prototype was used to edit ten of the primary fields on the questionnaire. Data collected via the NYC-HVS are used to determine rent control regulations for New York City. The variables that we used in edits were: TENURE, PUBLIC HOUSING?, TYPE OF CONSTRUCTION (TOC), TOC CODE, YEAR MOVED, YEAR BUILT, YEAR ACQUIRED, CO-OP OR CONDO, RENT AMOUNT, and OWNER OCCUPIED. With previous edits, these fields were edited sequentially, starting with the TENURE field. The TENURE field reports whether the occupant of the dwelling is

- the owner,
- pays rent, or
- lives there rent free.

Table 3.--First Example of Edit-Failing Record in Main Output from EDIT.FOR

```
      Record #    1 (    1)       ID: 1001

Implicit edit #  1 failed:
   1. VAR1       :  2
   4. VAR4       :  1
   5. VAR5       :  1

 Implicit edit #  5 failed:
   1. VAR1       :  2
   3. VAR3       :  2
   6. VAR6       :  1

Implicit edit #  6 failed:
   3. VAR3       :  2
   4. VAR4       :  1
   5. VAR5       :  1
   6. VAR6       :  1

    Deleted fields:
   ------------------
   6. VAR6            5. VAR5

The weight of the solution is  2.1100

imputation candidates for field  6. VAR6       :
   3. 3
   4. 4

imputation candidates for field  5. VAR5       :
   2. 2
```

Field names	Reported	Revised	Weights	Failed Edits
VAR1	2. 2	2. 2	1.100	2
VAR2	4. 4	4. 4	1.090	0
VAR3	2. 2	2. 2	1.080	2
VAR4	1. 1	1. 1	1.070	2
*VAR5	1. 1	-1.	1.060	2
*VAR6	1. 1	-1.	1.050	2

Table 4.--Second Example of Edit-Failing Record in Main Output from EDIT.FOR

```
        Record #    2 (    2)          ID:  1002

 Implicit edit #  1 failed:
   1. VAR1        :  2
   4. VAR4        :  1
   5. VAR5        :  1

 Implicit edit #  4 failed:
   2. VAR2        :  1
   4. VAR4        :  1

 Implicit edit #  5 failed:
   1. VAR1        :  2
   3. VAR3        :  2
   6. VAR6        :  1

 Implicit edit #  6 failed:
   3. VAR3        :  2
   4. VAR4        :  1
   5. VAR5        :  1
   6. VAR6        :  1

 Implicit edit #  7 failed:
   2. VAR2        :  1
   3. VAR3        :  2
   5. VAR5        :  1
   6. VAR6        :  1

 Implicit edit # 16 failed:
   1. VAR1        :  2
   2. VAR2        :  1
   5. VAR5        :  1

 Deleted fields:
 ---------------
   5. VAR5          6. VAR6          4. VAR4

    The weight of the solution is    3.1800
 imputation candidates for field  5. VAR5                    :
   2. 2
 imputation candidates for field  6. VAR6                    :
   2. 2
   3. 3
   4. 4
 imputation candidates for field  4. VAR4                    :
   2. 2

                                                                          Failed
   Field names              Reported        Revised          Weights      Edits
   ---------------          ----------      ----------       ----------   --------
        VAR1                  2. 2            2. 2           1.100          3
        VAR2                  1. 1            1. 1           1.090          3
        VAR3                  2. 2            2. 2           1.080          3
       *VAR4                  1. 1            -1.            1.070          3
       *VAR5                  1. 1            -1.            1.060          4
       *VAR6                  1. 1            -1.            1.050          3
```

A sequential edit, implies an edit based on if-then-else rules. The advantage of sequential edits is that they are often easily implemented. A principal disadvantage is that they are not easily checked for logical consistency. Another disadvantage is that there has to be a initial field from which the remaining fields will be edited. The TENURE field was the initial field for the Annual Housing Survey (AHS). The initial field is never edited in the sequential edit application but can be using a Fellegi-Holt model.

The prototype edit considers all fields simultaneously. The TENURE field was edited in the same manner as the other nine fields. It would be a correct assumption that most respondents are aware of their living arrangement, making TENURE a very reliably reported field. Therefore, TENURE did hold a higher weight. However, there are other circumstances that would cause it to be incorrect. It still needed to be edited.

The explicit edits needed for the DISCRETE prototype were developed from the edits of the prior set of sequential edits. Only the 24 edits that exclusively included the ten fields were considered. Because of the existing sequential edits, the explicit edits needed for the prototype DISCRETE edit were developed with very minimal support from the subject-matter specialists. These 24 edits were run through the edit generator, gened.for, and 8 implicit edits were computed. The edit generator reduced the number of explicit edits to 23, because it determined that one of the explicit edits was redundant. There was now a total of 31 edits for the ten data items.

The DISCRETE prototype produced edited data that were only slightly cleaner than the sequential edit because the data for the AHS were quite clean. The AHS is a long-term survey in which responses are obtained by experienced enumerators rather than via mail responses. The results of the prototype edit were similar to those of the previous sequential edit, except for one striking difference. Using the prototype edit, the TENURE field was in conflict with other fields more often than the subject-matter staff had anticipated.

A second prototype was developed for the Survey of Work Experience of Young Women. This prototype showed the power of the DISCRETE system because it allowed the editing of a large number of data items involving a very complicated skip pattern. No edits had previously been developed for these items because of the complicated nature of the edit situation. The core data items consisted of WORKING STATUS, HOURS/WEEK, HOURS/WEEK CHECK-ITEM, OFFTIME, OVERTIME, and CURRENT LABOR FORCE GROUP. Overall, this prototype was developed for 24 data items. Using previous edit systems, these data items were not edited because of their complex relationships and skip patterns. However, these skip patterns were incorporated into the prototype as explicit edits. This turned out to be a surprising advantage of the simultaneous edit. Working with subject-matter staff, 42 explicit edits were developed for the 24 data items. The edit generator computed an additional 40 implicit edits for a total of 82 edits. Because of the use of the method of data collection used for this survey, the data were very clean. However, the results of this prototype were not as important as was the fact that the prototype was able to edit relationships that were previously considered too complex.

Summary

The DISCRETE system is a Fellegi-Holt edit system for general edits of discrete data. It is a prototype system that is written in FORTRAN. As currently written, it is not maintainable and not

easily portable. Due to new theoretical/algorithmic characterizations (Winkler, 1995b), the system should be more generally applicable than any currently existing system. Although no speed tests have been done, the software should be approximately as fast as other currently existing edit systems.

References

Fellegi, I. P. and Holt, D. (1976). A Systematic Approach to Automatic Edit and Imputation, *Journal of the American Statistical Association*, 71, 17-35.

Garfinkel, R. S., Kunnathur, A. S. and Liepins, G. E. (1986). Optimal Imputation of Erroneous Data: Categorical Data, General Edits, *Operations Research*, 34, 744-751.

Winkler, W. E. (1995a). DISCRETE Edit System, computer system and unpublished documentation, Statistical Research Division, U. S. Bureau of the Census, Washington, D.C., USA.

Winkler, W. E. (1995b). Editing Discrete Data, *Proceedings of the Section on Survey Research Methods, American Statistical Association*, to appear. ■

Chapter 2

Generalized Edit and Imputation System for Numeric Data

Joel Bissonnette, Statistics Canada

Abstract

Statistics Canada's Generalized Edit and Imputation System (GEIS) serves the needs of the professional statistician, and can be used to satisfy the edit and imputation requirements of primarily quantitative surveys. For the editing phase, it checks the consistency of edit rules and produces many summary tables based on survey data, which help to determine the edit rules to apply; it also uses advanced linear programming techniques to flag the fields to be imputed, according to the edit rules. Outlier detection is also available.

For the imputation phase, a choice of three methods is offered. Deterministic imputation identifies fields for which there is only one possible value that allows the record to pass the original edits. Donor imputation replaces the values in error by imputing those from the valid nearest-neighbour record that is most similar to the record in error. Imputation estimators provide imputation for individual fields using a variety of estimators: ratio, current and historical means, and historical values with or without trend adjustments. ■

Chapter 2

The New SPEER Edit System

William E. Winkler and Lisa R. Draper,
U.S. Bureau of the Census

Abstract

This paper describes the methodology, the workings, and an application of the SPEER (Structured Programs for Economic Editing and Referrals) edit system.

The original SPEER, developed by Brian Greenberg, is a Fellegi-Holt system for editing ratios of economic data and has been used on some of the largest U.S. economic surveys. The advantages of Fellegi-Holt systems are

- they check the logical consistency of an edit system prior to the receipt of data,
- the main logic resides in reusable mathematical algorithms with control by easily maintained input tables, and
- with one pass against an edit-failing record, the minimal number of fields are changed to values such that the entire record satisfies all edits.

The current SPEER system consists of entirely new FORTRAN source code and computational algorithms, is exceedingly fast, and is portable across all known computer systems. Merely by changing a parameter file consisting of input and output filenames and input FORTRAN format statements, SPEER can be run against entirely different surveys.

The New SPEER Edit System

William E. Winkler and Lisa R. Draper
U. S. Bureau of the Census

Description of the SPEER Edit System

The following subsections describe aspects of the SPEER edit system.

Purpose, Model, and History

The SPEER edit system is designed for ratio edits of continuous economic data. The system utilizes the Fellegi-Holt model of editing. The first version of SPEER was written by Brian Greenberg (Greenberg and Surdi, 1984; Greenberg and Petkunas, 1990) and the current version was written by William Winkler (1995). The computational algorithms, much of the imputation methodology, and the source code in the current version is new.

Software and Computer Systems

The software consists of two programs, gb3.for and spr3.for. The software is written in portable FORTRAN which should recompile on a variety of computers. It currently runs on IBM PCs under DOS, Windows, or OS/2, DEC VAXes under VMS, DEC Alpha under Windows NT, UNISYS, and a variety of UNIX workstations. The programs run in batch mode and the interface is character-based.

The first program, gb3.for, generates the entire set of edit bounds. The main input is a file containing at least three lines. The first line is the name of the input file of explicit edits, the second is the name of the output file of implicit edits, and the third is the name of the output summary file. If a fourth line is present, it consists of the FORTRAN FORMAT for the input file of explicit edits. If a file of variable names, BNAMES.DAT, is present, then the variable names in it are used; otherwise, default names of the form VRnnn are used where nnn can range as high as 999. After the main input is read, the inputs and outputs are the usual ones associated with edit-generation programs. The most important input is the file of explicit edits that have been defined by an analyst. This input must be in a fixed format that is specified in the program documentation. As output, gb3.for produces the file of implicit edits that are logically derived from the explicit edits and also checks the logical consistency of the entire set of edits. With appropriate test data, an auxilliary program D-MASO (also in FORTRAN) can help an analyst determine the lower and upper bounds on the ratios that are in the set of explicit edits. The appropriate test data might consist of prior year's edited data or (a subset of) the current year's data.

The second program, spr3.for, performs error localization (i.e., determines the minimal number of fields to impute for a record failing edits) and then does imputation. The main input is a control file with at least five lines. The five lines are (1) the name of the input file being edited, (2) the name of the file containing implicit edits, (3) the name of the output, (4) the FORTRAN format of the quantitative data in the file being edited, and (5) the number of variables (fields) being edited. Five additional lines are also

read in. They are (6) the name of the file containing implicit edits, (7) the name of the file containing variable names, (8) the name of the file of beta coefficients, (9) the file of weights, and (10) optional FORTRAN FORMAT for file of explicit edits. The first six lines are mandatory. If the last five lines or the associated files do not exist, then defaults are used. The weights affect which fields are imputed. The variables with lower weights are imputed before those with higher weights. After the control file is read, the input files consist of the set of implicit edits produced by gb3.for, the data file being edited, and a set of "beta" values associated with ratios. The beta values are determined a priori using an appropriate test deck and consist of regression coefficients under the model $y = \beta x$. There can be as many coefficients as there are implicit edits. The imputation methodology consists of first determining an imputation range for a variable so that edits are satisfied. Within the range, the first choice of imputation uses a reported variable that is not being imputed and the corresponding "beta" coefficient. After the first choice, a hierarchy of defaults based on the imputation range is selected. Regression imputation is only used when the appropriate beta coefficient is available and the variable being imputed is associated with a variable that is reported. By the Fellegi-Holt theory, any values of fields chosen in the imputation range necessarily yield complete multivariate records that satisfy all edits.

The outputs from the second program consist of summary statistics, the file of edited (i.e., containing imputes) data, and a file giving details of each record that was changed. The details consists of the failed edits, the minimum fields to impute, and the imputation methodology that was utilized for each field.

Documentation

Three documents describe the overall SPEER methodology and capabilities. They are Greenberg and Surdi (1994), Greenberg and Petkunas (1990), and Greenberg, Draper, and Petkunas (1990). The documents do not describe details of the algorithms or how to create and run the system for specific data bases. New computational algorithms (Winkler, 1995) eliminate much of the redundant computation of earlier versions of the SPEER system. Major restructuring of the computer code makes the system much easier to apply in new situations because only one FORTRAN FORMAT statement describing locations of input fields in the file being edited must be changed. Winkler (1994) describes how to develop and run a SPEER system.

Documentation related to the details of the software and how to run the software has been created for the first time (Winkler, 1995). The main documentation consists of instructions on how to run the example that is included on the disk with the software. Each program has internal documentation (in comments at the end) describing the nature and structure of the inputs and the outputs. The internal documentation should be sufficient to allow all but the most naive users to apply the software in a variety of situations. The new source code is more easily understood because of its modular structure. In most applications it is unlikely that source code (with the possible exception of two parameters that determine that amount of allocated storage) will need modification.

Limitations

SPEER only deals with ratio edits. For a new user, the file of explicit edits may not be very easy to develop. A statistical package should be used to determine those variables that are linearly related and the associated regression ("beta") coefficients. The regression model is $y = \beta x$. Those "beta" coefficients that are placed in an external file are used for the default imputations. If "beta" coefficients are not available for two variables that are associated via a ratio edit, then the default imputation is based on

allowable range that satisfies the edits. The best imputations require survey-specific modifications in which the imputation module is replaced by special code.

The main output from spr3.for is a large print file that contains details of the failed edits, the error localization, and the imputations that were made. The program spr3.for does not produce an output file that has the same FORMAT as the main input file being edited and that has appropriate quantitative data (missing or edit-deletes) replaced by imputations. This is not done due to the difficulty in writing necessary generalized i/o routines, documenting the routines, and getting users to understand how to carry and output additional information from the input file that does not pass through SPEER edits. The program spr3.for does produce an output file EDIT.OUT that contains all the quantitative data fields that pass through the edits and that contains the newly imputed values. It is output in a fixed format and could be merged in with the original data that passes through the edits because it corresponds on a line-by-line basis.

The program spr3.for does not impute values for variables in connected sets in which all values are blank. A set of variables is connected if they are connected via ratio edits. Connected sets form a natural partition of the entire set of variables being edited. If all variables in a connected set are missing, then imputation cannot be based on ratios and must be determined via default procedures that might possibly be based on data from a prior time period.

Strengths

The software is very easy to apply because only one format statement describing the locations and sizes of the quantitative being edited needs to be changed (Winkler, 1995). In situations where storage does not exceed the default storage of the program, the FORTRAN format statement can be read in from an external file. Thus, the software does not need to be recompiled when it is used on different data files. While the software will handle a moderately large number of variables (200+), the present computational algorithms, with suitable modification, could allow it to handle more than 2,000 variables. The software is fast. For instance, to generate 272 pairs of implicit edit bounds in each of 546 industrial categories for the Census of Manufactures requires only 35 seconds on a Sparcstation 20. Because ratio edits are basically simple, algorithms and associated source code are quite straightforward to follow or modify. For most situations, source code should not need any maintenance or modification. All core edit algorithms are in debugged code that is reusable. Checking the logical consistency of the set of edits (via gb3.for) does not require test data. Default imputations are quite straightforward to set up. A new software program cmpbeta3.for will compute the "beta" coefficients for all pairs of variables (fields) that are associated via the ratio edits that are explicitly defined. The program cmpbeta3 is approximately 50 times as fast as commercial software because it contains no diagnostics or special features.

Developing and Running A SPEER Edit System

This section provides an overview of how to create and run the SPEER edit system. It describes some of the non-SPEER components that must be used in addition to the SPEER components. It also gives the type of personnel that are useful as an edit team developing a system.

Developing an Edit System Using SPEER

There are three facets to the development:

- ❒ analysis of the data using statistical and other packages,
- ❒ development of a pre-edit system, and
- ❒ development of a SPEER system. If data from a prior time period are not available, then data obtained during the collection can be used.

Stage 1 proceeds with a variety of steps. The analyst would begin by running various tabulations on the data to determine means, variances, ranges, and other values. Next the analyst would run a regression package to determine which continuous variables are linearly related and to get a variety of diagnostics. The pairs of variables that are linearly related and the associated "beta" coefficients from the regression need to be stored. When data from a prior time period is available, then analysts often have much of this information already.

Stage 2 consists of preliminary edits that often do not require sophisticated rules. These can involve checking whether a State code takes a value within a set of correct values, a variable takes a value in a specified range, and a group of variables adds to a desired sum.

Stage 3 begins with determining the edit bounds for ratios. To facilitate the process, we have a software tool, D-MASO, developed by David Paletz, that delineates potential bounds and a variety of diagnostics. The analysts can then quickly determine bounds. SPEER software consists of two components. The first, gb3.for, generates the logically implied edit bounds and checks the consistency of the entire edit system. It does not require test data. The second component consists of the SPEER edit, spr3.for. It determines edit failures, the minimum number of fields (variables) that must be changed so that the record satisfies edits, and then does imputation. The first program only needs the set of explicit edit bounds as input. The second program needs the set of implicit edit bounds from the first program, the set of "beta" coefficients from the regressions, and the data file that is being edited. A new program cmpbeta3.for will compute the "beta" coefficients for all pairs of variables that are connected via ratio edits. The program requires the file of explicit ratio edits, the main file being edited, the FORTRAN format of the quantitative data in the file being edited, and the number of variables (fields) being edited. It computes beta coefficients for all pairs of variables that can be associated via implicit ratio edits.

Maintenance of SPEER Code

The code may not require any maintenance. If larger data structures are needed, then the two parameters at the beginning of the code should be changed and the program recompiled. If the imputation module is changed or a new one is developed, then updating merely involves substituting the new subroutine for the old.

The code is very modular and contains much internal documentation. In particular, comments at the end of the code give details related to running the programs.

Other Maintenance of a SPEER System

The analyst must document how the "beta" coefficients from the regressions are obtained. The program cmpbeta3.for can quickly produce the set of "beta" coefficients.

Edit Team

An edit team is most useful when it consists of at least one individual in each of the following categories: methodologist, analyst, and programmer. Development of an edit system is primarily a programming project once subject-matter and analytic needs are identified. The methodologist could be an economist, demographer, or statistician who is familiar with the Fellegi-Holt theory and can facilitate the programming of the system. The methodologist can provide an important focal point if the methodologist can make sure that programmers are given knowledgable information about system requirements and understands details of programming such as how long it takes programmers to develop new, difficult skills. The analyst is a subject-matter specialist who is familiar with the industries for which data are being edited. Often analysts and programmers have worked together successfully on other projects. Teams often start slowly because of the time needed to develop common terminology and communication skills. Once team members are working closely together, however, final products are often better because individuals are stimulated by detailed knowledge provided by other team members.

Example

The example basically shows what the inputs and outputs from running the two programs of the SPEER system look like. The first program generates all the implied edits that are needed for error localization and checks the logical consistency of the entire edit system. An edit system is inconsistent when no data records can satisfy all edits. The second program uses the entire set of edits that are produced by the first program and edits data records. For each edit-failing record, it determines the minimum number of fields (variable values) to change to make the record consistent.

Implicit Edit Generation

The first program, gb3.for, takes a set of explicit edits and generates a set of logically derived edits. The edits consist of the lower and upper bounds on the ratios of the pairs of variables. Two tasks must be performed. The first is to create an input file of explicit ratio bounds. The bounds are generally created by subject-matter analysts who are familiar with the survey. An example is given in Table 1. The eight fields of the input file are: form number, edit-within-form-number, variable number of numerator, variable number of denominator, lower bound on ratio, upper bound on ratio, an intermediate value between the lower and upper bounds, and the four-character names of the variables. The form number describes the industry to which the edit refers. With U.S. Bureau of the Census surveys, the same form may be sent to all companies over a broad range of industrial classification categories. Separate ratio bounds need to be developed for each industrial classification.

Table 1.--Example of Explicit Ratio Bound Input File

110	1	1	2	.0212400	.0711125	.0369900	EMP1/APR2
110	2	2	3	1.5369120	6.8853623	3.2590401	APR2/QPR3
110	3	3	2	.1670480	.5273000	.3068400	QPR3/APR2
110	4	4	2	.0202880	.2717625	.0929800	FBR4/APR3

The second field refers to the edit number. It is primarily for the benefit of the analysts and is not used by gb3.for. The next two fields are the variable numbers of the fields in the ratio and the following two are the lower and upper bounds created by the analysts. The final two fields are not used by gb3.for but can be used by the analyst. The next-to-last field is possibly an average or median value that the analyst enters in the input file. The last field is a character representation that helps the analyst remember the variables. For instance, QPR3 might refer to "quarterly payroll" and APR2 might refer to "annual payroll."

The second task is only needed if default storage allocations are not sufficient. The task requires changing a parameter statement at the beginning of the program and recompiling the program. The statement has the form

PARAMETER (BFLD=45).

BFLD refers to the upper bound on the number of variables (here 45) being ratio edited. The number of variables being edited is assumed to be the same in every industry if more than one industry is edited. For the example, the output file primarily contains the ratio bounds (implicit edits) for the six pairs of the four variables.

Error Localization

The main edit program, spr3.for, takes three inputs. The first is the set of implicit edit ratios produced by gb3.for. The second is a set of "beta" coefficients that are created by a regression package that the analyst has used. The third input is the file being edited. A FORTRAN FORMAT statement that describes the locations of the input variables in the third file must be modified and placed in an external file. A parameter statement at the beginning of the program

```
PARAMETER(BFLD=45,BCAT=3,NCENVL=BFLD,NFLAGS=9,N_FLG=100,
 + NEDIT=BFLD*(BFLD-1)/2,MATSIZ=BFLD)
```

must also be changed. BFLD and BCAT are upper bounds on the amount of storage that is allocated. NFLAGS and N_FLG are upper bounds on storage for errors for a single record. In many situations, the default values of these parameters will be sufficient. If they are not, then parameter values will need to be increased and the program must be recompiled. Comments at the end of the source code give many details of setting up and running the program.

Two output files are produced. The first consists of summary statistics. The second (see Table 2) contains details of the edits, blank fields, and imputations for each edit-failing record. The output shows what edit has failed, the minimum number of fields that must be imputed, the imputation method that was adopted, and the revised and reported values of the record.

The program spr3.for is set up so that a more sophisticated imputation can easily be substituted for the existing one. Basically, analysts would have to do more modelling and determine a hierarchy of imputations that would be coded in a subroutine. The imputation subroutine would be added to the code and the eight lines associated with the existing (default) imputation would be replaced by a call to the subroutine. Documentation in the code clearly shows where the substitution should be made and what data must be passed to and from the imputation subroutine.

Table 2.--Example of Edit-Failing Record in Main Output from SPR3.FOR

```
Record #  1
    Failed edits:
        1.8964540  < APR2 / QPR3   <    5.9863030

Deleted fields:   3. QPR3
Imputation range for QPR3 :  Lo =        3.3410      Up =  10.5460
  QPR3  imputed using  QPR3 / EMP1  ratio

          Fields        Revised       Reported       Lower        Upper
          -------       ----------    -----------    --------     --------
          EMP1           1.000         1.000          .425         1.422
          APR2          20.000        20.000        14.062        34.207
          QPR3           5.714        13.000         3.341        10.546
          FBR4           3.000         3.000          .406         5.435

Record #  5
     Failed edits:
       .0402807 <   EMP1 / QPR3 <      .4257010
      1.8964540 <   APR2 / QPR3 <     5.9863030

Deleted fields:   3. QPR3
Imputation range for QPR3 :  Lo =        6.6819      Up =  21.0920
  QPR3  imputed using  QPR3 / EMP1  ratio

          Fields        Revised       Reported       Lower        Upper
          -------       ----------    ----------     -----        -----
          EMP1           2.000         2.000          .850         2.845
          APR2          40.000        40.000        28.124        68.415
```

Application

SPEER is currently being used in two large interactive applications. These applications are the Annual Survey of Manufactures and the Census of Manufactures and Mineral Industries. The applied system, named LRPIES (Late Receipts Processing and Interactive Edit System), is used primarily for basic data entry and editing, editing of late receipts, and processing establishment adds. The current version has features that facilitate analysts' review and correction of data records. Analysts in Washington can now enter and correct late receipts that arrive after the central data processing center in Jeffersonville, Indiana has shut down. Previously, late data were entered but generally left unedited. Analysts can also perform additonal review of the non-late data that were previously edited at the Jeffersonville location.

The SPEER application (LRPIES) involves the largest U.S. surveys of industry and manufacturing. As much analyst review of data is needed, custom software modifications that provide assistance and review capability have been added. The modifications are specific to Digital VAXes and the large

screen display capabilities of the types of VAX terminals in use. Records that have failed edits and that require imputation to make them consistent with the set of edits can be retrieved and processed interactively. For each edit-failing record, a number of values are displayed that facilitate the analysts' review and correction. The values are current values, a prior time period's corresponding values if available, suggested impute values, and ranges in which values can be imputed that are consistent with the set of edits. Analysts -- possibly after a call-back -- have the capability of entering a flag that causes an edit-failing value to be accepted. The custom code in LRPIES associated with the interactive edits is the majority of the code. The main SPEER subroutines merely need to be called and do not need to be modified.

The LRPIES application needs edit parameters and information for 546 SIC (Standard Industrial Classification) codes. The main edit parameters are the lower and upper bounds associated with the ratios being edited. Bounds from a prior year are often used as the starting point in producing the bounds for the current year's edits. Edit bounds and information can vary substantially across SIC codes. The specific parameters and information are the implicit edits for the current year and the prior year, the industry average value, and the beta coefficients obtained from regressing one of the variables (fields) in a ratio against the other variable. While the basic SPEER imputation merely uses a regression imputation, the LRPIES application uses a hierarchy of imputations based on the existence of prior data. The exact types of imputations and the hierarchy are determined by analysts familiar with the data.

Summary

The SPEER system is a Fellegi-Holt edit system for ratios of linearly related data. It is written in portable FORTRAN, easily applied, and very fast. Applications of SPEER include some of the largest U. S. economic surveys.

References

Greenberg, B. G.; Draper, Lisa; and Petkunas, Thomas (1990). "On-Line Capabilities of SPEER," presented at the Statistics Canada Symposium.

Greenberg, B. G. and Surdi, Rita (1984). "A Flexible and Interactive Edit and Imputation System for Ratio Edits," SRD report RR-84/18, U.S. Bureau of the Census, Washington, D.C., USA.

Greenberg, B. G. and Petkunas, Thomas (1990). "Overview of the SPEER System," SRD report RR-90/15, U.S. Bureau of the Census, Washington, D.C., USA.

Winkler, W. E. (1994). "How to Develop and Run a SPEER Edit System," unpublished document, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C., USA.

Winkler, W. E. (1995). "SPEER Edit System," computer system and unpublished documentation,Statistical Research Division, U.S. Bureau of the Census, Washington, D.C., USA. ■

Chapter 3

On-Site Data Capture

Chair: George Hanuschak, National Agricultural Statistics Service

Wouter J. Keller ◆ W. F. H. Ypma

Janet Sear

Anne Rhodes ◆ Kishau Smith ◆ Peter Goldstein

Chapter 3

Electronic Data Interchange for Statistical Data Collection

Wouter J. Keller and W. F. H. Ypma, Statistics Netherlands

Abstract

This paper gives a brief description of some of the information-technological developments within Statistics Netherlands. After an overview of the effects on the production process, it focuses on one aspect, Electronic Data Interchange (EDI). Among the many projects currently running at Statistics Netherlands, "Pilot 2" is described. This concerns EDI on the financial accounts of enterprises. We will focus on the role of the meta-information as a tool to control the process. We will see how technology changes this role and generates new possibilities to enhance the effectiveness of the meta-information.

Electronic Data Interchange for Statistical Data Collection

Wouter J. Keller and W. F. H. Ypma
Statistics Netherlands

Introduction

Statistics Netherlands is at present under the influence of several developments. As everywhere else, it no longer operates as an untouchable organisation of civil servants. Efficiency and market-orientation are the key-words now. We need to produce at lower costs. Furthermore, we need to lower the costs we inflict upon our suppliers of data. The outcome should be a product that, although not actually sold on a market, our clients eventually want.

Furthermore, we are confronted with new developments in Information Technology (IT). They will give us the opportunities to construct the necessary tools to meet the new demands.

In a situation like this, (NSI) needs to make the right strategic choices.

Demand Pull

The production process is, on the one hand, influenced by the growing demands of our clients and respondents.

There is a strong political demand for a decrease in the respondent burden, as a part of alleviating the administrative burden of enterprises. Statistics Netherlands sends out 1.25 million questionnaires to enterprises and other institutions per annum. Large and medium-sized enterprises may receive as many as 50 questionnaires per year, including repetitive monthly and quarterly surveys. In particular, larger companies in manufacturing are subjected to many (about 20) different types of surveys. The conclusion is clear: Statistics Netherlands has "to fight the form-filling burden."

Budgets are shrinking, so there is a demand for higher efficiency and higher productivity.

Concerning our output, we see a demand for a higher user-friendliness. One particular aspect is a demand to improve the coherence of all of the information we offer. Another aspect is that our clients will want to be able to use the new media IT has to offer.

Technology Push

On the other hand, we are blessed with information-technological developments or the technology push.

In the first place, these developments give us new technical possibilities, the means to construct new tools for our production process. We see large improvements in the possibilities of data processing data storage and data transmission. The latter aspect will probably have the most striking influence on our work: the communication of data between our respondents and the NSI on the one hand and the communication of data between the NSI and its clients on the other.

In the second place, these new developments create their own demand. The new technology will be used anywhere. Our suppliers of data will use it. Our clients will use it. They will no longer be satisfied to communicate with us in the old way -- that is, on paper. Our suppliers produce their data by electronic means and will want to use those means to deliver those data directly to us in order to minimise their own costs. Our clients process our data by electronic means. They will demand to be able to select and receive those data with the tools that IT has to offer.

These two factors lead to the conclusion that the NSI will have to make those strategic choices in its production process that make the best use of the possibilities IT has to offer.

Strategic Choices

New demands and new tools will affect all the aspects of our production process. To describe them, let us first discern, within this production process, three stages. The input phase is where the data are collected in contact with the respondents. In the through-put phase these data are processed to produce the information with the characteristics we are actually looking for. In the output phase this information is offered to and disseminated among our clients.

Let us begin with the input side -- the collecting of data -- start with data collection among individuals and households. It is not saying too much when we state that a major step forward has already been taken at Statistics Netherlands. We have introduced all kinds of Computer-Aided Interviewing (CAI) and developed BLAISE to do so. (Needless to say, BLAISE does more than develop and present electronic questionnaires.) The gains of these developments were mainly in terms of an increase in productivity or efficiency. The number of staff needed for coding, data entry and checking decreased dramatically. This efficiency also shows itself in the much faster production of results. There is greater efficiency of the production process, itself, but, also, in the statistical sphere, where improvements are still possible: new ways of interviewing: CASI, computer-aided self-interviewing, and, not directly a matter of IT, more efficient sample designs. Much more, however, is still to be done in the field of collecting data among enterprises. The demands here are stronger. Response burden has become an issue. It is the driving factor behind our strategic choices here. When we see at the same time that almost everywhere automation and IT has invaded the bookkeeping systems of the respondents involved, it is clear what our task for the nearby future will be: the "Edi-fication" of the collection of information from enterprises by the NSI. What CAI is for interviewing among households, EDI (electronic data interchange) will be for data collection among enterprises. Later in this paper we will go deeper into EDI with enterprises.

In the through-put phase, we are looking for more efficient ways of processing our data. Of course, CAI and EDI make much of the editing superfluous. Less errors will be made. Still, we expect much from more efficient or rational ways to handle the editing process. Here data processing is the key. The choice will be that we will no longer edit each individual record. It should be possible to use the computer to find the worst errors and help to correct them. At the same time, the computer can prevent us from spending time and money on correcting unimportant errors. The gains here are primarily productivity gains.

Finally, in the output phase, the new developments probably get the most attention from the public. We see the new media by which information can be presented to its users. Paper publications may continue to play their role, but especially the more professional user will want to select and receive his data by electronic means. Statistics Netherlands is producing or developing both data on CD-ROM and data on Internet.

More important and maybe more difficult is the way data should be presented with those new media. The amount of information will be much larger than we had in our paper publications. Thus, the management of the meta-information becomes crucial. For this purpose, Statistics Netherlands is developing STATLINE. This should lead to a database intended for the end-users that should give access to "all" our data. As could be expected, structuring those data is the main problem. At the same time, we are confronted with a lack of coherence due to lacking statistical coordination. Still, we aim at a first complete version of STATLINE by the beginning of 1997.

STATLINE is intended to play a key role in the dissemination process of our data. The strategic choice has been made that we aim for a structure, wherein all publications and all there dissemination of data goes through STATLINE.

Restructuring the Production Process

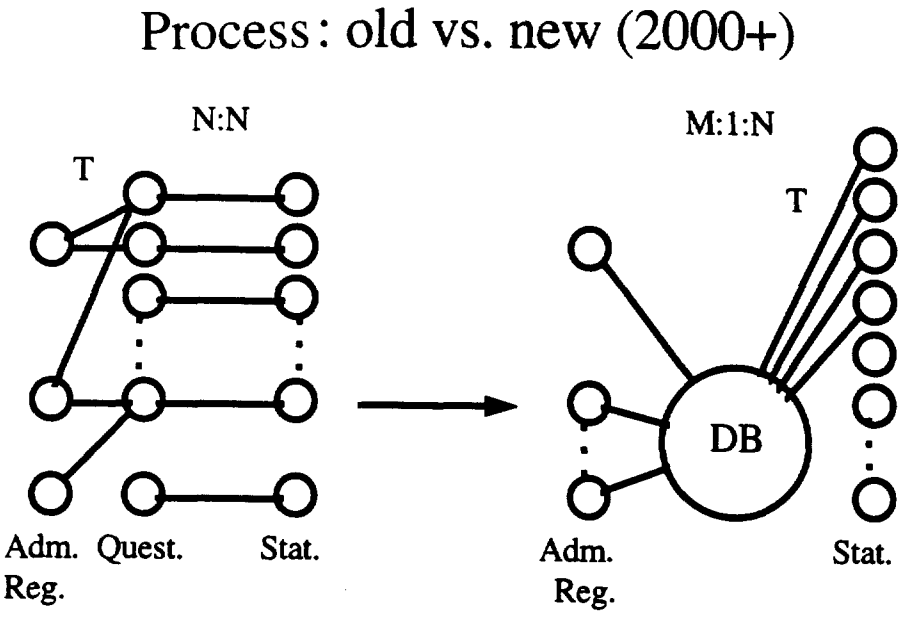
In the previous section we described the strategic choices we made regarding the different phases of our production process. Those choices go further than just the development of a new tool. They will affect the structure of the production process, itself. One should be prepared to take those consequences, as well.

The present or the "old" way the production process is structured is along the lines of the individual statistics. For each statistic -- an end product -- a new questionnaire is designed, respondents are selected, data are processed, and a publication is produced. Especially on the input side this is inefficient.

In the new situation -- we are talking about more than 10 years from now -- especially the data collection will be re-ordered. No longer the demand for information but the supply -- the available actual data sets -- will dictate the organisation there: the sources. Each source will be tapped once and completely for any possible use within the NSI. The collection is technically and conceptually adapted to that source. (In the remaining sections of this paper, we will give some indication regarding the nature of those sources.)

Having collected the data, we may have to translate them to statistically suitable concepts, integrate them, and we will have to distribute them among users. They may be inside the NSI -- e.g., the integrative systems like the National Accounts -- or outside the NSI. This means that somewhere those data will have to come together for distribution.

For the input side this can be illustrated as follows:



On the left we see the old situation, with a separate production line for each individual statistic. On the right is the future situation. There, all the possible sources contribute to a central database of relevant information. From that database the actual statistics are produced by combining the relevant information. It is evident that in order to combine information one should be certain that the characteristics of that information are such that combination makes sense. Those characteristics are specified in the meta-information.

Electronic Data Interchange (EDI)

From now on we will focus on EDI with enterprises and institutions.

An NSI collects data to produce statistical output. What needs to be done is making a translation from the data of the respondent to the data of the output. This is done in several steps. The first step may be left to the respondent. If so, it leads to a certain response burden.

The first step of the translation involves two parts. First there is the conceptual translation, the mapping of the concepts of the source, the administrative concepts on the concepts to be delivered to the NSI. This is the most difficult part. Not only do business records differ from statistical information, but they also differ among themselves. The second part of the translation is a technical one. We

would like to receive data in a suitable technical form; especially so that we and our respondents avoid data entry.

Modes of EDI

Electronic data interchange will be one of the strategic tools to meet the challenge of lowering the response burden and improving our productivity. In every individual case, we should decide whether to use it and in what mode.

We will describe several modes of EDI and judge them by their effect upon the response burden. Of each possibility we will indicate the nature of the translation and, especially, who is going to make it. We concentrate on the conceptual translation.

EDI on Centrally Kept Registers

Here, we do not approach the individual respondent at all. We are dealing with centrally kept information on individual units, collected for other purposes than statistics and yet of interest to the statistician. In itself, this kind of data collection creates no response burden.

There are, however, disadvantages. The most important is that there is very limited choice as to the conceptual contents of the data the NSI receives. In other words, one cannot ask for much translation towards statistical concepts. That will have to be done by the NSI, itself.

The second problem is closely connected and is that of units and populations. Here, also, one cannot but accept what the register keeper is able to supply. If the units he uses do not comply with the statistical units there is a problem. The same is true regarding the classification of those units. How can we connect the register population to our total statistical population?

A third problem regards the sampling strategy. If the register provides us with yearly data on, let us say, 70 percent of a population we formerly used to describe with a rotating sample of 1 out of 5, then what should our strategy be regarding the remaining 30 percent? In the Netherlands there are several examples of usable registers. There are centrally kept registers of enterprises with the Chambers of Commerce. The tape of these registers feeds our own register of statistical units. Statistical data can also be had from fiscal (company tax, value-added tax or VAT) or social security sources. In several cases (Chambers of Commerce, company tax and VAT), the possibilities are used or being researched.

Commercial Bookkeeping Bureau

A related option is tapping the information from the Commercial Bookkeeping Bureau. They keep the records on financial information or regarding the wages of sometimes a large number of individual enterprises. This possibility also is attractive because of the large number of respondents involved with only one link. Furthermore, these service bureaus will be capable of providing us with more information than, e.g., the fiscal records contain. A disadvantage is that these service bureaus probably will charge their clients for answering the questions of the NSI. Not every client will be prepared to pay.

While these bureaus often hold much of the information the NSI needs, there are two possibilities regarding the question of who will make the translation. The answer is a matter of cost benefit analysis. There is an example at Statistics Netherlands of one bureau that does the bookkeeping of 40 percent of the enterprises in one particular branch. In that case, it is profitable for the NSI to make the necessary translation. In other cases, we propose to provide software by which the bureau itself, makes the necessary translation.

EDI on Individual Respondents

When the above described options are not available, we will have to approach the individual respondent. In doing so, we should be aware of the fact that sometimes we will have to discern within one statistical unit, often an enterprise, several sets of administrative records. We will see that we will have to approach these subsets separately and in a different manner. Within commercial enterprises, we find the financial records, the logistical information (foreign trade, stocks), and the records on wages and employment. Especially the financial records and those on wages are strictly separated in the Dutch situation.

Here we classify by the translator of the information.

The NSI Translates

One of our EDI projects -- EFLO -- works along these lines. It deals with the data from the Dutch municipalities. They deliver a set of records directly tapped from their own complete set of records. The translation is done at Statistics Netherlands. The advantages in terms of respondents' burden are evident. Although extra work by the NSI is needed, this extra work can be seen as an investment, depending on the stability of the translation scheme. It is expected that this form of EDI will lead to an improvement of productivity once the translation schemes are completed. It is important that we are dealing with a limited number (600) of respondents.

The Respondent Translates to a Standard Record

Here, a standard record of information is defined. The standardisation relates to both the conceptual and the technical aspects. To produce the record, writing the software is left to the respondent. Working with a standard record is not always possible. It can only be done when the information is already standardised among respondents to a certain degree. Furthermore, to make a standard record possible, the NSI sometimes may have to move towards the concepts of the respondent. In that case, a larger part of the total translation to the final statistical output has to be done by the NSI.

Especially when the standard record is available in the bookkeeping software the respondent uses and regularly updates, this mode of EDI has a clearly favourable effect on the respondents' burden.

There are two examples. One is IRIS, the EDI on INTRA-EC trade. The standard record developed here is implemented in over 40 software systems available on the Dutch market, after certification by Statistics Netherlands. The EGUSES project is the other example. It concerns wage informa-

tion. That subset of company records is highly regulated in the Netherlands, making it possible to define a standard record.

The Respondent Translates -- No Standard Record

Still, a very large part of the information we are looking for is left out. The respondent has it in a form that conceptually and technically differs from what the NSI wants and from what other respondents have.

- **Paper Questionnaires**.--This clearly is no form of EDI. We mention it as a possibility to be complete and to emphasise the point that, here, the respondent does all the translating by himself and each time has to do it all over again.

- **Electronic Questionnaires**.--Although strictly speaking at most partial EDI, this method proves very successful with IRIS, the software on INTRA-EC trade. (IRIS works with a standard record, as well as with data entry.) By providing extra help functions and the possibilities of adapting the questionnaire to the individual respondent, it also helps to lower the response burden.

- **"Full" EDI**.--The last possibility is that the NSI provides the software by which the respondent can set up a translation scheme for both the technical and the conceptual translation. Once set up, and in so far as no changes occur, the scheme can be used to produce data to be delivered to the NSI. The example here is EDI -- Pilot 2, directed at the financial records and described in the next section.

Before we go into that, we give a summary of the characteristics of the several possibilities of EDI on individual enterprises:

(Sub)sets of records	Financial Wages Logistics All records
Translator	NSI Respondent
Output of Respondent	Not translated data Standard record Non-standard record Data entry: electronic questionnaire paper questionnaire

EDI -- Pilot 2

We will now describe the project EDI -- Pilot 2, directed at the financial records of individual enterprises as an example. It shows the problems one has to face. While describing Pilot 2, we can refer to the scheme in the previous section.

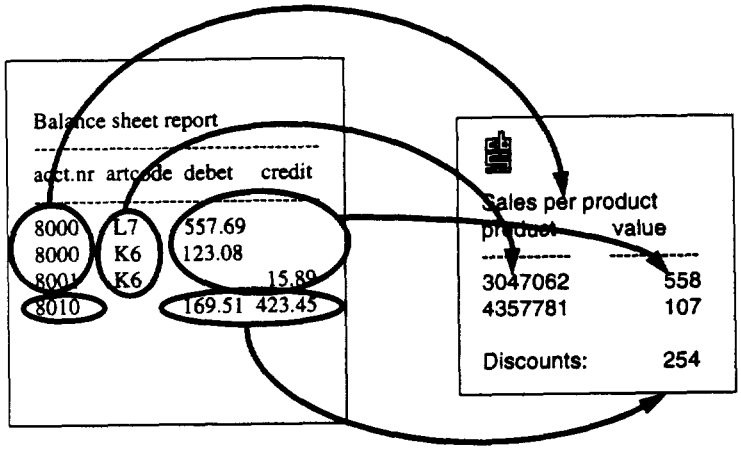
Pilot 2 is directed towards individual financial accounts. In the Dutch situation these are only a part of the accounts of an enterprise. Specifically, the accounts on wages and employment are excluded. This is not a choice voluntarily made by Statistics Netherlands, but one forced upon us by the way the bookkeeping systems are organised in our country. Leaving out detailed questions on wages, we combine within Pilot 2 all the questions that are put to the financial accounts. The result is the combined questionnaire.

The contents of the combined questionnaire are dictated by what is available in the financial accounts. Regulated as our society may be, the financial accounts may diverge strongly in internal organisation and in the concepts used. In the first place this means that we will have to adapt our questions towards the possibilities of the automated system of the enterprises. This may imply more statistical work for the NSI to reach the same output. If one wants more, it will probably be necessary to ask for additional information to be given explicitly by the respondent; that involves data-entry. In the second place, the diversity of respondents means that a unique translation scheme will have to be set up and maintained for each respondent.

Financial accounts also differ in their technical layout. A large number of bookkeeping software systems is in use. There is no standard record for information to be selected electronically from the software and it is not expected that it will be possible to define one within the near future. As the main goal of Pilot 2 was the lessening of the respondents' burden, it was decided that the amount of data-entry was to be minimized.

That means that some ingenuity was needed to create the automated link we were looking for. This is done by using the reports or printouts of the software system. Instead of printing them, they are sent to a file, a printfile, to be read by the translator, the main part of the software module that will run on the respondents' computers; it is now being developed as part of Pilot 2. The layout of the reports, and thus of the printfiles, is fairly stable. The respondent communicates this layout to the translator. He defines rows and columns within the report. Subsequently, he tells the translator how to manipulate the rows and columns in order to transform the information in the report to the statistical information asked for by the combined questionnaire. The resulting records are sent over to Statistics Netherlands.

The Translator



We see then the two parts of the translation scheme. The first part lays down the layout of the printfiles to make the technical transformation. The second part defines the conceptual transformation of the information to be found on the printfile towards the statistical information asked for on the combined questionnaire.

The final question is who will make that translation scheme. One of the principles of Pilot 2 is that "the respondent translates." This means that the respondent himself has to set up the translation scheme. This, of course, makes it less respondent-friendly. It seemed, however, impossible to set up those translation schemes at Statistics Netherlands. It is clear that this is not an easy task for the respondent. On the one hand, this means that a strong help-desk and a fairly large field service is needed; and, on the other hand, this means that even with Pilot 2 we will not yet reach the ultimate user-friendliness of EDI.

We expect the translation scheme to be fairly stable or, in other words, that technical and conceptual changes will not be too frequent. In subsequent years, the translator can use the already available translation scheme to produce the statistical information. Answering the combined questionnaire then becomes a matter of minutes instead of hours and can be handled by a less qualified employee. That is what makes the concept attractive and the initial investment worthwhile to the respondent.

Scope of Pilot 2

As said, Pilot 2 is directed towards the financial accounts. The principle is that all the information that is tapped from the financial accounts by any statistic of Statistics Netherlands will go through Pilot 2 if automated retrieval of that information is possible. In practice, this means that several large statistics will switch completely to EDI. For industry, our main target, we find:

- Monthly statistics on total turnover
- Monthly statistics on foreign trade, by product
- Quarterly statistics on turnover by product
- Yearly statistics on gross investment
- Yearly statistics on the production process
- Yearly statistics on the financial processes, inc. balance sheets.

The participation of foreign trade is a pilot within the pilot. Not only does Statistics Netherlands already have a successful EDI on this area in IRIS, but also the possibilities of getting enough foreign trade data, when aiming primarily at the financial accounts, still have to be researched.

Some questions in the above-mentioned statistics are dropped -- e.g., the questions on quantities of energy used in the production statistics. They cannot be addressed by this form of EDI. Probably a separate paper questionnaire on this subject will be sent.

On the other hand, some questions originating from other statistics -- mainly aimed at other subjects and accounts (e.g., the labour and wage accounts) -- are included, because the answers are typically to be found within the financial accounts of the enterprise.

The domain of EDI consists of those commercial enterprises that have set up financial accounts by means of computer software that satisfies certain technical specifications. In practice, this means

that we direct ourselves towards the profit sector within industry, trade, and services. We start with industry, because there the gains in terms of lessening the respondents' burden will be the largest. Individual smaller enterprises are not included, because their bookkeeping and automation capacities are expected to be too low. In view of the relative small amount of information asked here, more is expected from centrally-kept records (VAT, corporate tax)and from bookkeeping bureaus which often keep books for hundreds of smaller enterprises. The very large enterprises are also excluded. Because of their complexity, they need an individual approach; of course, in the end, also by means of EDI but then "tailor made."

Regarding the number of respondents participating in this kind of EDI, we should mention that in Pilot 1, 12 respondents participated and still do. Pilot 2 will start with a field test next March aimed at 20 respondents. Starting September 1996, we aim at larger numbers. By the end of 1996, Pilot 2 should handle several hundred respondents. Pilot 2 will also be used to approach the bookkeeping bureaus. That will lead to larger numbers of statistical units described with one EDI-link. If EDI -- Pilot 2 is successful, we will, following Pilot 2, in 1997 aim at 25,000 units to be approached with this instrument, partly through the bookkeeping bureaus.

The revenue of Pilot 2, if successful, will primarily be a relief of the respondents' burden. Productivity gains will not be that large. In the first place, all kinds of activities remain. Not every respondent will participate, data will still have to be checked, etc. Secondly, new activities arise in the form of a growing help-desk and a field-service that will not only have to cope with bookkeeping problems but also with technical automation problems.

Controlling Pilot 2 -- The Meta-System

Eventually, Statistics Netherlands aims to reach several thousands respondents. This, of course, will require a control system to deal with the production of the appropriate questionnaire, sending it to the respondent,checking the (timely?) response, checking and storing the incoming data, and controlling possible feedback, etc. This means that a lot of information -- meta-information -- on the respondents has to be kept updated.

Another part of the meta-information deals with the contents of the combined questionnaire. As an example we will focus on that part.

In order to construct the combined questionnaire, we need to coordinate the approach of the different statistics aimed at the financial records among each other and, also, with the bookkeeping practices of the respondents. Of course, the latter already happened before, but with EDI it will become more explicit. This needed some negotiation. It is clear that with EDI up and running, much of the former autonomy of the individual statistics, especially regarding their questionnaire, disappears.

The module containing the translator gives us better opportunities for supplying meta-information to the respondent than before. There are the usual on-line help-functions. By means of hypertext the explanations are linked. For the help-desk and for the field service probably a more detailed system of help-functions and explanations will be set up. The system not only contains cross-linkages, but also simple computational rules so that, for instance, totals can be computed.

To that end, a set of variables was laid down in a database, with names, questions texts, explanations and, if necessary, computational relations with other variables. From this database, variables, question-texts, explanations, etc., are selected and combined to construct questionnaires.

Respondents are classified into clusters by size, branch of activity, and type of financial records kept. Sometimes sale records are kept by the enterprise, itself, but the yearly balance sheets are set up by a bookkeeping bureau. For that statistical unit, the total of the information needed will have to be collected by two different questionnaires directed towards two different reporting units. Each cluster gets its own combined questionnaire.

The Changing Role of Meta-Information

In this way, a large set of meta-information on concepts emerges. This meta-information controls the process of data collection. A question aimed at the financial records can only get there through the central database of variables. When entering the variable, the relation with the rest of the contents will have to be made clear. It has to fit in.

In the first place, we now see that the character of meta-information has changed. In most of the literature we often find meta-information as a mere descriptive piece of information, only available if the statistician has found the time to set it up, mostly after he has produced his statistics, for the benefit of the user. If later on the statistician diverges from his earlier meta-information, there is nothing to stop him and nothing that guarantees that the meta-information will be adapted.

Here, we find a piece of meta-information that has to be set up before the production process starts. The statistician cannot but use the meta-information system. The meta-information has become a tool in the production process. From being descriptive it has come to be prescriptive. Earlier we saw the same thing happening with data collection among households through BLAISE.

This, however, has further reaching consequences. We can now go back to the first sections of this paper. There we spoke of the extra demands put to Statistics Netherlands. One of them was less respondents' burden. That was the first goal of EDI -- Pilot 2. But we also see here how the technology push gives us some opportunities to answer another demand -- namely, that for more coherence. It goes without saying that the way EDI is implemented, here, will lead to a larger extent of statistical (conceptual) coordination. We mentioned the power of the meta-system and we also see that within EDI a number of statistics are combined that were earlier produced in separate, independent processes. It is also worth noting that this growth in statistical coordination is not reached by an increase in central directives, but as a side-product of the tools used in the production process. We do not think that all the problems of the coherence of our end-product -- that is, all the problems of statistical coordination -- can be solved by devising the proper tool. We do think, however, that further improvements can be made in this field by applying the possibilities of the technology push in the right way. ■

Chapter 3

PERQS (Personalized Electronic Reporting Questionnaire System)

Janet Sear, Statistics Canada

Abstract

In an effort to reduce response burden and data capture and edit costs, the Retail Trade Section of Statistics Canada has developed a Windows-based, diskette version of their Annual Chain and Department Store Survey Questionnaire. The system allows respondents to download information directly onto diskettes from their own spreadsheets or databases (record layout is pre-specified by the Retail Trade Section). Interactive edits are built into the system alerting respondents to possible problems. Respondents can enter comments explaining unusual circumstances, helping to reduce the need for post-collection follow-up.

A pilot test was conducted in 1994 and met with very positive response. The number of companies using PERQS was expanded for the 1995 survey reference year.

Ms. Sear discussed PERQS in relationship to many of the topics that are listed under Productivity; Systems Development; and Organizational and Management Issues. She also discussed the results of the pilot test. As well, a working version of PERQS will be demonstrated as a software exhibit.

PERQS (Personalized Electronic Reporting Questionnaire System)

Janet Sear, Statistics Canada

Introduction

Response burden has long been a major concern of Statistics Canada and its survey respondents. At the same time, statistical agencies face demands from users for increasingly detailed data (e.g., by geography, by industry, and by commodity). The challenge faced is to find innovative ways of balancing these contradictory goals.

The use of administrative data records is one solution to this dilemma, but so, too, is the use of electronic data collection. As computer technology evolves and spreads throughout the business community, opportunities arise to reduce response burden and still collect detailed information.

PERQS

One such effort in the direction of electronic data collection has been the development by the Retail Trade Section of Statistics Canada, of a WINDOWS based electronic questionnaire for its annual retail trade survey. Known as PERQS (Personalized Electronic Reporting Questionnaire System), this collection vehicle offers a number of features designed to help reduce respondents' efforts while still collecting detailed information.

Features

PERQS is a WINDOWS based FOXPRO application. It provides an electronic questionnaire with interactive editing and import and export features. Developed by Statistics Canada, PERQS is offered at no cost to respondents of the survey. Each interested respondent is sent a set of diskettes (two "systems" diskettes and one "data" diskette). To aid in questionnaire completion and allow for year-over-year editing, non-imputed company specific information is pre-loaded onto the "data" diskette. All outgoing and incoming data are encrypted as well as password protected. On-line Help is provided as well as a user guide and a 1-800 Help Line telephone number.

Annual Retail Trade Survey

The Retail Trade Section of Statistics Canada conducts an annual survey of retail trade in Canada. At the company level, respondents are asked to report some basic revenue and expense items; cost of goods sold; class of customer; and kind of business ("Part A" of the questionnaire). A provincial breakdown of operating revenue; salaries, wages and benefits; and cost of goods sold is also requested. In addition, all retail chains [1] and department stores are required to complete a location-based supplement ("Part B"). For ***each*** location, they are requested to report operating revenue, floor area, distribution of sales by class of customer, industrial classification (SIC code), address, and host department store and/or shopping

center name. As a consequence, companies with a large number of locations are subject to an onerous response burden. In the past, they have been willing to complete the questionnaire, mainly because they are users of the data and find the detail provided very useful.

Because of the heavy response burden imposed by this survey and because it was strongly suspected that most of the requested location data was already being stored electronically, it was decided that this survey was a perfect trial application for an electronic questionnaire.

Development

Development of PERQS began in the fall of 1994. The primary goal was to explore the possibilities of electronic reporting as a means of reducing response burden. Improved quality, timeliness and perhaps response rates were also anticipated. The potential for reduced costs was also a longer-term consideration.

The approach was to start modestly (a pilot test of 25 large chain and department stores) and expand slowly if successful. Development took place within Industry Division led by a survey manager and a systems analyst and programmer.

For the pilot, a list of potential participants was drawn and each company was called and questioned as to willingness to participate or at least to have a demonstration of the system before deciding. Twenty-five of the first 27 companies called agreed to either a demonstration or to have the diskettes sent directly to them. The two that did not participate did not have WINDOWS. Initially there was some confusion as to what exactly a diskette questionnaire was, (a picture, or in this case a demonstration, was worth a thousand words). Before seeing the demonstration, some company contacts, with minimal computer experience, expressed concern as to the possible complexity of the system. Their managers, on the other hand, were quite eager to participate. Some respondents also expressed concern as to whether the system would be accessing their data files directly. Some stated that it was about time that Statistics Canada developed such a product!

Pilot Outcome

In the end, all 25 companies (representing over 25 percent of the total number of chain and department store locations operating in Canada) agreed to participate in the pilot. Their reaction was very positive. All participants commented on a reduced workload and found the interactive edits a helpful feature. All found PERQS to be very user friendly. Some even stated that it was a lot of fun! Completed diskettes were returned by 23 of the 25 participants. One company went bankrupt during the collection period, the other changed ownership and was going through a restructuring. All PERQS reporters expressed the desire to continue with diskette reporting.

The one-on-one contact with respondents brought about by PERQS promotion, provided some beneficial results. In a number of cases, the meetings resulted in contact with a higher level employee than would otherwise happen with a paper questionnaire. Clarification of survey concepts occurred; marketing opportunities arose; goodwill was generated. Managers and their staff were appreciative of our concern about response burden and efforts to try and reduce it. One manager stated that it was the first time they had ever met with Statistics Canada and ended up with less work to do. They asked if they would be able to report electronically for other surveys for Statistics Canada as well as other government

departments). They acknowledged seeing cost benefits for themselves as well as Statistics Canada. One company expressed the strong desire to send the equivalent of one electronic report containing all government required data.

Respondents were encouraged to participate in the future development of PERQS. They were asked to suggest changes or improvements.

More Features

One of the most popular features of PERQS is the ability to import location data directly into PERQS from a choice of spreadsheets or database files. The record layout of the import file is pre-specified by Statistics Canada. If the respondent does not already store their location data in a spreadsheet or database file, PERQS will create one for them in the required record layout and containing all of the location information (except for operating revenue) as PERQS "knows" it at the time of file creation. This file can then be "exported" from PERQS; updated with current period information; and then imported back into PERQS to complete the location part of the survey. Respondents are not forced to enter values for cells that require no change, blanks are defaulted to whatever data PERQS currently has stored. Respondents, if they so desire, can maintain this spreadsheet throughout the year making any necessary updates as they occur. As the respondent moves through PERQS answering the questions, PERQS will pre-fill (link) any related unanswered questions with data from completed (saved) questions. For example, for chain and department stores, PERQS will calculate provincial totals of operating revenue using the location data (from "Part B") and pre-fill the appropriate "Part A" question. Interactive edits (inter-year and inter-field) are included in the system, alerting respondents to possible problems and allowing them to immediately correct any "errors" or enter comments or explanations. Data capture errors and costs are thereby reduced as well as the need for post-collection follow-ups.

Features which would benefit the respondent (filters, sorting options, etc.) were also incorporated into PERQS to give the respondent further incentive to use the system.

Phase II

The number of companies using PERQS for the annual retail trade survey was greatly increased for the 1995 reference year. Participation by retail chains was expanded to 185 companies (representing over 60 percent of the total number of chain and department store locations). As well, PERQS was offered to a test group of large "independent" retailers [2]. PERQS "participation rates [3]" were significantly different between the two groups. Approximately 85 percent of contacted chain and department store organizations were willing and able to try electronic reporting. In contrast, 55 percent of contacted independents ended up being sent a PERQS package. The unavailability of WINDOWS, or even a computer, was a significant factor in the independent group. Thirty percent of those independents who said they were interested in trying an electronic questionnaire did not have WINDOWS and therefore had to be excluded from using PERQS.

For the most part, PERQS has so far been offered to large companies. We expect that the PERQS "participation rate" will diminish as the size and complexity of the companies diminishes. It is, however, the large, complex companies that will benefit the most from our electronic questionnaire and it is they that contribute the most to our retail estimates.

Electronic Transmission

Currently, it is thought that the sending and receiving of diskettes will outpace the transmission of electronic files (via a modem or the Internet), at least for the next few years. When asked if they would want to report electronically, some participants in the pilot said they might consider it but that a modem, although one exists somewhere in the company, was not readily available to the person completing PERQS. Some asked if we would consider buying them a modem if we wanted the data transmitted electronically. At the time of the pilot, the Internet was still quite new to most respondents. Many were concerned about security issues associated with sending confidential data.

Future Plans and Considerations

Because of the success of our pilot, the use of electronic reporting for other surveys is being implemented. As well, for some very large enterprises, the possibility of collecting consolidated information is being examined.

Conclusion

Electronic questionnaires are a viable option in data collection. The investment in development will yield multiple benefits (improved respondent relations; reduced capture and collection costs). A certain number of lessons learnt are worth pointing out. Whatever application is developed to meet survey needs should be ***very*** user friendly. It should be simple and, from the respondent's perspective, it should not appear to be intimidating. Although instructions should be provided, it should not be assumed that the respondent will take the time to read them. The computer literacy of the respondent should not be overestimated. Some respondents will be very knowledgeable, in fact some may be too knowledgeable. A significant portion, however, will have limited computer experience.

In today's market place, as companies struggle to reduce costs to stay competitive, they are challenging data collectors with finding more efficient means of gathering information. Many, or at least the largest, are discovering the cost saving tools that modern technology has to offer. As respondents, as well as clients, become aware of what is possible electronically, they increasingly demand that statistical agencies accommodate their desire and ability to use modern developments in information technology to provide and receive data in a less costly and time consuming manner.

In an era where it is not only the private sector which must find ways to reduce costs, today's advances in information technology provide vehicles which the public sector can also utilize in its efforts to reduce spending. Demand for the use of modern information collection technology will come both from outside as well as within the public sector.

Footnotes

[1] Companies operating four or more retail locations within the same kind of business.

[2] Companies operating less than four retail locations.

[3] The number of contacted companies willing and able to use PERQS, divided by the total number contacted as to PERQS participation. ■

Chapter 3

Electronic Data Collection: The Virginia Uniform Reporting System

Anne Rhodes and Kishau Smith
Virginia Commonwealth University

Peter Goldstein, NI-STAR Data Systems

Abstract

The Virginia Commonwealth University Survey Research Laboratory has implemented a system to collect data from organizations in Virginia who serve clients with HIV and/or AIDS. This data, which includes demographic characteristics and information on services provided to clients, is used for decision making at the local, state and federal level. In an effort to improve reporting time and data quality, a system of automated data entry and error checking has been designed. This system allows for "real time" transmission of data and correction of errors. Providers fax forms in after a client encounter to a central computer where the form is checked against a database quality assurance program. The database application produces a data verification report, which also details any errors found, and which is automatically faxed back to the provider site. Providers check the data, correct any errors, and fax the report back to the central computer. Improvements have been found in the quality of the data and in response time from providers. This type of system has applications in survey work where there is a need for data confirmation and corrections from respondents.

Electronic Data Collection: The Virginia Uniform Reporting System

Anne Rhodes and Kishau Smith
Virginia Commonwealth University

Peter Goldstein, NI-STAR Data Systems

Background

In 1991, Virginia began receiving Title II funds from the Ryan White Care Act to provide services to HIV affected and infected individuals. The Virginia Department of Health subcontracted with the Virginia Commonwealth University Survey Research Laboratory to manage the collection of client level data from providers receiving Title II funds. This data collection, the Virginia Uniform Reporting System, is modeled after Federal reporting guidelines and consists of demographic and medical characteristics of clients, as well as the types and number of services provided. Service providers report on each client who is served with Title II funding. These data are unduplicated at the provider, regional, and state levels to provide quarterly and annual data on numbers of clients and services provided.

At the Federal level, Title II funds are administered by the Health Resources Services Administration (HRSA) of the Department of Health and Human Resources. HRSA currently does not require quarterly client-level data reporting from Title II providers, but does require an annual count of services and clients. In 1994, HRSA awarded a number of contracts to demonstrate the usefulness of client level data for the evaluation of HIV/AIDS service programs. Virginia was one of seven sites to receive this contract, which called for electronic data collection from all service providers over a three year period. This electronic data collection was to replace the old system of paper form submission that had been used in Virginia from 1991 to 1994. The next section describes the old system and the problems that were encountered with it.

Uniform Reporting System Prior to 1995

Under the previous system data collection was done quarterly in five regions of Virginia. Each region has its own consortium representative who manages fund allocation and usage for all providers in the region. Providers filled out their forms and mailed them to the consortium representative at the end of each quarter. The consortium representative had the responsibility to do initial error checking on the forms and send them on the Survey Research Laboratory (SRL).

Once forms were received at the SRL data entry was completed and a report was generated detailing errors found in the forms. This report was sent back to the consortium representatives who forwarded them to the providers. The process of sending forms, processing, and report writing generally took

about 4 to 6 six weeks to complete. As a result, providers were being asked to correct data forms that they had filled out 2 months ago. This led to very few corrections being made and a high rate of "unknown," or blank, data in the system.

Another problem with the old system was that it was resource intensive, both for providers and personnel at the SRL. Providers often filled out all forms at the end of the quarter, with some providers submitting 500 or more forms. At the SRL, all received forms had to be logged, data entered, and error reports typed. With approximately 60 providers submitting forms from around the state, data entry and reporting time was considerable.

The Faxable Forms System: Concept

In 1994, the Virginia Department of Health, in conjunction with the SRL, received funding from HRSA to switch the Title II data reporting to an electronic system over a three year period. Where feasible, providers were given the option of using computer software to enter client intakes and encounters. Where providers could not or did not want to use computer software, a new type of electronic submission was implemented.

The faxable forms system was designed to allow for real-time data collection and feedback. Providers fill out the data reporting forms immediately following a client contact and fax them to a dedicated computer at the SRL. This computer system reads the forms, checks for errors and immediately generates a report back to the provider, detailing all data received and asking for corrections on any errors. The provider checks the report, makes any necessary corrections and faxes the report back to the SRL computer, where it is automatically printed out. SRL personnel make any corrections to the database and file the report.

This system significantly decreases the amount of personnel time needed to check and enter the data forms. It also allows providers to check and correct data immediately following a client visit, when the client file should still be easily accessible. Also, this system makes data reporting for the provider more manageable as it does not require the provider to take a large amount of time at the end of the quarter to fill out all the forms for the previous three months.

Design of the System

The faxable forms system was designed by the SRL, with assistance from NI-STAR Data Systems, who did the initial programming and editing of the system. The system uses a scanning software program, Teleform, in conjunction with a FoxPro system, which does extensive error checking and houses the final dataset.

Teleform is used to design and print all data reporting forms. The forms are keyed to each provider so that when forms are faxed in, Teleform recognizes the provider number by a box in the upper left hand corner of the form. Teleform interprets each form, recognizing the type of form and exporting the data from the form into a FoxPro database file.

The FoxPro system reads in the database file and runs an error checking program which looks for missing or inconsistent data. FoxPro saves the data to a database and writes the errors to another data-

base, which is merged with the confirmation report form. This form is generated by Teleform and automatically faxed out the provider. The FoxPro system incorporates a data entry module which SRL personnel use to make corrections in the database.

Other features of the system include:

- a log report printed by FoxPro every day, listing all forms received and any errors on those forms;
- a fax log which details each fax sent and any problems with fax transmission;
- a Teleform Verifier log which stores an image of each form received along with the status of the form; and
- an exception report generated by FoxPro which list all outstanding errors for each provider.

Implementation and Results

This system was implemented with providers in the Central Virginia region in 1995. Approximately 15 providers began using the faxable forms in April 1995. SRL personnel did on-site training at each provider site, demonstrating how to fill out the forms, fax them in, and interpret and return the confirmation report.

This initial test of the system was successful. Some minor changes were made to the system, but providers indicated overwhelming approval for the real-time error checking and reporting. Some providers who had older fax machines could not fax their forms in, as the forms became too distorted in the machine. These providers mailed their forms to the SRL, where they were scanned in and reports were faxed back to the providers. This type of submission, while not ideal, still allows for better, and more timely, data confirmation than under the old system.

The system was implemented statewide in 1996. Currently, approximately half of the providers are using faxable forms, with the rest using computer software. Improvements in the quality of the data have been dramatic over the first twelve months of the project. The number of outstanding errors has decreased by about 50 percent and providers have indicated that they feel more responsible for the data as they are constantly receiving feedback on it.

Additional Considerations

The quality of the data reporting forms at the provider site has become an issue. Forms must be clean copies which maintain the exact state of the original or Teleform will be unable to read them. SRL personnel have had to retrain some providers who continually submitted forms that were not clean.

As the system has gone statewide, the number of forms being received on a daily basis has increased to the point where the system needs to be expanded to more than one computer. Currently, the system is being converted to work on a network so that maintenance of the system can be performed on a computer separate from the one where forms are received, checked, and reports faxed out.

While personnel time is decreasing on the paper management side of the project, there has been increased staff time in the development and maintenance of the system. This is expected to decrease as the system becomes more automated over time. Plans for future automation include an automatic quarterly report to providers, which summarizes the number of clients seen and services provided. It is expected that this type of data feedback will increase data quality by providing organizations with data that can be used for program planning and evaluation. As they begin to utilize the data in this manner, they will have a greater stake in producing accurate and timely reports. ■

Chapter 4

Case Studies -- I

Chair: Leda Kydoniefs, Bureau of Labor Statistics

Glen Ferri ◆ Tom Ondra

Richard J. Bennof ◆ M. Marge Machen ◆ Ronald L. Meeks

Toward A Unified System of Editing International Data

Glen Ferri and Tom Ondra, U.S. Bureau of the Census

4
Chapter

Abstract

CONCOR is the editing component of the Integrated Microcomputer Processing System (IMPS). This module was originally a stand-alone procedural language used to identify and/or correct invalid or inconsistent information. The microcomputer provided a DOS-based platform to integrate all the major tasks of survey and census data processing which IMPS accomplished. CONCOR and CENTRY, the IMPS data entry module, have been combined to provide interactive editing. IMPS and CONCOR are being redesigned to run under Windows. CONCOR will move from a procedural language used by programmers toward an edit specifier used by subject matter specialists. What parts are to be carried over from the old version and what parts need to be re-engineered?

The survey questionnaire diskette is designed from data entry into 18 tables containing Federal obligations by intramural and extramural performers, by fields of science and engineering, and by geographic distribution. ■

Chapter 4

Data Editing Software for NSF Surveys

Richard J. Bennof, M. Marge Machen, and Ronald L. Meeks, National Science Foundation

Abstract

The National Science Foundation (NSF) uses automated data entry programs to collect research and development (R&D) data for its annual national surveys. Three of these programs are the topic of this paper:

- academic expenditures program used to collect data for the Academic Science and Engineering R&D Expenditures Survey,
- FSS program used to collect data for the Federal Support to Universities, Colleges, and Nonprofit Institutions Survey, and
- FEDFUNDS program used to collect data from the Federal Funds for R&D Survey.

This presentation will describe each of these studies and demonstrate how each data entry program is suited to its users.

Data Editing Software for NSF Surveys

Richard J. Bennof, M. Marge Machen, and Ronald L. Meeks,
National Science Foundation

The National Science Foundation's (NSF) Division of Science Resources Studies (SRS) has a mission to produce and disseminate high quality data and analyses related to science, engineering, and technology. SRS is responsible for conducting surveys on a wide variety of areas. Three of those surveys that are outlined below are managed in SRS' Research and Development Statistics Program.

NSF uses automated data entry programs to collect research and development (R&D) data for its annual national surveys. Three of these programs are the topic of this paper:

- ASQ program used to collect data for the Academic Science and Engineering R&D Expenditures Survey;
- FSS program used to collect data for the Federal Support to Universities, Colleges, and Nonprofit Institutions Survey;and
- FEDFUNDS program used to collect data from the Federal Funds for R&D Survey.

ASQ

The Academic R&D Expenditures Automatic Survey Questionnaire (ASQ) is a user-friendly, PC-based program that requires at least 384K of RAM. The ASQ program provides help at all points of data entry and allows the user to make choices by selecting menus. Automatic editing will check the ASQ for arithmetic errors or inconsistencies; and, such problems will be pointed out to the respondent. Respondents have the opportunity to manually correct them or allow the ASQ to total automatically. The program allows the user to enter data in any order and does not have to be completed at the same time. This feature allows the user to compile information at different times. When the data have been entered for any item, the user will be asked whether they wish to edit or not. When all data have been entered and edited, the user is prompted to select the trend checking option for comparing the previous year's data with current year's data to identify major increases or declines. All data are stored back on the ASQ diskette as they are entered.

The ASQ program will let the user print out a facsimile of the institution's questionnaire response, both for the current and previous years.

The most recently completed survey collected data from over 500 institutions of higher education in the United States and Outlying Areas and 18 university-affiliated Federally Funded Research and Development Centers (FFRDCs).

FSS

FSS includes a PC-based program (written in Visual Basic for DOS) used as the survey instrument by many agencies in reporting its data. Since the program requires at least 490K of conventional memory, 8M of RAM, and about 2M of hard disk space, it must be loaded on the PC's hard disk in order to work. That's because there are 6,538 specific institutions eligible to provide data for in FSS, and each institution's name, code, and geographic location are stored in FSS. The program also has the capability of adding new institutions. A "CHKMEM" software feature, which allows the program to check the hardware to see if expanded memory is running, has a "soft-boot" program which allows the participant to run the FSS program without having to reconfigure the hardware. FSS is a user-friendly, menu-driven program with extensive built-in instructions for users.

The data collected from Federal agencies includes:

- total program support of both science and engineering (S&E) and non-science and engineering (non-S&E) activities to academic institutions;
- total S&E support to FFRDCs administered by academic institutions; and
- R&D and R&D plant support to nonprofit institutions and FFRDCs administered by nonprofit institutions.

The contractor has held a series of annual hands-on respondent workshops on FSS in which participants were generally enthusiastic about working with the FSS PC survey disk. FSS contains data edit checks, a function to search for an institution by name or institutional code, a convenient lookup capability for field of science and engineering detail (respondents do not have to search through a hard copy of science and engineering taxonomy), and a function for data trend analysis. An "Import" feature of the FSS PC survey disk allows agencies with a large volume of data downloaded in an already-formatted database to import that data directly into the FSS PC survey database. The data are completely edited during the import procedure. Respondents can print out a summary report which displays the total obligations of each type of institution with field of science and engineering totals, a detailed report by individual institution, and a trend report which lists individual institutions which have a large increase/decrease in obligations between the current and prior fiscal year. The prior year totals can also be displayed in summary and detailed reports.

FEDFUNDS

FEDFUNDS is a user-friendly, menu-driven system that can be used on any IBM compatible microcomputer. The entire program is written in visual basic and uses "forms" to organize and group like data for display on the monitor. The entire program is stored on a diskette to allow survey respondents to enter and edit data directly from their microcomputers. The program contains 47 "forms" that are displayed separately (each is a separate screen). The FEDFUNDS program displays a "form" containing data items form a particular questionnaire table and allows the user to enter or modify the data on display. Combined, the "forms" contain all of the data items to be maintained by the survey.

This disk based system is also equipped with extensive built-in instructions and help facilities to aid the user in completing the approximately 2,000 data fields and narrative statements. Recent efforts

to make the system more efficient included redesigning the data entry questionnaire program to build in more internal data checks (e.g., within table and cross table checks, trend analysis function). When the program finds discrepancies during data checking, an error message is displayed and identifies the exact items in question for the prompt attention of the respondent. Further advances incorporated in the data entry program include automatic totals of fields without respondent's intervention. The program also allows the respondent to combine data from several sources into a consolidated agency report.

The survey questionnaire diskette is designed for data entry into 18 tables containing Federal obligations by intramural and extramural performers, by fields of science and engineering, and by geographic distribution. The respondents are instructed to complete the tables in order to avoid table checking errors, since, for example, the detailed data requested on one table must add to the aggregated data on another table.

The respondent also has the option to run a trend report that will produce a list of all "large" differences in data from the prior survey submission and the current survey input.

The participants at the latest NSF data entry demonstration workshop for FEDFUNDS users were enthusiastic about working with the FEDFUNDS program. Several commented on the automatic totaling of subtotals and grand totals and asked if this feature could be further enhanced by eliminating the need for the survey respondent to move through the subtotal and grand total fields (currently a user must move the cursor to a total field before the program completes the automatic total). The respondents felt that since these amounts are already computed, then they should appear as the detailed level is entered. Also, it may soon be possible to transmit the FEDFUNDS program electronically to respondents. Respondents can now electronically send completed survey results via Internet.

Additional Information

For more information about the three data entry programs (and the surveys associated with them) described above, please contact the appropriate author on (703) 306-1772 or via Internet at *mmachen@nsf.gov, rbennof@nsf.gov*, or *rmeeks@nsf.gov.* ■

Chapter 5

Censuses

Chair: Clyde Tucker, Bureau of Labor Statistics

Olivia Blum ◆ Eliahu Ben-Moshe

Jan Thomas ◆ David Thorogood

Frank van de Pol ◆ Bert Diederen

Chapter 5

Automated Record Linkage and Editing: Essential Supporting Components in Data Capture Process

Olivia Blum and Eliahu Ben-Moshe,
Central Bureau of Statistics, Israel

Abstract

The data capture process of the Israeli 1995 Census of population and housing is based on an optical character recognition (OCR) technology. The data capture system has been designed and developed while bearing in mind short run targets and long-run goals. The short-run targets are concerned with the data capture itself. These include getting an accurate and reliable file in a relatively short period of time; decreasing the subjective, human component in census data capture; and simplifying control and quality assurance processes. The long-run goals address future needs and uses. These include shifting from microediting to macroediting, to avoid overediting and to make editing more efficient; allowing for reprocessing the data starting with the raw, input data file; and linking census records with previous census files and administrative records. The use of macroediting and the ability to return to a basic file permit recreation of the main census file on a different basis, as future needs become clear. An accurate and reliable file, with the exact values given by the respondents, is also necessary for reprocessing. The search for additional support for the data capture process was motivated by these concerns and the limitations of OCR.

Although automatic processes are embedded in an OCR system, it lacks as a substitute for a human eye-brain mechanism in two main respects. First, the mathematical function used in the recognition process does not specify the whole scope of handwriting styles. As a result, the reliability of the OCR values varies.

Second, enumerators' errors hamper the automatic definition of the process units. Correcting these errors involves altering values in ways beyond the OCR capabilities. Record linkage has been incorporated into the data capture process as external support for OCR in determining values in the questionnaires' fields. This benefits the data capture process and enhances the final file.

Automated Record Linkage and Editing: Essential Supporting Components in Data Capture Process

Olivia Blum and Eliahu Ben-Moshe,
Central Bureau of Statistics, Israel

The objective of this paper is to show that implementing advanced technology in the data capture process makes possible the redefining of goals and facilitates and simplifies data capture, thereby improving the quality of the product. It also allows for addressing technical problems and ideological debates that accompany or arise from the traditional process. In this paper we explain the advantages of the Israeli data capture process, comparing it with the traditional one; list the unique features of the system that overcome the limitations of the new technology; describe briefly the structure of the data capture process; and give first outcomes of the outgoing quality checks.

Planning the data capture process for the 1995 census of population and housing in Israel was based on two principal caveats:

- To anchor the process in an optical scanning and character recognition technology, which allows for simultaneous presentation of digital images of the questionnaires, and their fields' ASCII values.

- To produce a structured raw data file, that presents, as closely as possible, the actual responses of the respondents.

It is the combination of goals and technology that opens new options for planning and developing processes. Moreover, along the way, changes of needs and means allow for a productive interaction between planners and programmers: needs trigger further technological development and vice versa; technological developments present a wider range of potential solutions for existing problems, and therefore stimulate and create needs. The Israeli data capture system that has emerged from such interaction presents preferable alternatives to several components in the traditional data capture process.

Background

For the purposes of this paper, the traditional process is defined to include the following steps: preparing documents, editing enumerator's and respondents' fields, coding alphabetical fields, keypunching, and shredding the paper questionnaires.

This process is work intensive and time consuming, suffers from high workers' turnover and lacks in its product's quality. Furthermore, the end result of the process is a data file whose values come from

three different sources: actual respondents' answers, **edited answers, and errors added throughout the data capture process**.

The 1995 Israeli optical data entry system was designed to overcome the obstacles of the traditional process.

- ❒ The need for document preparation is eliminated because the scanning process captures the data as they appear on the questionnaire. In addition, each page carries a unique identifying number which makes the process indifferent to the order in which pages are captured.

- ❒ The ability to transform the data from a written paper to a digital image at the beginning of the data capture process implies that the paper questionnaire can be discarded immediately after scanning, while still having an accurate and lasting optical archive.

- ❒ The human-machine interface makes it possible to see simultaneously the image of the write-in responses and the corresponding values residing in the database which facilitates the comparison of values. This contributes enormously to the ability to implement data capture and quality control tasks.

- ❒ As for the subjective component, it is significantly reduced, since the first ASCII values are given by the optical character reader. Automatic and computer-assisted processes can be executed as soon as the scanning stage has been completed, with no previous human intervention.

- ❒ The working environment is comfortable and easy to operate in. Basic actions involved in handling a questionnaire, such as turning pages, browsing, moving from one record to the other are computerized and, therefore, easier and much faster.

- ❒ Beyond the technical-practical advantages of the optical system, the target file is unique in its character. It is a raw file in which the variables' values reflect the respondent's answers **as they are** on the questionnaire. Respondents' answers are not edited, even if logical contradictions are embedded in them. Moreover, no editing is done, even in cases where the respondent has marked two closed categories instead of one. When two answers make no sense (in yes/no categories) and when more than two answers has been marked, the field is emptied but the audit trail carries a status indicating the number of answers given for the question. This procedure ensures that no information is lost, regardless of its quality.

This first census raw file addresses several important issues:

- ◆ Avoiding over-editing during data capture process, by minimizing micro-editing within a record.

- ◆ Opening the option to reconstruct more than one census file, by using different editing methods.

- ◆ Isolating the data capture process from the debate over the issue whether or not statistical agencies should release an unedited data file.

From an Optical Reader to an Optical Data-Capture System

In spite of the visible merits of the optical character recognition (OCR) facility, it is not a full substitute for a human eye-brain mechanism. Moreover, the system has been developed under restricted budget; thus, the full potential capabilities have not been realized.

Consequently, supporting components have been added to the optical reader to transform it into an optical data capture system. These components have come to specificly address the following **limitations of the optical reader**:

- The recognition process does not cover the whole scope of hand-written styles. Therefore, the reliability of the OCR-suggested values is variable, meaning that relying on the OCR as the only source of identification can be error prone.

  In order to utilize the process and channel resources to where they are needed, each OCR value is accompanied by a status, indicating its level of confidence (Super-Sure, Sure, Doubt or Fail). These statuses determine the nature of the treatment needed in the following stage.

- The census processing units are records of an individual, an household, or an enumeration area (EA). A record is defined once it is exclusive, exhaustive and unique, meaning that it contains ALL the data of only ONE unit and that there are no duplicate records.

  The need to define a record arises when the enumerator's fields on the questionnaire carry wrong values. Hence, defining a processing unit implies altering values in the enumerator's fields. This task is beyond the capabilities of the OCR.

- The optical reader has been designed to identify bar codes, preprinted numbers, handwritten numbers and marks (Xs). Because of the limited potential market for recognition of Hebrew letters -- and, therefore, high development costs -- the OCR does not include alphabetic character recognition.

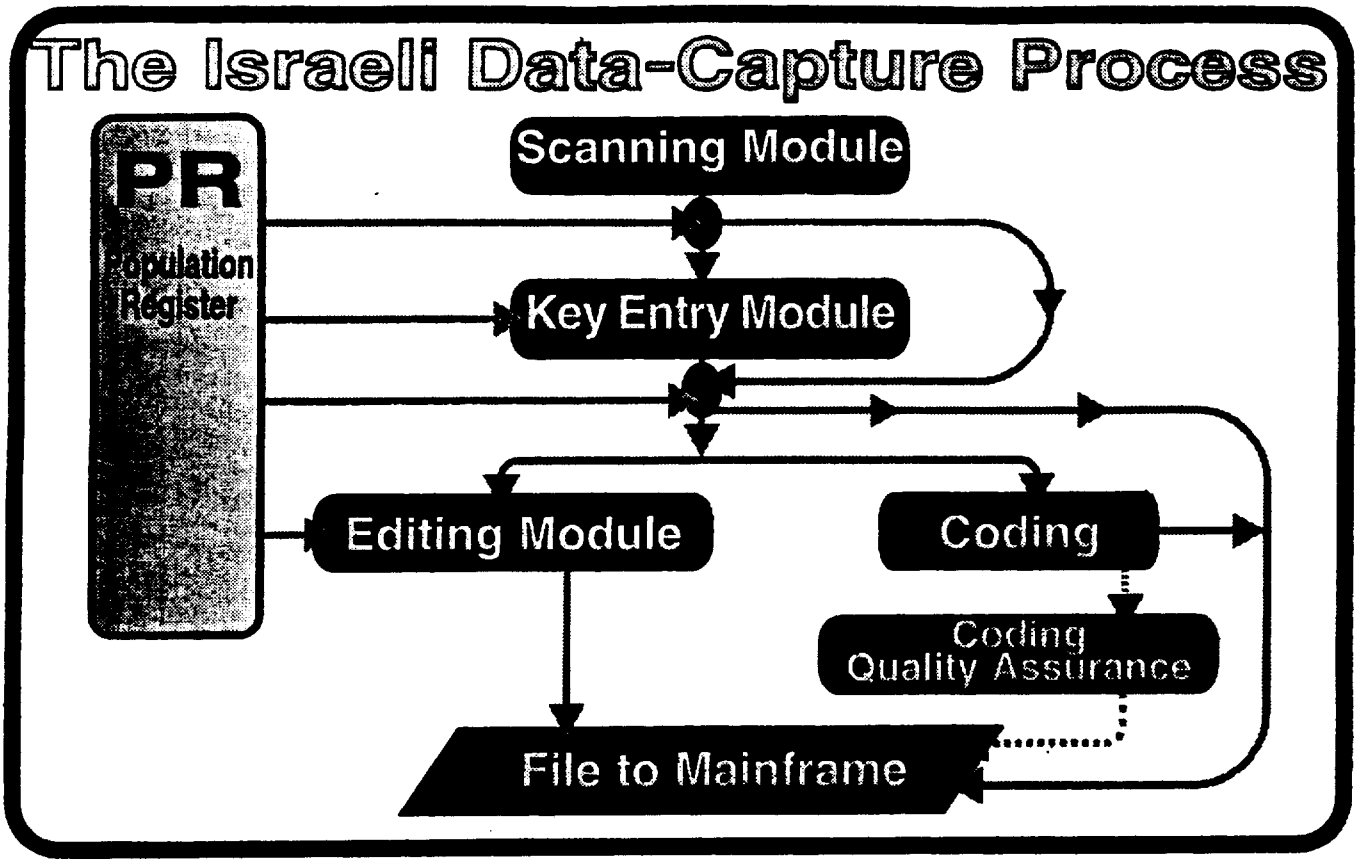
These limitations, along with the requirement to have a raw file, dictate a non-conventional data capture process (see next page).

This process has several unique interrelated features:

- Each component in the process supports and relies on adjacent components.

- There are no homogeneous stages, in a sense that each one includes operations that traditionally belong to other stages. For example, throughout key entry procedure, edit checks are performed and the individual processing units are defined by an automatic record linkage with the population register. Both tasks are traditionally editing operations, but, in this instance, they support the key entry stage.

- In addition to the intertwined stages, the population register of Israel has been added to serve as an external support throughout the process.



- The definition of editing during the data capture process is altered altogether; the conventional editing tasks in which data are changed, deleted or added are postponed and take place in the central computer, on a macro level. Editing within the data capture process consists of three assignments:

  - Correcting enumerators' errors, in order to define those processing units that have not been defined automatically.

  - Corroborating or correcting values in respondents' fields, when captured values fall out of an expected range or when they produce logical contradictions; and

  - Coding residual, open categories, where answers are in not optically recognized, alphabetic characters.

The description of the data capture system is beyond the scope of this paper; however, a description of **the flow of data** throughout the process is needed. The first step in transferring the written paper into a magnetic medium is done in the **scanning module**. It is not a sheer scanning procedure but also includes Form Drop Off (FDO), where all preprinted non-unique texts, numbers and graphics are removed from the image of the questionnaire; Optical Character Recognition (OCR), which is performed on marks and numbers written on the questionnaire, compression of the data and inserting it into the database. At this point, the paper questionnaire is repacked and stored.

In the **key-entry module**, the system looks for two sources of identification in order to confirm a field's value. The four available sources in the order of application are: OCR, the Population Register (PR), 1st key-entry, and 2nd key-entry. The Israeli PR contains data on nearly all Israeli citizens and permanent residents (over 95 percent of the population). Here it serves to corroborate the OCR suggested values of variables included in both records, PR and census. By so doing, human intervention is reduced significantly, since even doubtful values that are usually directed to a relatively intensive key-entry process can skip this step altogether. It also serves as an independent source for defining an individual record; once a census record is linked to the PR, it is considered as a defined reference record. This operation is taking place throughout the data capture process: in the key-entry and editing stages, before each round within each stage, and after it (key-entry is performed in, at most, two rounds while editing can be performed in up to four rounds).

The **editing module** deals with non-conventional tasks, while it omits the traditionally defined editing tasks. The editors correct errors in enumerators' fields in order to complete the definition of processing units that have not been defined automatically. They also correct or validate values assigned in previous stages. It should be noted that no editing, in its traditional form, is performed. As was explained at the beginning of the paper, respondents' answers are not changed, even if logical contradictions are embedded in them.

A parallel activity to the editing step is **coding**. There are three types of coding in the system: addresses, occupation, and economic activity. The last two are followed by a quality control process.

At the end of the data entry process, an ASCII file is formed and sent to the main frame, and the images of the questionnaires of each enumeration area, along with the ASCII values of the fields, are saved on an optical disk.

Concluding Remarks

The data capture system of the 1995 Israeli census started in February 1996 and will be completed in August 1996. Although data capture is still in process, there are two important points to note:

- The questionnaires of 1.6 million households are captured by 123 workers. They are expected to complete the task in less than 140 working days.

- The current error rate in the raw data file, as measured by the out-going quality checks, is 0.558 percent. The permissible error rate is 1 percent.

The reduction of human involvement in the data capture process and the high quality of the raw data file are the outcome of careful planning, using the advantages offered by the technological improvements, avoiding or solving problems that have plagued the traditional data capture process, and anticipating -- and, therefore, giving solutions for -- the new system-related problems. ■

Chapter 5

Editing and Imputation Research for the 2001 Census in the United Kingdom

Jan Thomas and David Thorogood,
Office for National Statistics, United Kingdom

Abstract

The United Kingdom Census Offices are working on a development programme for the 2001Census. This paper outlines the research being undertaken on editing and imputation as part of this programme.

New methodological and technical developments are being investigated to see if they offer improvements over previous systems. These include the possible application of neural computing to imputation, and the use of generalised editing software to create editing rules. Research will also address the extent to which editing and imputation processes can be integrated to reduce the occurrence of inconsistencies caused by imputation. The impact on the editing and imputation systems of possible changes to other aspects of the census procedure will also be considered.

Editing and Imputation Research for the 2001 Census in the United Kingdom

Jan Thomas and David Thorogood,
Office for National Statistics, United Kingdom

Introduction

The three UK Census Offices are working on a development programme for the next decennial population census in 2001. The Census Offices are:

- the Office for National Statistics (ONS) in England and Wales;
- General Register Office for Scotland; and
- Northern Ireland Statistics and Research Agency.

The same editing and imputation systems will be used in both England and Wales, and Scotland. Northern Ireland may adopt different processing systems.

The census has traditionally used a system of enumerator delivery and collection of forms to/from households. The possibility of asking respondents to return completed forms by post will be examined in a census test in 1997. Changes such as these may mean that different demands are made on the editing and imputation systems.

As in any data collection exercise census data will contain errors. "Tidying up" the data helps to check the validity of the entry, and ease computer processing. Editing is performed on the raw data which is received from the public. This will contain missing answers, answers which are inconsistent with others on the form, or coded answers which are outside of a pre-defined range.

Imputation aims to fill in gaps in data caused by missing answers and items rejected as invalid or inconsistent by edit checks.

The roles of editing and imputation systems are closely linked. In the 1991 Census, certain imputed items were inconsistent. Checks will be incorporated into the imputation process for 2001 to avoid this. Re-editing the data after imputation is problematic as this may lead to "looping" with data repeatedly failing the post-imputation edit. The option of closely integrating edit and imputation processing will be investigated as this may offer an efficient way of ensuring consistency.

Editing

Policy

Editing should be considered an integral part of the data collection process. In addition to the role of fixing errors, editing can also play a valuable part in gathering intelligence about the census process.

The overall editing (and imputation) policy is to make the minimum number of changes to the database, whilst ensuring that it is complete and error free (as defined).

A review has taken place of the statistical and operational requirements for editing systems and processes for 2001. The main findings were:

- the methods chosen need to be practical and statistically sound. Editing must not cause bias or distortion in the data;
- the methods must allow pre-determined Data Quality levels to be met;
- the processes need to provide a complete, consistent, comprehensive, valid dataset.

There are a number of stages at which editing can take place within the overall process. These are:

- clerical editing by manual scrutiny of the forms;
- at the Data Capture stage (in the processing office), with simple stand alone checks built into the Data Capture software;
- at the Coding stage, when certain decisions may need to be made, for example, preferences where two answers are given to a question but only one is allowed;
- a Post Capture/Coding main edit process to carry out checks within records and between records, and ensure consistency of the database; and
- within or post-imputation checking to ensure that the imputation process has not created inconsistencies.

In summary, errors are:

- ***invalid (out of range)***: relatively simple for any data capture system to spot;
- ***missing data***: i.e., no answer given. A missing code needs to be supplied to identify such cases, again at data capture; and
- ***inconsistent data***: i.e., answers to questions that conflict with one other. These can be:

  definite such as a 1 year old married person;
  less definite such as a 15 year old married person; or
  doubtful such as a 60 year old student.

Inconsistencies can occur:

- **within records** (of the type described above);
- **between person records in the same household** -- for example, relationship to person 1 ticked as husband or wife, but person 1 having ticked single; or

- **between households** -- for example, if there are two households within one building, one ticks use of bath as shared, and the other ticks use of bath as exclusive.

Editing Options

The edit options being considered for 2001 fall under the following main headings.

1991 "Edit Matrices" Approach

In the 1991 Census, the edit system checked the validity of data and performed sequence and structure checks. Invalid, missing and inconsistent items were identified for the imputation process. The editing process filled in a few missing items. The edit matrices were constructed so as to consider every possible combination of values for relevant items and to give the action, (if any) required should that combination arise, by making the least number of changes.

Simplified 1991 Edit Matrices

An assessment of the 1991 approach is underway to identify areas of excessive complexity that can be simplified. In 1991, it was found that missing items are usually genuinely missing and so could go straight to imputation (providing a quality check takes place to ensure that there are no quality assurance problems, for example, with software). Only inconsistencies need to be handled at editing stage.

Stand Alone Edits at Data Capture

Within record checks are being defined which could be carried out at the data capture stage. There is a school of thought which says that the sooner errors are detected and eliminated the better. However, it may prove more effective to eliminate all inconsistencies as one main process.

Interactive Editing

Simple logic and data validation edit checks could take place via clerical intervention, using software packages to load the main database with correct values.

Fastpath Editing

After capture editing could be carried out on "closed" questions (those covered by tick box answers) only. This would produce an earlier partial database. The "hard to code" questions such as occupation could follow later. Such an approach would only be adopted if there was a clear customer requirement for information on certain "easy to process" questions.

Selective Editing

Selective editing prioritises which fields should be edited and then applies edits to those priority fields only.

The selective editing approach is to calculate a score for each field with one or more detected errors. If the score is low it is expected that correction would have little impact on the resulting edits.

Generalised Editing Rules

Specific software, designed and developed to generate editing rules is available. Generalised software systems have great advantages when compared with *ad hoc* applications: they obviously reduce application costs, but, more importantly, they allow the correct application of given methodologies to each suitable situation. The system considers the edits for all fields simultaneously and the response reliability of each field is assessed. The explicit edits needed to edit combinations of these fields are then automatically generated, even if the relationships between the fields are very complex. Work is planned to investigate the use of these generalised systems.

Editing and Imputation Combined

Imputation needs to take place on a consistent database, but can itself cause inconsistencies. If, as a result of new approaches, fewer errors have to be dealt with at main edit time, then imputation and editing could be carried out in conjunction, with a final consistency check at the end of processing.

Imputation

Policy

The Census Offices have access to all available information at individual record level, and so are best placed to guide the imputation of missing data. The imputation system removes the need for "missing" or "not-stated" categories in statistical outputs, which can be inconsistently used and interpreted by users. It is therefore accepted that some form of imputation must be undertaken by the Census Offices prior to the production of outputs.

Imputation Options

Three options are being considered. These are:

- Donor imputation (primarily hot deck systems as in the 1991 Census);
- Neural networks; and
- Multi-level modelling (MLM).

Of these, a hot deck system or a neural network solution are the most likely to be adopted. However, the MLM approach is being investigated further to see if it can be used, in whole or in part. The boundaries between the various types of system are not always clear, particularly those between different forms of donor imputation.

It is intended that the imputation options will be trialed using a common set of test data from the 1991 Census. This should assist with the comparison of results.

It is possible that different variables might be imputed using systems of different types and/or differing levels of sophistication. For example, a complex hot deck or neural network system might be used to impute key variables (age, sex, marital condition), with other variables imputed using a simpler system.

Some imputation can be carried out by the editing system itself, in cases where only one code is possible. For example, the marital status of a one year old person can only be single.

Donor Imputation

In donor imputation methods a value is selected from a valid record (the donor) and copied to fill in the missing item(s) of another record (the recipient). Donor methods offer the benefit of imputing plausible values as they are copied from real records. However, these values are not always consistent with other parts of the recipient record. Differences between types of donor method centre mainly on how the donor is selected. These methods are outlined below. Although the most likely donor option for 2001 is a sequential hot deck, aspects of the other methods below might be adopted.

Sequential Hot Deck

The imputation system used in 1991 was based on the hot deck method developed by Felligi and Holt (1976).

A series of tables were designed reflecting the relationship between the variable to be imputed and other census variables. For example, it was known from the 1981 Census and intercensal tests that a good indication of the number of cars available to a household could be found from housing tenure, the number of people in the household, and whether the accommodation was in a permanent building or not. The imputation table for "number of cars" therefore held the observed values for number of cars available to households with all combinations of these reference variables. For example, if the number of cars was missing for a particular household, the most recently processed record with the same tenure, building type and number of persons as the recipient, was selected as the donor. The number of cars available to the donor household was copied to the recipient.

Fifty separate imputation tables (or **decks**) were used: 13 for household items; 24 for persons in households; one for communal establishments; and 12 for persons enumerated in communal establishments. For each cell in the imputation table, a series of values were held. These were updated continuously, with new values being taken from the most recently processed wholly valid record, and the oldest in that series of values being discarded. When an item required imputation, the newest value in the appropriate cell series was copied to take the place of the missing value. However, if this value had already been used to impute, the next oldest was used. For household variables, such as tenure and accommodation type, each cell held a series of 3 values, whereas cells in tables used to impute individual variables, such as a person's age and sex, held 6 values.

This hot deck method is known to work. It makes efficient use of computer processing capacity as each data file is read once only, although there is an additional storage requirement to hold data in both the imputation deck and main data file.

However, hot deck systems can be complicated and time consuming to program. This is closely related to the number and complexity of the imputation decks which are included in the system.

If more than one case in succession contains missing items, certain donor values may be used several times. The likelihood of this is reduced by storing several donor values in each class. Where there are many classes, the likelihood of having to re-use donor values increases, particularly where there is much missing data.

Simplified Hot Deck

This is essentially the 1991 system reduced in complexity, with fewer imputation decks and cells. There could be significant reductions in the time and resources need for system development, and there may also be reductions in the computer processing and storage resources required.

As simpler imputation classes are used, each cell should tend to be updated more frequently. There is therefore less chance of re-using donors. This can be seen as a trade-off: as classes are simplified, donors may be less similar to the case to be imputed but reuse of donors is less likely.

Hierarchical Hot Deck

Here, the data file is sorted into a much larger number of detailed imputation classes in a hierarchical structure. If no suitable donor is found at the finest level of the classification, classes can be collapsed into broader groups until a donor is found. A pattern of "hard" and "soft" class boundaries can be programmed into the hierarchical structure, e.g., to ensure that an item is always imputed from a donor of the same age group, even though the area of residence classes may be collapsed.

Hierarchical hot deck imputation frequently allows items to be imputed from very similar cases. However, the method is less efficient in its use of computing resources compared with sequential hot deck imputation. The system development process is difficult and time consuming, with possibly little benefit. For these reasons it is unlikely that the method would be of direct use for 2001. Aspects of the method, such as collapsible class boundaries for certain key or hard to impute variables, could be of possible use.

Statistics Canada -- New Imputation Methodology (NIM)

The NIM system is being investigated to see if aspects of this could be used. NIM allows forward searching within a data file to select a suitable donor record. This differs from the hot deck system used in the UK in 1991 which could only select donors from already processed records. This system may offer a way of avoiding re-use of donor values in areas where there is sparse data. NIM may also offer more plausible imputations.

Neural Networks

Unlike traditional computing approaches which need to be explicitly programmed, neural computers automatically learn solutions from the data itself. A neural computer can be taught and can learn about personal and household profiles provided in the census data in order to impute missing values.

Initially, the neural computer will go through this learning process, commonly known as "training." By using analysis tools, a model which has learnt profiles from the data may be analysed to show the relationships it has learnt.

Neural computing can impute by forming a model using examples showing how the imputed variable is related to the other variables, and then applying this model to cases where a value for the imputed variable is not known. One model is constructed for each variable to be imputed, where the model takes the form of a function that takes the known data as an input, and delivers the imputed value as an output.

A related development for 2001 could be the adoption of the One Number Census (ONC) approach to disseminating census and validation data. A ONC approach means only adjusted (for coverage) census results are output, as opposed to census tables and separate coverage correction factors. One way of doing this might be to impute missing person and household records of a number and type indicated by the validation estimates. These are people and households **not** identified by the enumerator, but estimated to exist by the use of alternative sources (such as administrative registers) or a follow-up survey. This would represent a significant increase in the role of the imputation system. It has not yet been decided if a ONC approach to dissemination will be adopted.

Disclosure Control

There are proposals that some form of additional imputation should be undertaken as a disclosure control technique. By deleting and re-imputing valid records, additional uncertainty of identification and matching is introduced which may reduce the need for other disclosure control techniques to be used. Although there are problems with this approach to disclosure control, this use of imputation will be considered.

Reference

Fellegi, I. and Holt, D. (1976). A Systematic Approach to Automatic Edit and Imputation, *Journal of the American Statistical Association*, 71, pp. 17-35.

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys, New York, Wiley. ■

Chapter 5

A Priority Index for Macro-Editing the Netherlands Foreign Trade Survey

Frank van de Pol and Bert Diederen,
Statistics Netherlands

Abstract

A macro-editing index for the Netherlands Foreign Trade Survey is described. This index is intended to trace errors by merely confronting current publication figures with the publication cell's history. A provisional experiment was carried out to determine the index's power in pointing to errors in the data. Information on pseudo errors was obtained from the present record-oriented editing process. Despite the big variability in many of the time series concerned, first results show a good association between index values and pseudo errors.

A Priority Index for Macro-Editing the Netherlands Foreign Trade Survey

Frank van de Pol and Bert Diederen,
Statistics Netherlands

Introduction

The Netherlands Foreign Trade Survey (NFTS) used to be a complete enumeration of all trade above a certain threshold, until in 1993 the EC put an end to restrictions on trade within the EC. Because of this, administrative customs data on the Netherlands trade with EC countries were replaced by survey data with the usual low response rates, ranging from 44 percent for small firms to 85 percent for large firms. Therefore, a shift of attention to increasing the response rate is necessary. Resources for this will be found in a complete redesign of the NFTS, especially in more efficient data editing.

Editing the 2 million transactions that come in each month, mostly via electronic data interchange, will be reduced to the bare minimum of valid value checks. The main instrument to trace errors will be macro-editing. An index is designed to prioritize inspection of the 58,000 nonempty publication cells. Only publication cells that deviate clearly from the value we expect from the past will be checked for errors.

A special feature of the data that shapes the priority index is that these many detailed trade time series contain lots of zero observations. Therefore our index is built on two distances, one relating the observed value to zero and the other relating the observed value to the expected non-zero value. The probability to observe a zero is also taken into account. A section on the data gives some more detail about the data. The next section on macro editing contains the formula we chose, and the section on tracing errors by macro editing gives results on comparisons of edited and unedited data. The final section gives the conclusions.

The Data

According to EC regulations, firms have to use a very detailed classification of goods and countries. Publication of monthly imports and exports in this amount of detail for the Netherlands turns out to give quite irregular figures for most publication cells. Berends et al. (1995) recently decided to publish monthly data on a somewhat more aggregated level, with about 50 country groups and about 800 product categories. When data on value and weight are aggregated to publication cells, the data matrix as displayed in Table 1 is obtained.

Table 1.--Notation for the NFTS Data; Import in Month *t*

	country 1, product 1	country 1, product 2	country c, product p	country C, product P	unit-total
unit 1	v_{111} q_{111}				v_{1++} q_{1++}
unit 2					
unit *u*	v_{u11} q_{u11}		v_{ucp} q_{ucp}		v_{u++} q_{u++}
unit *U*					
publication cell total			v_{+cp} q_{+cp}		

Rows are firm units and columns are publication cells. For firm unit u the import datamatrix holds value v_{ucp} and quantity q_{ucp} of product p from country c in month t. For exports, there is another matrix of the same format. In fact, both matrices are split in two submatrices, one for EC trade and one for non-EC trade. Although the number of publication cells has been reduced, these data matrices still hold hundreds of thousands of non-empty cells, as Table 2 shows.

Table 2.--The Size of the Datamatrix of the Netherlands FTS

	EC imports	EC exports	Non-EC imports	Non-EC exports
Country groupings, *C*	10	10	44	44
Product categories, *P*	798	797	796	796
Nonempty publication cells	6,681	7,538	17,688	26,088
Firm units, *U*	25,694	20,184	? 2,000	1,099
Nonempty matrix cells	347,356	273,109	? 100,000	83,598

?: Educated guess.

We want to use macro-editing principles to trace errors in these many publication cells (Granquist, 1994, 1995). The editors will look for the current publication figures that deviate most from what is expected, and look for errors in those columns of the datamatrix only. In the following we will describe the index we use and present some first results on its ability to trace errors.

A Macro-Editing Index Which Takes Zero Observations into Account

An index was devised to quantify the deviation between current and expected publication values. It should be approximately uniformly distributed between 0 and 100 in order to enable an interpretation in terms of a percentage of the cells. Moreover, we have to take account of the fact that one third of the publication cells that are nonempty for yearly figures, will remain empty with monthly figures.

The same index will be applied to value and quantity, for imports and for exports. Therefore we simplify notation to x_{tuc}, with t for month, u for firm unit and c for publication cell. A publication total is written as x_{t+c}. An unedited value is written with a prime, as x'_{tuc}.

An unedited publication value x_{t+c} might be in error, when it differs a lot from the value we expect, $\hat{x}_{t+c}$, on the basis of exponential smoothing of the cell c history (Michels, 1996; Siver and Peterson, 1985). However, with a very stable time series, differences are sooner suspect than with a variable one. Therefore we standardized the difference with the standard deviation of the time series concerned, $s_{x_{tc}}$, thus obtaining as distance measure

$$d_{tc} = \left| x'_{tc} - \hat{x}_{tc} \right| / s_{x_{tc}}.$$

Next, we observed that many time series have zero observations in some months, which greatly boosts the standard deviation and thus makes the d_{tc} less sensitive for an outlying current observation in zero-ridden time series. As a consequence, too small, but non-zero observations will not be noticed with this distance measure. To avoid this, time series are considered to have two regimes, zero observations and non-zero observations. An observation should be compared with the zero and the non-zero regime of the time series. We use on the one hand the distance measure between the observed value, x'_{tc}, and the predicted non-zero value, $\hat{x}_{tc} | x_{tc} \neq 0$,

$$d^{\varnothing}_{tc} = \frac{x'_{tc} - (\hat{x}_{tc} | x_{tc} \neq 0)}{s_{(x_{tc} | x_{tc} \neq 0)}},$$

and, on the other hand, the distance between the observed value and zero,

$$d^{0}_{tc} = \frac{\left| x'_{tc} - 0 \right|}{s_{(x_{tc} | x_{tc} \neq 0)}}.$$

These two distances are combined into a single measure using the probability to observe a zero, p_o, as predicted from the time series' history,

$$M^{*}_{tc} = v^{*}_{tc} \left[\left(1 + d^{0}_{tc}\right)^{\hat{p}_o} \times \left(1 + d^{\varnothing}_{tc}\right)^{(1-\hat{p}_o)} - 1 \right],$$

with value $v^*_{tc} = \max(v'_{tc}, \hat{v}_{tc})$. This M^*_{tc} measure has a minimum of 0 and no maximum. Its distribution is left skewed and has a long tail to the right. To obtain a more evenly distributed measure between 0 and 100 we transform it with

$$M_{tc} = \frac{100\ M^*_{tc}}{median\ (M^*_{tc}) + M^*_{tc}}$$

With this macro editing index some tests were carried out, which will be presented in the next section.

Tracing Errors by Macro-Editing

Before using real data we first did some testing with artificial data. Table 3 shows four time series of six months, each with non-zero mean 100, but with varying amounts of zero observations.

Table 3.--M_{tc} Values for Four Series and Four Current Observations, 0, 50, 100, and 200

Month	Series 1	Series 2	Series 3	Series 4
January	105	105	0	0
February	110	110	110	0
March	95	0	0	0
April	105	100	100	100
May	95	95	0	0
June	90	90	90	0
p_o	0	0.17	0.5	0.83
$\hat{x}_{tc} \mid x_{tc} \neq 0$	100	100	100	100
$s(x_{tc} \mid x_{tc} \neq o)$	7.07	7.07	8.16	10[1]
M_{tc} (July= 0)	66.5	54.8	26.8	6.6
M_{tc} (July= 50)	50.0	50.0	46.2	41.3
M_{tc} (July=100)	0	7.8	26.8	47.4
M_{tc} (July=200)	66.5	66.5	63.2	58.5

[1]With only one non-zero observation $s(x_{tc} \mid x_{tc} \neq o)$ was set to $(\hat{x}_{tc} \mid x_{tc} \neq 0)/10$.

As estimator for the predicted non-zero value, the mean of the non-zero values in the time series was used. The probability of a zero was estimated as the proportion zeroes in the time series. In a time series without zeroes and with mean 100, an observation of 200 turns out to be as alarming as an observation of 0 ($m_{tc} = 66.5$). When the series has a zero in it, an observation of 200 still has $m_{tc} = 66.5$, using the distance measures $d^{\emptyset}_{tc}$ and d^{o}_{tc}, which treat zero observations separately. A distance measure, which treats zero as an ordinary observation, would, due to a higher standard deviation of the series, wrongly consider 200 as less exceptional than our measure does.

An observation of 100, which is the mean non-zero value of all series considered, gives index M_{tc}= 0 when the time series has no zeroes in it. The more zeroes the time series has, the higher the index value will be for an observation of 100 ($M_{tc} = 7.8$ for $p_o = 0.17$, $M_{tc} = 26.8$ for $p_o = 0.5$ and $M_{tc} = 47.4$ for $p_o = 0.83$).

In order to test the macro-editing index with more realistic data, an arbitrary sample was drawn from the unedited data of August 1995. A limited number of products and countries of varying type was selected from the 44 country groupings and 798 product groupings available. Moreover, only large firms were included in the sample, which holds about 2000 firm units. This test sample was then matched with its history, time series of 24 months length, and with the corrected data after data editing.

In this file we computed pseudo errors, that is the difference of unedited values minus edited values. Our major concern was to test whether the macro editing index M_{tc} would be able to trace publication cells with large pseudo errors. If so, editing of non-EC trade will rely on this index for error detection

Table 4.--Relation Between Macro-Editing Index (Vertical) and Absolute Edit Size (Horizontal); Observed Non-zero Import Values of a Selection of Publication Cells in August 1995

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Total
$0<M_{tc}\leq 5$	14							13														12
$5<M_{tc}\leq 10$	9									13												8
$10<M_{tc}\leq 15$	6		100				9					10										6
$15<M_{tc}\leq 20$	5			67						13		10										5
$20<M_{tc}\leq 25$	4					8			17	13												4
$25<M_{tc}\leq 30$	5								8						33							4
$30<M_{tc}\leq 35$	2					8																2
$35<M_{tc}\leq 40$	3					8	9															3
$40<M_{tc}\leq 45$	3					8																3
$45<M_{tc}\leq 50$	3					23		13		25												4
$50<M_{tc}\leq 55$	3									25												3
$55<M_{tc}\leq 60$	3											10										3
$60<M_{tc}\leq 65$	3					8																3
$65<M_{tc}\leq 70$	3						9				100			50								3
$70<M_{tc}\leq 75$	3						18	25				10										4
$75<M_{tc}\leq 80$	4						9		17			10			33							5
$80<M_{tc}\leq 85$	3					8	9	13				33										4
$85<M_{tc}\leq 90$	7			33								10				25						6
$90<M_{tc}\leq 95$	6					15	18	13	25			10	50			50		50	50			8
$95<M_{tc}\leq 100$	11					15	18	25	33	13		30	50	50		25	100	50	50	100		14
Total %	85		.	1		2	2	2	2	2	.	2	.	.	1	1	.	.	.	.		100
Total freq.	465	0	1	3	0	13	11	8	12	8	1	10	2	2	3	4	1	2	2	2	0	550

Labels absolute edit size in Dutch guilders (≈ $0.68):

0: 0	6: 5000-10,000	12: 200,000-500,000	18: 7,000,000-10,000,000
1: 1- 10	7: 10,000-20,000	13: 500,000-700,000	19: 10,000,000-50,000,000
2: 10- 100	8: 20,000-50,000	14: 700,000-1,000,000	20: ≥50,000,000
3: 100-500	9: 50,000-70,000	15: 1,000,000-2,000,000	
4: 500-1000	10: 70,000-100,000	16: 2,000,000-5,000,000	
5: 1000-5000	11: 100,000-200,000	17: 5,000,000-7,000,000	

from June 1996 onward. EC trade is planned to follow in January 1997. An Oracle database is being built to guide the editors from a suspect publication cell to a suspect firm unit, a cell in Table 1. Underlying transactions can be looked up and corrected.

Table 4 shows a cross-table of non-zero reported import values with the absolute size of the pseudo errors on the horizontal axis and the M_{tc} index of all 550 non-zero publication cells considered on the vertical axis. The export counterpart concerns 772 non-zero publication cells and is shown in Table 5.

Table 5.--Relation Between Macro-Editing Index (Vertical) and Absolute Edit Size (Horizontal); Observed Non-zero Export Values of a Selection of Publication Cells in August 1995

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Total
$0<M_{tc}\leq 5$	16					10	6		4			7										13
$5<M_{tc}\leq 10$	7					3	19															6
$10<M_{tc}\leq 15$	5							5														4
$15<M_{tc}\leq 20$	4							5	12			7	11									4
$20<M_{tc}\leq 25$	4					10		5	4													4
$25<M_{tc}\leq 30$	5							5														4
$30<M_{tc}\leq 35$	2					3		9	4		9				33							2
$35<M_{tc}\leq 40$	5					3			4													4
$40<M_{tc}\leq 45$	3							5		13		7										3
$45<M_{tc}\leq 50$	4					7	6		8			13										4
$50<M_{tc}\leq 55$	3						6	5	4			7				13						3
$55<M_{tc}\leq 60$	3							5			9											3
$60<M_{tc}\leq 65$	3					7	6	5				7										3
$65<M_{tc}\leq 70$	4							9			9											3
$70<M_{tc}\leq 75$	3					3	13	9	4	38			11									4
$75<M_{tc}\leq 80$	3					14	6					7	22									4
$80<M_{tc}\leq 85$	3					14	6	5	12	13	18	7	11	33	33							4
$85<M_{tc}\leq 90$	6					3	6		4		9	13	11									5
$90<M_{tc}\leq 95$	6					7		9	12		9		11				33		50			7
$95<M_{tc}\leq 100$	11				100	14	25	23	31	38	36	27	22	67	33	88	67		50	100	100	15
Total %	79				.	4	2	3	3	1	1	2	1	.	.	1	.		.	.	.	100
Total freq.	612	0	0	0	10	29	16	22	26	8	11	15	9	3	3	8	3	0	2	3	1	772

Labels absolute edit size in Dutch guilders (≈ $0.68):

0: 0	6: 5000-10,000	12: 200,000-500,000	18: 7,000,000-10,000,000
1: 1- 10	7: 10,000-20,000	13: 500,000-700,000	19: 10,000,000-50,000,000
2: 10-100	8: 20,000-50,000	14: 700,000-1,000,000	20: ≥50,000,000
3: 100-500	9: 50,000-70,000	15: 1,000,000-2,000,000	
4: 500-1000	10: 70,000-100,000	16: 2,000,000-5,000,000	
5: 1000-5000	11: 100,000-200,000	17: 5,000,000-7,000,000	

We first focus on the frequency distribution of the edits that were made in the present, transaction-record oriented approach. It turns out that in most of the publication cells no edits have been done, 85 percent for imports and 79 percent for exports. This is the column labeled '0' in tables 4 and 5. Few cells have an accumulated effect of less than dfl.1000, which probably means that no very small corrections are carried out. Most edited cells, about 12 percent for import values and about 18 percent for output values, have undergone medium sized alterations, that is less than half a million guilders.

About one quarter of these medium sized edit effects are predicted by a macro editing index larger than 90. When editors would be looking at cells with an index value larger than 80 percent only, about 50 percent of the medium sized edits will take place. For high impact edits, causing publication cell changes of more than half a million, the "hit rate" comes close to 100 percent.

Tables 6 and 7, finally, refer to those publication cells that firm units reported to be zero in August 1995. Few zero observations are altered in the present editing process, in our sample 3 percent for import values and 1 percent for export values. These few cases are well predicted by the macro editing index.

Table 6.--Relation Between Macro-Editing Index (Vertical) and Edit Absolute Size (Horizontal); Observed Zero Import Values of a Selection of Publication Cells in August 1995

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Total
$0<M_{tc}\leq 10$	24																					23
$10<M_{tc}\leq 20$	7																					7
$20<M_{tc}\leq 30$	8																					8
$30<M_{tc}\leq 40$	5																					5
$40<M_{tc}\leq 50$	6																					6
$50<M_{tc}\leq 60$	5																					5
$60<M_{tc}\leq 70$	5																					5
$70<M_{tc}\leq 80$	7						100															7
$80<M_{tc}\leq 90$	11					50																11
$90<M_{tc}\leq 95$	8				100																	8
$95<M_{tc}\leq 100$	13					50								100				100	100	100		14
Total %	97				1/2	1	1/2							1/2				1/2	1/2	1/2		100
Total freq.	262	0	0	0	1	2	0	0	0	0	0	0	0	1	0	0	0	1	1	1	0	270

Labels absolute edit size in Dutch guilders (≈ $0.68):

0: 0	6: 5000-10,000	12: 200,000-500,000	18: 7,000,000-10,000,000
1: 1- 10	7: 10,000-20,000	13: 500,000-700,000	19: 10,000,000-50,000,000
2: 10-100	8: 20,000-50,000	14: 700,000-1,000,000	20: ≥50,000,000
3: 100-500	9: 50,000-70,000	15: 1,000,000-2,000,000	
4: 500-1000	10: 70,000-100,000	16: 2,000,000-5,000,000	
5: 1000-5000	11: 100,000-200,000	17: 5,000,000-7,000,000	

Table 7.--Relation Between Macro-Editing Index (Vertical) and Edit Absolute Size (Horizontal); Observed Zero Export Values of a Selection of Publication Cells in August 1995

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Total
$0<M_{tc}\leq 10$	18																					17
$10<M_{tc}\leq 20$	11																					11
$20<M_{tc}\leq 30$	8																					8
$30<M_{tc}\leq 40$	7																					7
$40<M_{tc}\leq 50$	6																					6
$50<M_{tc}\leq 60$	5																					5
$60<M_{tc}\leq 70$	8																					8
$70<M_{tc}\leq 80$	8																					8
$80<M_{tc}\leq 90$	7																					7
$90<M_{tc}\leq 95$	8																					8
$95<M_{tc}\leq 100$	14												100				100		100			15
Total %	99												1/2				1		1/2			100
Total freq.	338	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	2	0	1	0	0	342

Labels absolute edit size in Dutch guilders (≈ $0.68):

0:	0	6:	5000-10,000	12:	200,000-500,000	18:	7,000,000-10,000,000
1:	1- 10	7:	10,000-20,000	13:	500,000-700,000	19:	10,000,000-50,000,000
2:	10-100	8:	20,000-50,000	14:	700,000-1,000,000	20:	≥50,000,000
3:	100-500	9:	50,000-70,000	15:	1,000,000-2,000,000		
4:	500-1000	10:	70,000-100,000	16:	2,000,000-5,000,000		
5:	1000-5000	11:	100,000-200,000	17:	5,000,000-7,000,000		

Concluding Remarks

The results in the previous section showed that our macro-editing index will trace almost all large errors in the foreign trade data, but a sizable proportion of the mid-sized errors will remain undetected. Although most of these errors are too small to have a serious effect on publication figures, we still would like to boost the performance of our index with the larger ones.

One improvement will be that we will use a better time series model to predict the current value. In this preliminary experiment we simply used the arithmetic mean of the non-zero values in the previous 24 months. First results with exponential smoothing (Silver and Peterson, 1985; Michels, 1996) promise to be better.

Secondly, a robust estimate for the variability of the series would make the index more sensitive with series which behave relatively stable, but have one extreme outlier in their history.

Finally, we consider an alternative for the distance measures we presently use. Instead of dividing the absolute difference between an observed and a predicted value by a measure of the variability of the series, one could subtract 1.96 times the standard error of the predicted value,

$$d_{tc}^{\varnothing *} = \max\left[\, 0, \left|x'_{tc} - \left(\hat{x}_{tc} \mid x_{tc} \neq 0\right)\right| - 1.96 s_{\left(\hat{x}_{tc} \mid x_{tc} \neq 0\right)} \right].$$

With this measure, macro editing would only look at significant deviations from the expected value. The main problem with this measure is that an estimate of the standard error of the predicted value is not available for the case of exponential smoothing. Moreover, this measure will be 0 for all publication cells with nonsignificant deviations. This property is at variance with our wish to have a measure which has a nearly uniform distribution.

References

Berends, M. L. Visser; R. Janssen; and G. Slootbeek en N. Nieuwenbroek (1995). Onderzoek bij de Statistiek van de Internationale Handel (Eindrapport), Centraal Bureau voor de Statistiek, Heerlen.

Granquist, L. (1994). Macro-Editing -- A Review of Methods for Rationalizing the Editing of Survey Data, United Nations Statistical Commission and Economic Commission for Europe: Statistical Data Editing, vol. 1, *Methods and Techniques*, United Nations, Geneva, pp. 111-126.

Granquist, L. (1995). Improving the Traditional Editing Process. In Cox, Binder, Chinnappa, Colledge, Knott, (Eds.), *Business Survey Methods*, John Wiley, New York, pp. 385-401.

Michels, P. (1996). Voorspellingen Met Effeningsmethoden voor Controle/Correctie bij de Internationale Handel, Internal Research Note, Statistics Netherlands, Heerlen.

Siver, E. and R. Peterson (1985). Decision Systems for Inventory Management and Production Planning, second edition, Wiley, New York. ■

Chapter 6

Graphical/Interactive Systems

Chair: Cynthia Z. F. Clark, National Agricultural Statistics Service

Bill Goodman ◆ Laura Freeman ◆ Mike Murphy ◆ Richard Esposito

Paula Weir

Mary Kelly

Chapter 6

Experiences on Changing to PC-based Visual Editing in the Current Employment Statistics Program

Bill Goodman, Laura Freeman, Mike Murphy, and Richard Esposito, Bureau of Labor Statistics

Abstract

A demo and talk were presented on the ARIES system, as used in the Current Employment Statistics Program at the Bureau of Labor Statistics. The system uses a top-down graphical and query search technique to identify outliers in estimate-level data, and then isolate and treat outliers in the corresponding sample data. This talk discusses the experiences of both developers and users in adopting ARIES, including the problems encountered and suggestions for future development.

Experiences on Changing to PC-based Visual Editing in the Current Employment Statistics Program

Bill Goodman, Laura Freeman, Mike Murphy, and Richard Esposito, Bureau of Labor Statistics

Background

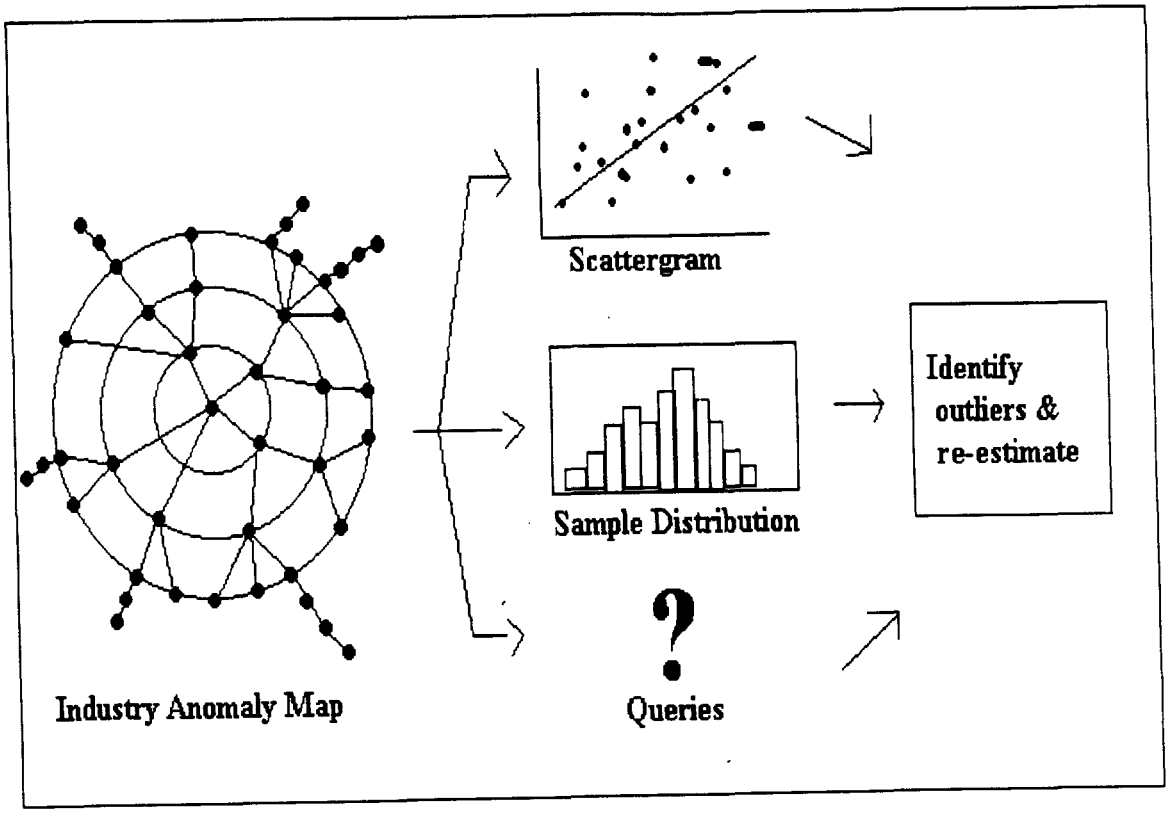
To convey the advantages of the ARIES system, some background on the survey in which it is used and on the preceding system will be necessary. The Current Employment Statistics survey is the largest monthly survey of employers in existence and one of the first major economic indicators each month. It samples nearly 400,000 establishments employing nearly 40 percent of all payroll workers, with over 45 million workers in the sample. It produces data for over 600 industries, including major divisions and more detailed industry levels. Our data were primarily collected by mail until we recently increased the use of electronic media. Now, phone collection, touch-tone self response, computer-assisted interviews, fax technology, and voice recognition are being used to obtain higher and faster response rates. The Current Employment Statistics survey of employment, hours, and earnings entails estimates of six to seven fields in 1,700 strata, for a total of approximately 10,800 estimates, monthly. Because of demands for quick responses, three sets of estimates -- two preliminary sets based on partial samples available earlier and the third set using the entire set of usable responses -- are produced each month.

The previous system, designed in the early to mid-seventies and written in COBOL, relied heavily on the use of listings, which were reviewed by nine industry analysts. Once questionable estimates were identified, it was generally necessary to review all individual reports for that stratum, as listings were sorted only by industry, state, and reporter identification number, so that finding all outliers entailed a complete review of the reports for the stratum. While certain preliminary screening of responses was performed, the screening did not catch all errors.

The ARIES System

ARIES is a graphical and query-based PC review system, designed to more easily identify and treat both estimate outliers at the macro level and sample outliers at the micro level. Review using ARIES consists of a winnowing through successively more micro-level data to isolate the questionable individual sample reports among the approximately 400,000 reports which may be responsible for unusual movements in estimates for the current month. For each data analyst, review starts with an anomaly map, which is a graphical tree-structure of the industries and estimates for which that analyst is responsible. Colors on the nodes of the map indicate the location of questionable industry estimates. Once significant problems are identified at this macro level, review continues through individual industry

scattergrams of reports representing sampled establishments, distribution graphs, and interactive queries, to isolate the individual sample members most likely to have contributed to the questionable estimate. Decisions on data weighting or rejection are made at the individual report level, and estimates are automatically re-generated on the PC and displayed on the PC screen. (A more complete description of this process can be found in the *Journal of Computational and Graphical Statistics*, Vol. 3, Number 2, June 1994.) The process is represented in the following schematic:



Loosely termed the "cell garden," the anomaly map is used to perform data validation at the macro level during the estimate review process. An excellent tool at each "closing," the anomaly map is especially helpful during the initial preparation of a given month's first preliminary estimates, or "first closing."

During the first closing, analysts must check their estimates at the macro level, decide which cells should be reviewed, and then search for reports that may be causing the abnormal estimates. Abnormal estimates may be due to typos in reported data and are often corrected. But if the estimate appears to be correct the analyst should become familiar enough with the situation such that it can be readily explained to interested parties. Finally, necessary changes are implemented, and the data are uploaded for merging into the production database. At first closing, the analyst has about three hours to perform these tasks and any others related to preparing and understanding the estimates for that month.

The anomaly map does two great things for the time-sensitive production analyst. The first is to identify cells which are out of range compared to the historical data for that particular month. The second is to make possible rapid review of all cells, regardless of whether ARIES found the cell's estimate abnormal. The ARIES anomaly map provides the historical over-the-month change data for each estimate. Estimates, or cells, can be selected and reviewed based on this history. The cell garden serves to facilitate the analyst's review of each cell in a rapid manner, and without the use of the somewhat cumbersome paper over-the-month change book.

Using the scattergrams in ARIES, one can select perhaps ten outlying individual reports for review, instead of reviewing all reports in the stratum, perhaps several hundred. The selected reports can immediately be reviewed in the ARIES system. An additional advantage of ARIES is its capacity to present, immediately on user demand, the latest sixteen months' data for a selected reporter. The old listings showed only the three latest months' data and the same three months a year earlier. With more data for the reporter, one can make a more informed decision about the validity of its latest data.

The Advantages of ARIES

One of the advantages of using the ARIES system is that members of our staff can take data editing a step further. For example, the Query function allows the analyst to investigate estimation problems more thoroughly by asking specific questions pertaining to the sample. For example, you could ask to see all the reporters that increased their employment by a certain percentage. To do so was impossible using the old paper method without going over thousands of pages of data, risking mistakes made on a calculator and using up valuable time. Another advantage to the query function is that it adds a new realm of research as to why estimates came out the way they did. It gives solid facts, for example, on reporting trends by state and comment codes. We are now able to tell in a matter of seconds how many of our reports came in coded, for example, with an effect of bad weather. So ARIES has not only improved our efficiency, but has also improved our analysis of the data. Our analysis of the data with the assistance of the ARIES system has become more accurate, more efficient, and more thorough. ■

Chapter 6

Graphical Editing Analysis Query System (GEAQS)

Paula Weir, U. S. Energy Information Administration

Abstract

The Graphical Editing Analysis Query System (GEAQS) is a software tool developed by the Energy Information Administration to graphically edit respondent level data using a top down approach with drill down capability. GEAQS combines exploratory data analysis and visualization techniques with the directional power of the anomaly map concept of ARIES, within an object oriented Windows application using PowerBuilder.

Graphical Editing Analysis Query System (GEAQS)

Paula Weir, U. S. Energy Information Administration

Background

In 1990 the Data Editing Subcommittee of the Federal Committee on Statistical Methodology released the Statistical Policy Working Paper No. 18, *Data Editing in Federal Statistical Agencies*. The paper presented the Subcommittee's findings that median editing cost as a percentage of total survey costs was 40 percent for economic surveys. The Committee felt that the large proportional cost was the direct result of over identification of potential errors. Hit rates, the number of identified potential errors that later result in a data correction divided by the total number identified, were universally very low. As a result, a lot of time and resources were spent that had no real impact on the survey results. The report cites research by the Australian Bureau of Statistics concerning the use of graphical techniques to find outliers at both the micro and macro level. A similar graphical approach to editing used by the U.S. Bureau of Labor Statistics for the Current Employment Survey is also described. The Automated Review of Industry Employment Statistics (ARIES) system helps to identify true errors quicker and results in fewer man-hours to edit the data. Graphics, particularly screen graphics, were found to be a preferable approach by the data processors and greatly reduced the amount of paper generated during the survey cycle. The recommendations of the subcommittee included the need for survey managers to evaluate the cost efficiency and timeliness of their own editing practices and the implications of important technological developments such as microcomputers, local area networks, and various communication links, as well as the expertise of subject matter specialists.

Subsequent to the efforts of the Data Editing Subcommittee, a working group of analysts, research statisticians and programmers was formed within the Bureau of Census to examine the potential use of graphics for identifying potential problem data points in surveys. It was felt that the existing procedure of flagging cases failing programmed edits and reviewing each edit on a case-by-case basis, had three main disadvantages. Examination of each case individually allowed the analysts to neither see the bigger industry picture nor see the impact of the individual data point on the aggregate estimate. The analysts, therefore, examined more cases than necessary. Thirdly, edit parameters or tolerances were derived from previous surveys which implied the relationships were constant over time. The group felt that the tools of exploratory data analysis combined with subject matter specialists' expertise were well suited for identifying unusual cases. The group considered box plots, scatter plots and some fitting methods, as well as transformations. This graphic approach could also be combined with batch-type edits while simultaneously evaluating dynamically set parameters or cutoffs. The working group concluded that a successful system requires that the system be acceptable to the people who use it. This requires training and incorporating the tools into the production environment and system. Two other systems, the Graphical Macro-Editing Application at Statistics Sweden, and the Distributed EDDS Editing Project (DEEP) of the Federal Reserve Board, have further demonstrated the efficiency of graphical editing.

The Concept

The Graphical Editing Analysis Query System (GEAQS) is being developed by EIA as a tool to reduce survey costs and reduce the amount of paper generated. It combines and builds on the features of the four other systems mentioned above--the ARIES system, the Census Working Group prototype, the Graphical Macro-Editing Application, and DEEP. The GEAQS borrows from the ARIES system the concept of an anomaly map which summarizes the relationship of various levels of aggregates and flags questionable aggregates through the use of color. This top down method of editing provides the user the ability to drill down through the aggregates to the respondent level. From the Census Working Group prototype and recommendations, GEAQS makes use of the tools of Exploratory Data Analysis. Box-Whiskers graphs summarize aggregate changes from the previous period to the current period through multiple boxes for the "children" of the select higher level aggregates. Further subaggregates are visible and identifiable within each box. Scatter plots are used to further drill down and display respondent level data for the current period versus the last period for the select aggregate. Actual reported data is distinguished from imputed data by the use of circles and triangles. This allows the user to pursue different follow-up procedures accordingly. The additional benefit of different symbols for respondents and imputed data is the visualization of the distribution of imputed data with respect to reported data and confirmation of whether respondents are similar to nonrespondents. Data points with high influence are indicated by color. High influence points that visually deviate the most from the trend contribute the most to the overall change. Outliers of low influence, if not systematic, are not as cost effective to pursue and contribute to over editing. Batch edit flags can be passed to the system to further prioritize the failures, as well as evaluate and help determine parameters or cutoffs. GEAQS builds upon the need for a Windows' application as developed by Statistics Sweden. This allows the user to point-and-click on an aggregate in the anomaly map or the Box-Whiskers, as well as a data point on the scatter graph. The user can take advantage of tool bars, dialogue boxes, and icons. Resizing and zooming are built in to enable the analyst to focus on particular parts of a graphic. Tiling, on the other hand, allows the analyst to maintain the previous graphic while operating on the next graphic of the same drill down effort. An icon for a legend is also provided to assist the analyst in distinguishing colors, shapes, etc. In order to maximize the usefulness of GEAQS to other surveys, additional time and effort was taken to make GEAQS object oriented. This allows for minimal costs to modify or enhance GEAQS to operate on surveys other than the survey originally piloted. It will also allow for ease of integration with the rest of the data processing system.

GEAQS will also build on the work done for the DEEP system of the Federal Reserve Board by capitalizing on time series information. It allows the analyst to view the respondent data over an extended period of time. What may appear as an anomaly with respect to other respondents in that cell may be consistent with that respondent's historical reporting. This capability supplemented with pull down text comments helps the analyst determine if the respondent's reporting difference has been verified previously. Like the Federal Reserve System, GEAQS was developed in PowerBuilder and uses Pinnacle graphics server to help generate the graphs. The use of PowerBuilder and Pinnacle resulted in quicker development time and less cost. In addition, in order to capture the recommendation of the Census Working Group that the system is acceptable to the people who use it, the development of GEAQS emulated the iterative user feedback process used by the Federal Reserve Board through testing by users at various stages of development. Unidentified requirements were quickly discovered and modifications made. This made the product more useful to the analysts by allowing their direct input throughout the process.

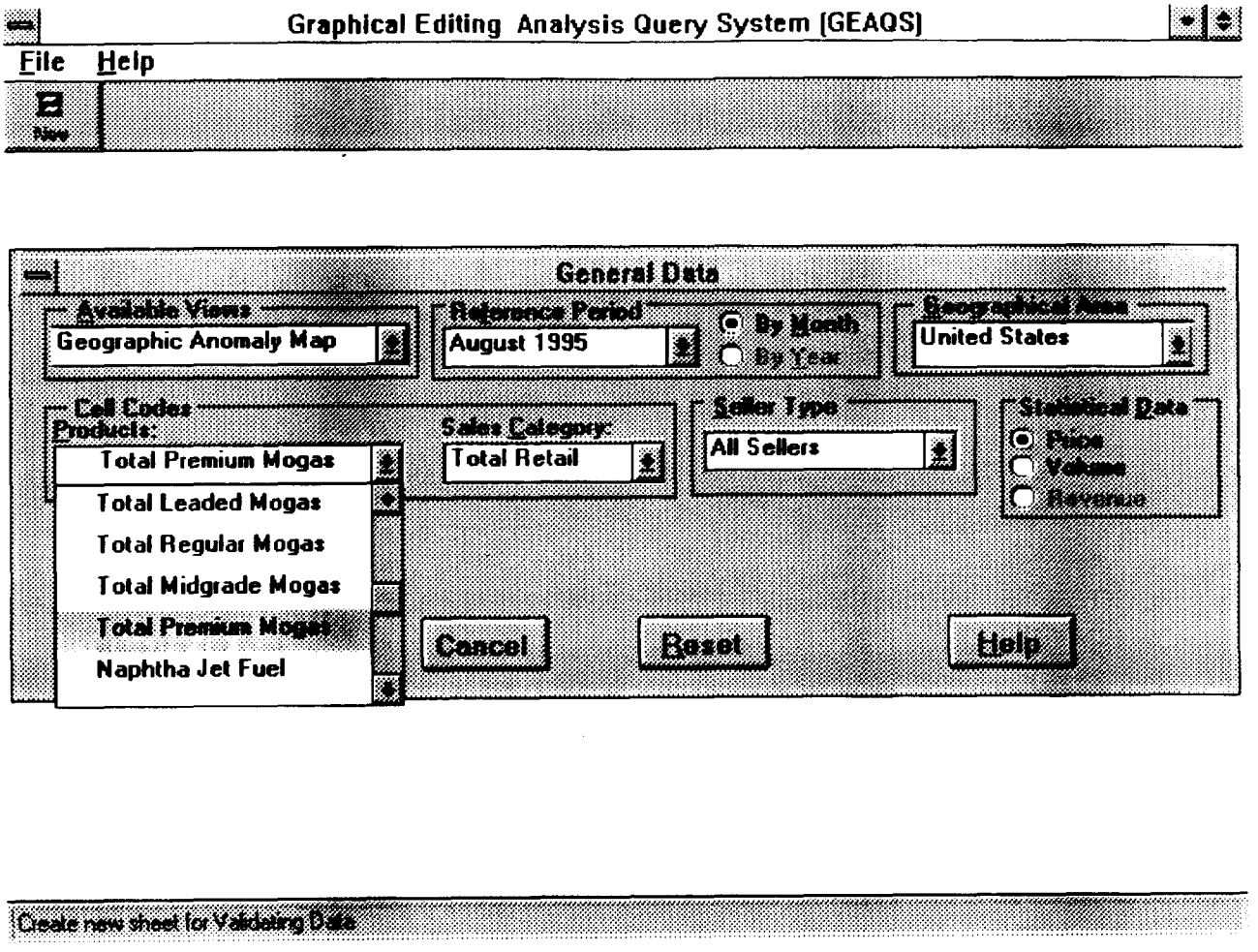
GEAQS also incorporates many of the visualization techniques described by William Cleveland. The top-down approach is an iterative process. Edit failures are not just listed prioritized or ranked by some predetermined variable. The analyst discovers which aggregates deviate the most, which next level aggregates directly contribute, and then which respondents are outliers and which have a high impact on that aggregate. Circles are used for data points to minimize darkening with the exception that triangles are used for imputed data. Only two colors, limited to four shades each, are used in the anomaly maps, while the scatter graphs contain only three colors. Colors are used to distinguish different levels of severity. Even though legends are provided, the limited number of colors allows for "effortless perception." That is, it lessens the need to use the legends which would be a cognitive process. Limiting the number of shades allows for clear distinction between shades within a color. In addition, visualization in scatter graphs of data also requires fitting the data. The fit may not be immediately apparent. GEAQS displays a least squares regression line in addition to the no change or current-equals-prior line for orientation. Transformations, particularly power transformations, of the data may also be necessary to uncluster the data, reduce the spread of the data, or reveal an underlying linear relationship. Logarithms make the data more symmetric and reduce skewness, monotone spread and multiplicative effects which make it difficult to visually determine the true outliers. To further assist in unclustering and identifying individual responses, zoom and resize capabilities are provided by a mere click on the respective icon. Tiling of the windows is also possible, allowing the analyst to keep the bigger picture in mind or a road map of where the analyst is in the process. The scatter graph automatically brings up the data table/spreadsheet into the right half of the window. Clicking on individual data points highlights the data in the spreadsheet and vice versa. Analysts can choose to focus on certain parts of the graph by drawing a box around the points of interest and then selecting either the inside box or outside box icon. The graph is then redrawn showing only the chosen set of data points. Similarly, the data table will reflect only those points.

The pilot survey used in the development of GEAQS was chosen because of its complexity. It was felt that if graphical editing could be successfully accomplished for this survey, it would be a small task to modify the system for other surveys. The survey chosen collects state level prices and volumes of petroleum products sold monthly from a census of refiners and a sample of resellers and retailers. Volume weighted average prices are published at the state, Petroleum Administration for Defense District (PADD), and U.S. level for a variety of sales types and product aggregation levels. Volume totals and volume weighted average prices for refiners are also published. Approximately 60,000 preliminary and final aggregates are published each month.

The Application

The user of GEAQS is provided the flexibility to decide where in the system to start. After clicking on the "new" icon, the opening dialogue box (Figure 1) allows the user to choose from various views. Four of these views are associated with aggregates -- three anomaly views and a delta graph (Box-Whisker on change). The anomaly views are available for geographical, product, and sales type, the three main dimensions of the pilot survey, in addition to time. The geographical view requires the user to also select a product, sales category, seller type, statistical data type, and reference period from the drop-down lists provided by clicking within the respective boxes. As the user makes the view selection, the system adjusts the possible product selection, according to the combinations of aggregates calculated by the survey's processing system. Similarly, as the user selects the product, the list of

Figure 1

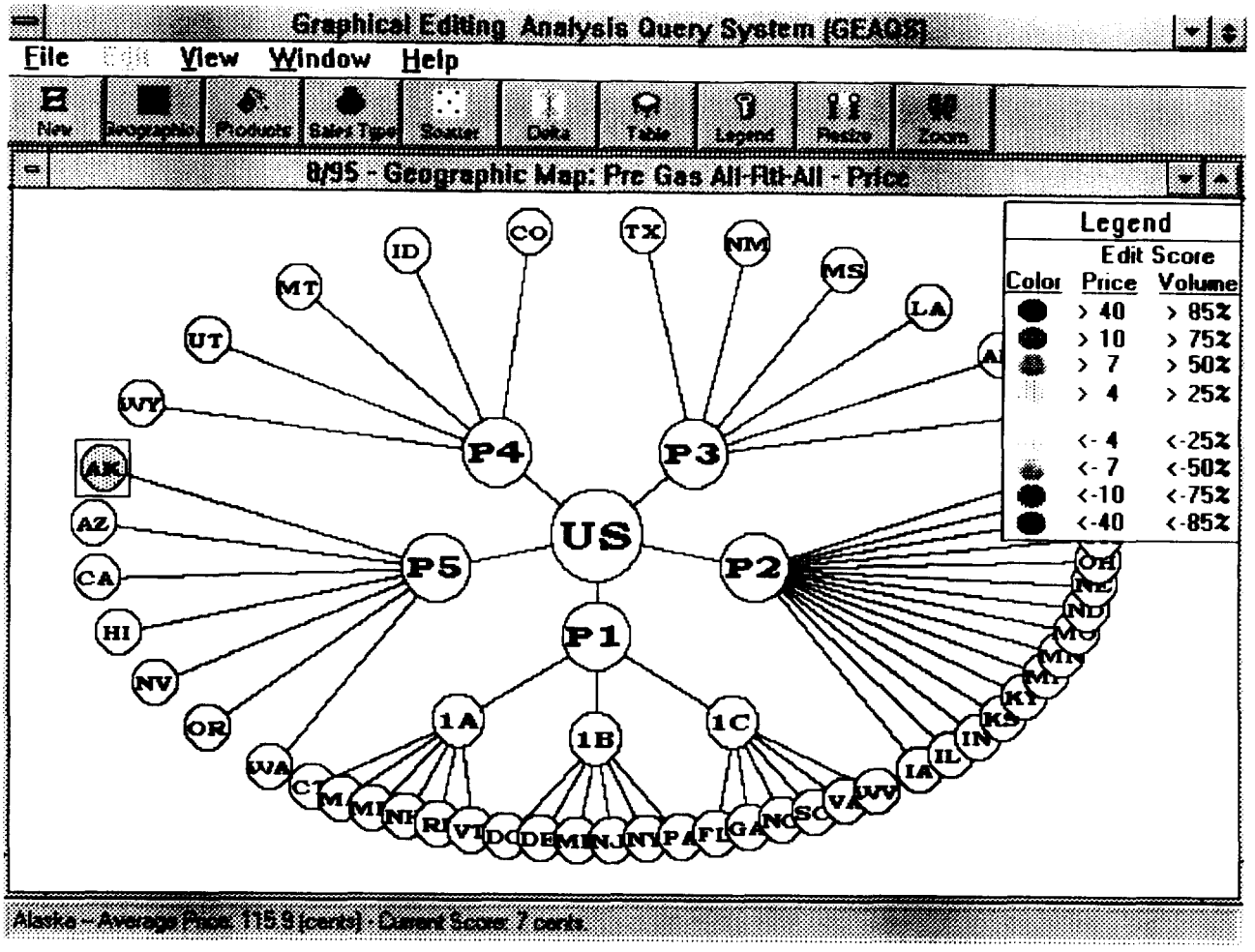


possible sales categories is adjusted accordingly. Once all selections have been made, the user clicks the OK button. The graphic is then displayed (Figure 2). The geographical anomaly view graphically represents aggregate cells of the selected data by placing a node for the highest level aggregate, the U. S., in the center of the map. Orbiting out from the center are nodes for the next level of aggregates, five regions of the country called PADDs. One PADD is broken out into three more nodes for subPADDs. From each PADD or subPADD node, state level nodes are used to represent the lowest level of geographic aggregate. Each node, regardless of the level, is colored according to its current edit score. For price data, the current score is the difference between the price change (current price minus the previous period price) at the state level and the price change at the PADD or subPADD level calculated without including that particular state; the edit score for state k, at time period t is:

$(P_{k,t} - P_{k,t-1}) - (P^*_{.,t} - P^*_{.,t-1})$, where $P^*_{.,t}$ is the PADD average price excluding state k.

Volume and revenue current scores are similar, but use the difference in percent change between the state and the PADD or subPADD. The current scores for the U.S. and PADDs are just the price change between the previous and current period. Four shades of blue are used to represent scores that indicate the price change is greater for that area (state or subPADD) than the more aggregated geographical area

Figure 2



(PADD or U.S.) by 4, 7, 10, or 40 cents as the darkness of the color increases. Similarly, four shades of green are used to represent area price changes that are less than the more aggregated geographical area by 4, 7, 10, or 40 cents. Areas where data do not exist are shaded grey. The analyst may click on the legend icon to clarify the color distinctions. The legend may be moved around the window or turned off as the user desires. If the user had chosen volume or revenue, rather than price for the statistical data selection, the shades of blue and green would represent different levels of percent change. A user may click on any node of the map to activate a geographical area colored to indicate a large price increase or decrease relative to the PADD. The tool bar at the bottom of the window will show the name of the state, subPADD, PADD, or U.S. node activated, along with the weighted average price and the score for the area. The user may drill down by either clicking on the products or sales type icon. If the user had previously selected a product that can further be broken down to the reported product level, the user would choose the product's icon. The window would be replaced by a new graphic, a product anomaly map (Figure 3), that shows for the activated geographical area node all products broken down to the reporting level component products. The nodes are shaded the same way as the geographical anomaly map to indicate the levels of the edit score. The user can click on the appropriate component product to activate the reporting level product and then click the sales type icon to further drill down. The screen is then replaced with the sales type anomaly map (Figure 4) which shows retail and wholesale sales type components for the activated state and product. Colored nodes are again used to signify the levels of relative change for the various sales types.

Figure 3

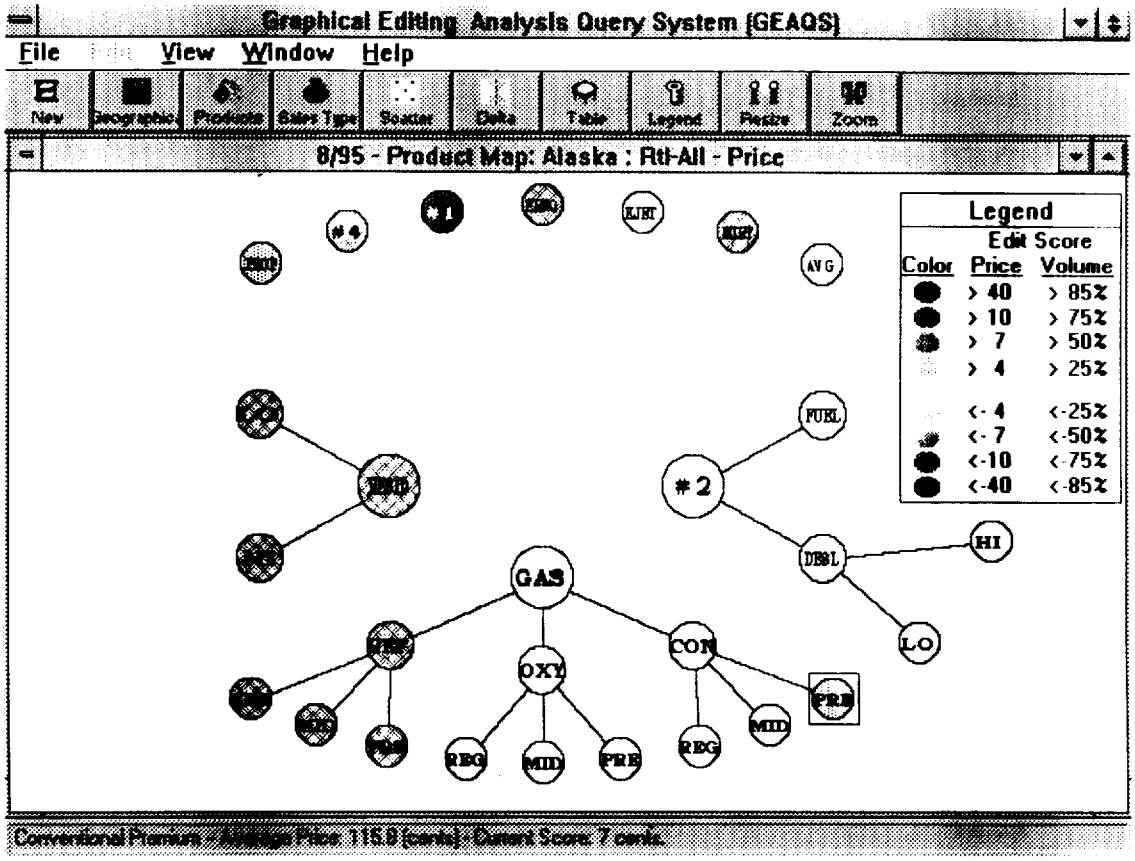
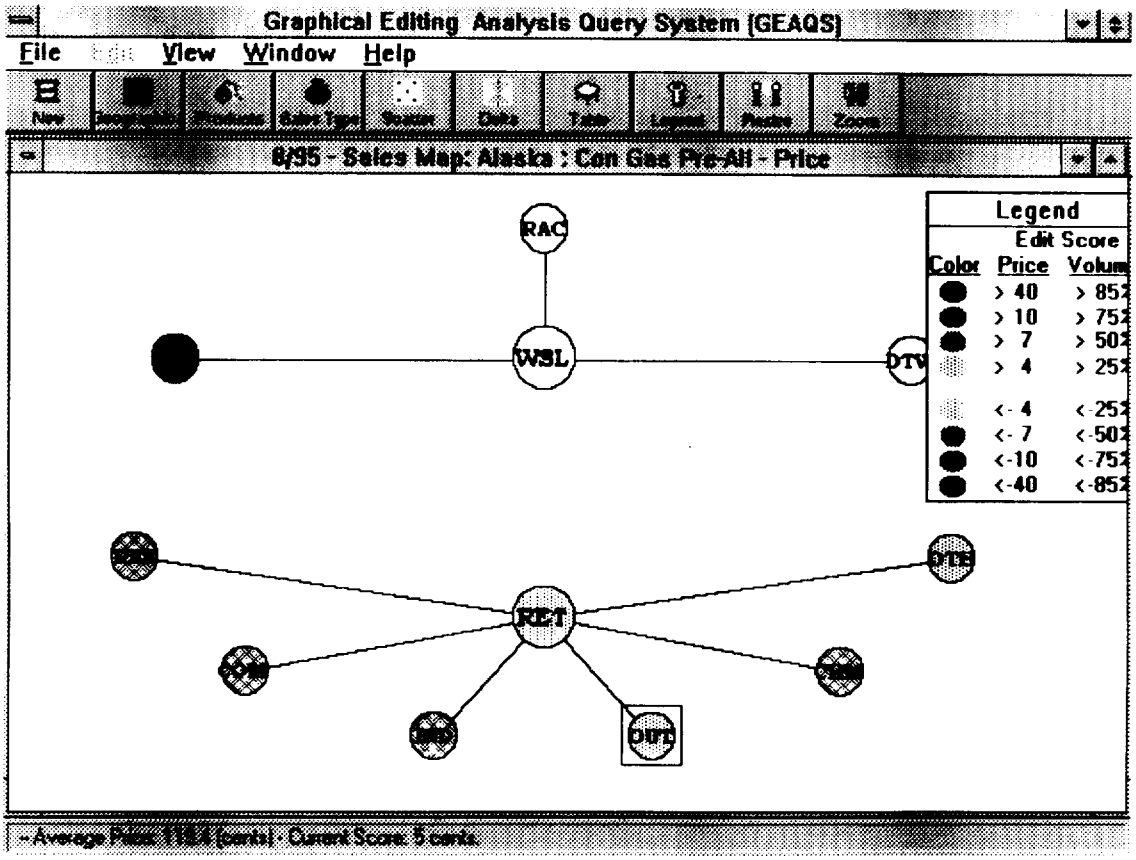
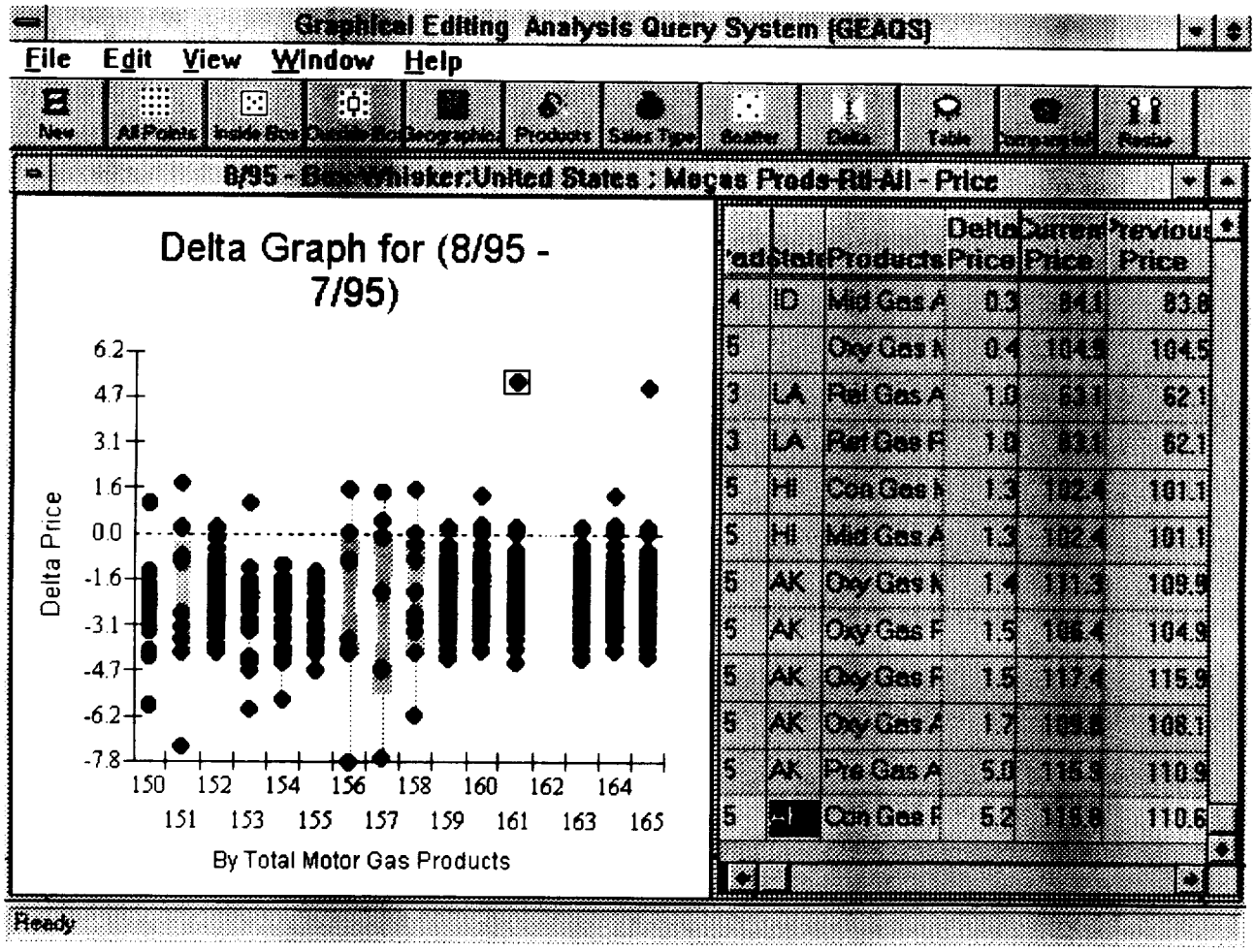


Figure 4



GEAQS allows the user to determine the path for drilling down. A user can start with a product or sales type anomaly rather than a geographical anomaly. The procedure is the same, only the order changes. An alternative procedure for drilling down is provided through the delta graph, a Box-Whisker graph of change -- price, volume or revenue -- between reference periods. In the opening dialogue box, the user selects delta graph under the available views. The user would next select a group of related products through a product selection preceded by "all," a high-level sales category, total retail or total wholesale, and all sellers for seller type. Once all selections have been made, the user clicks the OK button. On the left side of the window, the Box-Whisker graphic (Figure 5) displays a box plot for each individual product in that product group, allowing the user to compare the spreads of the changes across those products. The vertical axis represents the change (price, volume or revenue), positive and negative, between the current and previous reference periods. Each box plot is labeled at the bottom by the product code associated with it. The "waist" of the box signifies the median for that product across geographical areas, including the aggregate areas of subPADD, PADD, and U.S. Individual circles plot the change for the geographical areas within the box, the middle 50 percent of the values for the geographic changes, within the whiskers, and outside the whiskers, which are called outside values.

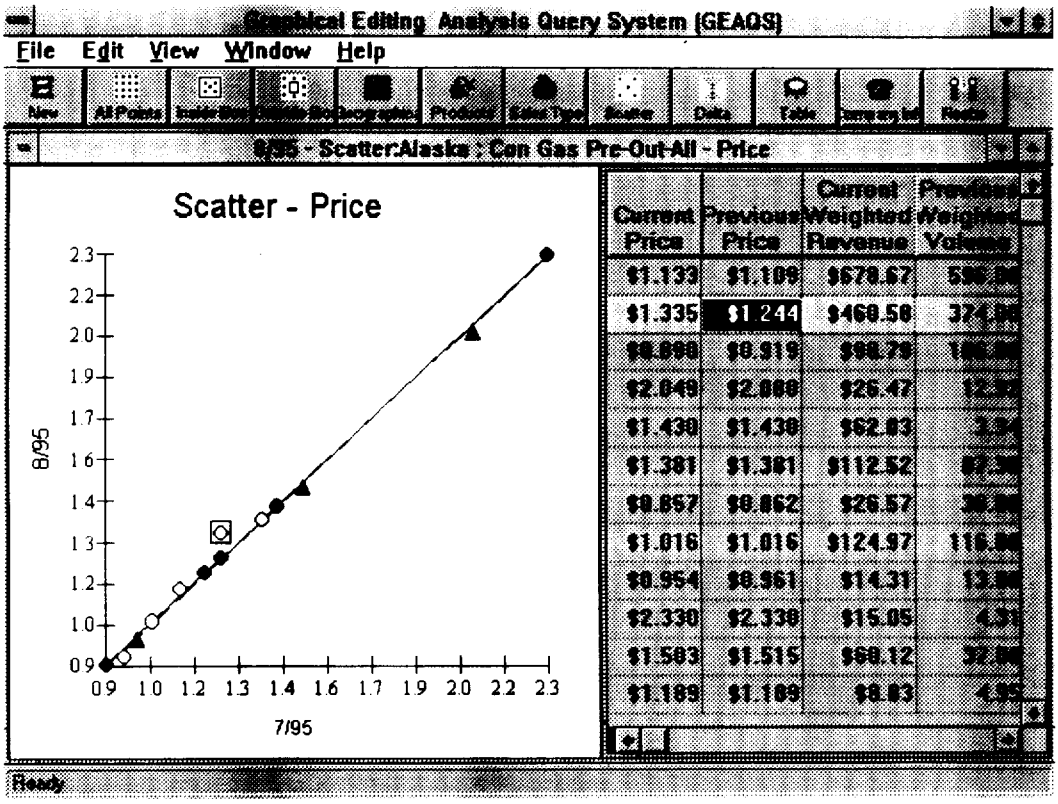
Figure 5



The values within the whiskers are those values less than or equal to (greater than or equal to) the upper quartile plus (lower quartile minus) 1.5 times the distance between the upper and lower quartiles. Values beyond the whiskers, outside values, may not exist if the largest (smallest) valued geographic area within the whisker is the maximum (minimum) of the changes of the geographical areas. If outliers exist, they would be outside values. The additional information gained from the Box-Whisker is the summary of the distribution of change. If the distance between the top of the box, the upper quartile, and the median is very different from the distance between the bottom of the box, the lower quartile, then the distribution of change is skewed. The right side of the window contains a spreadsheet of the information for each circle on the plot. The analyst can click on any circle and the associate row of information for that value will be highlighted in the spreadsheet. The change for that aggregate cell, the current period's actual value, the previous period's actual value, as well as the label of the cell's state and/or PADD/ subPADD and other relevant data are provided in the highlighted row. Utilizing windows' functionality, the analyst can scroll across, up or down the spreadsheet by clicking on the appropriate window's arrow buttons. Columns in the spreadsheet can be rearranged by the usual click, and drag method, clicking on the column title at the top of the column. Column size can be changed by clicking and dragging the line that separates the columns. Leading columns can be held fixed while scrolling across the rest of the spreadsheet by clicking on the shaded area left of the arrow button at the bottom of the screen and dragging it to the end of the last column to be held fixed. An icon is also provided for the Box-Whiskers graph. After an analyst has identified a particular product and sales category through the anomaly maps, the analyst can click on the Box-Whiskers' icon to see a single box plot representing the distribution of change across geographical areas between the previous and current periods. Regardless of the path chosen, at this point the analyst has determined the lowest level aggregate(s) that contributed the most to the higher level aggregate anomaly.

The analyst can further drill down to the respondent level by clicking on the scatter icon. For the activated geographical, product, and sales type, a scatter graph of the data will be displayed in the left half of the window (Figure 6). The y-axis is the coordinate for the current period and the x-axis is the

Figure 6



coordinate for the previous period. Each respondent-level price, volume or revenue is plotted using a circle and each nonrespondent's imputed value is plotted using a triangle. Data values whose contribution to the aggregate are 50 percent or more are depicted by red, values that represent 5 percent or more, but less than 50 percent, are yellow, and the remaining values, less than 5 percent share, are blue. A dashed line is provided that indicates no change; the current period's value equals the previous period's value. Data falling above this line indicate increases in the current period, while data below represent decreases in the current period. In addition, a least squares regression line is also provided, represented by a solid line. The analyst can draw a box around points of interest by clicking to the left and above the respective points, holding down the button, and dragging to the bottom right of the respective points and releasing the button (Figure 7). The user then clicks on the "inside box" or "outside box" icon to have the graph redrawn according to the selection, using only those points in the box or those points outside the box (Figure 8). The "inside box" icon allows the analyst to uncluster points and focus on particular values. The "outside box" icon allows the user to examine the scatter without certain points. The original graph can be obtained by clicking on the "all points" icon. The right side of the window shows the information relating to each point on the graph. Each row of this spreadsheet represents a respondent. The spreadsheet contains the values of each point, respondent identifier information, sample weights and volume weights, and other relevant information. The analyst can click on a row in the spreadsheet, highlight it, and a box will appear around the corresponding point on the scatter graph. Alternatively, clicking on a point in the scatter graph, which boxes the value, will result in highlighting the corresponding row in the spreadsheet associated with that value. Further information for contacting the respondent can be obtained by clicking on the "company" icon. The analyst can scroll up, down, or across the spreadsheet and rearrange columns as previously described for the spreadsheet associated with the Box-Whiskers plot. The combination of the scatter graph and the spreadsheet provide the user the tools needed to identify the specific respondent(s) causing the aggregate cell to be an anomaly.

GEAQS was designed to be interactive with the data base of the processing system. Once a particular respondent value has been identified, the analyst could change the response directly in the spreadsheet if so desired. The analyst would then be able to reexamine the newly computed aggregates to determine if it were still an anomaly. At this time, because GEAQS is not tied in with the processing system, and the pilot survey's estimation system is too complex to duplicate within GEAQS, the changing of respondents' data and recalculation of the aggregates cannot be demonstrated using the pilot's downloaded Watcom SQL database. It should be clear, however, that changes could be made, even temporarily, to determine the effect of the change.

Future Enhancements

Additional enhancements are still to be made in GEAQS. Work is ongoing to incorporate a more sophisticated measure of each respondent's contribution to the aggregate change. In particular for the pilot survey, because price is the ratio of revenue to volume, a respondent's contribution can be measured by the shift in the respondent's market share of revenue between months multiplied by the difference in price between that respondent and the respondent who inherits (or gives up) the majority of the market share in the corresponding month. This contribution to the change would be an improvement over a simple market share measure for influence which only indicates potential for contribution to the aggregate change. The other major enhancement to GEAQS is called "bubble up." This functionality provides the user anomaly information at the highest levels of aggregates concerning the associated

Figure 7

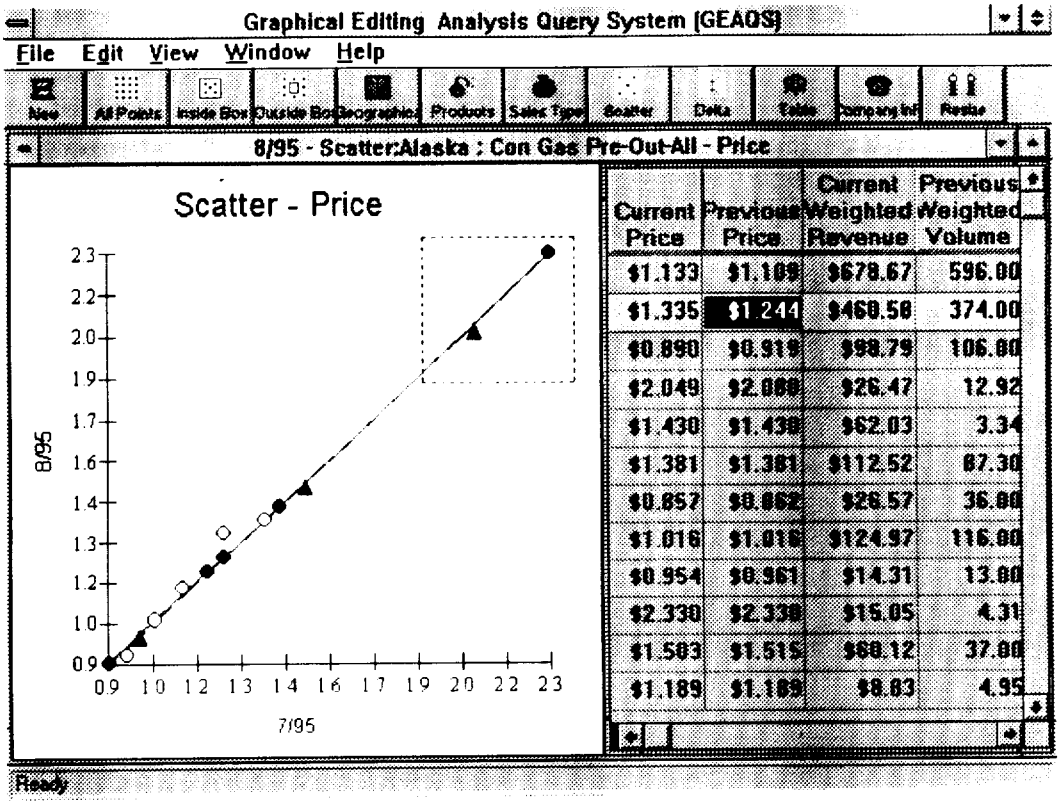
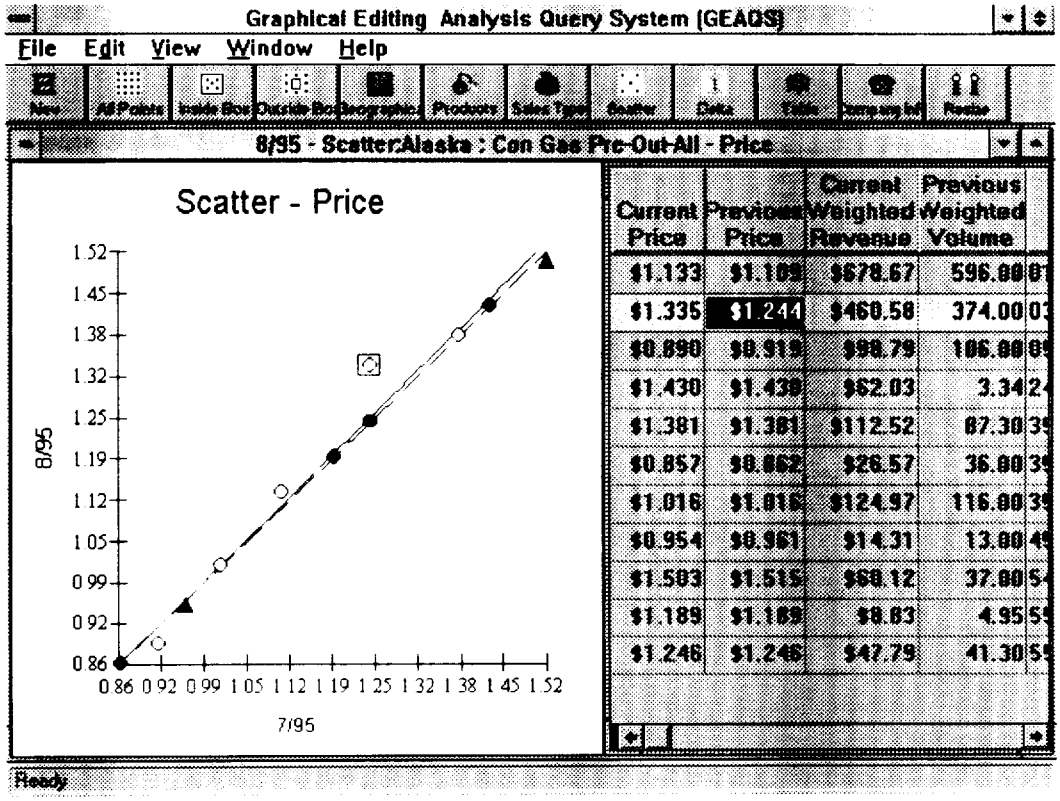


Figure 8



lower levels of aggregation. It graphically signals the user that even though the current aggregate is not anomalous, a component of that aggregate is anomalous. The user would immediately see where drilling down was necessary. This would remove from the user the burden of having to bring to the screen lower level published aggregates to determine if there are outliers at that level. It is expected that when GEAQS is incorporated into the processing system other variables in the data base will be available to it. Respondents who failed edits in the batch process can be flagged in the spreadsheet and scatter gram. Transformations such as logarithms and roots will also be possible. Recorded comments obtained by contacting respondents will be accessed by clicking on the "company" icon. Time series data for aggregates and respondents will also then be possible. Standard errors of aggregate estimates will also be incorporated.

Summary

The Graphical Editing Analysis Query System (GEAQS) built upon the concepts developed in four other systems. A top down approach to data editing and validating, macro-editing, enables the analyst to efficiently focus on outliers that impact the published aggregates. GEAQS provides anomaly maps and Box-Whiskers plots to identify aggregate level outliers. The anomaly maps summarize the relationships of various levels of aggregates and highlights outliers through color as determined by the current edit score. In comparison, the Box-Whiskers plot summarizes the distribution of change across geographical aggregates, allowing comparison of distributions within product groups, and highlights outliers as the outside values, outside the whiskers. Either path that is chosen directs the analyst to drill down to the lowest level aggregate. The scatter graph of the lowest level aggregate depicts the respondent level data that contribute to the aggregate. Outliers are identified by their position relative to the other respondents' values and the fit line, while color is used to emphasize respondents' influence on the aggregate estimate. The split window with the spreadsheet mapping to the scatter graph provides immediate identification of the values.

References

Bienias, J.; Lassman, D.; Scheleur, S.; and Hogan, H. (1995). Improving Outlier Detection in Two Establishment Surveys, ECE Work Session on Statistical Data Editing, Athens 6-9, Working Paper No. 15.

Cleveland, William S. (1993). *Visualizing Data*, Hobbart Press, Summit, New Jersey.

Engstrom, P. and Angsved, C. (1995) A Description of a Graphical Macro Editing Application, ECE Work Session on Statistical Data Editing, Athens 6-9, Working Paper No. 14.

Esposito, R.; Lin, D.; and Tidemann, K. (1993). The ARIES Review System in the BLS Current Employment Statistics Program, *ICES Proceedings of the International Conference on Establishment Surveys*, Buffalo, New York.

Mowry, S. and Estes, A. (1995). Graphical Interface Tools in Data Editing/Analysis, (1995). Washington Statistical Society Seminar presentation.

Subcommittee on Data Editing in Federal Statistical Agencies, Federal Committee on Statistical Methodology (1990). *Data Editing in Federal Statistical Agencies*, Statistical Policy Working Paper 18, Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget. ■

Chapter 6

Integrated Data Capture: A System for All Office of Compensation and Working Conditions Surveys

Mary Kelly, U. S. Bureau of Labor Statistics

Abstract

The Office of Compensation and Working Conditions (OCWC) conducts surveys of wages and other compensation by occupation and industry. Until recently, the office collected these data via three surveys: the Employment Cost Index, the Employee Benefits Survey, and the Occupational Compensation Survey Programs.

Each survey was completely independent of the other, with its own method of sampling, set of field collections staff and survey administrators, and its own set of computer systems for data entry, editing, transmittal and publication.

Approximately one year ago, the Office began a major reconceptualization of its survey programs, effectively collapsing the three sets of surveys into one consolidated survey program, along with a single Integrated Data Capture (IDC) system.

This presentation and associated exhibit will highlight the Integrated Data Capture system. The presentation will be built around answers to the following questions:

- Why was the IDC built?
- How was IDC designed and tested?
- What does IDC do?
- What is unique about IDC?

We offer the following overview of IDC -- The system was designed to consolidate all aspects of data capture for the three sets of surveys within OCWC. The data capture portion of IDC is a Windows-based Power Builder

Abstract (Cont'd)

system with data being fed into a relational SYBASE database residing on a Unix server. Prior OCWC computer systems were primarily mainframe platforms built as large batch systems requiring significant resources to implement modifications. IDC is a modular database system, with each component (e.g., data capture and edits, database, transmittal) residing independently of the others. Such a system has obvious advantages with regards to updates to the software.

Prior OCWC computer systems were designed so that many of the edits occurred at the same time as data entry. Thus, the system would "stop" and query the user if an important data element was missing or "out of scope" of an edit.

Such a design rendered the system unusable within a "live" in-person collection environment. IDC permits the user to enable the edits at the end of data entry, thus making the system viable for use during in-person collection.

Finally, IDC is compatible with the latest programmatic priority -- moving away from centralized data review and towards schedule review and edits under the greater control of individual field staff.

Integrated Data Capture: A System for All Office of Compensation and Working Conditions Surveys

Mary Kelly, U.S. Bureau of Labor Statistics

Introduction

The Bureau of Labor Statistics' Office of Compensation and Working Conditions (OCWC) conducts surveys of wages and other compensation data by area, industry, and occupation. The data for the surveys are collected across the country by about 160 field staff located in 8 regional offices and 9 smaller outstations. Survey data are collected by the data collectors, reviewed at the regional offices by the review staff, and sent to the main office in Washington for final processing. Until recently, the office collected data via three surveys: the Employment Cost Index, the Employee Benefits Survey, and the Occupational Compensation Survey Program. Each survey was completely independent of the other, with its own method of sampling, set of field collection staff and survey administrators, and its own set of computer systems for sampling, data entry, editing, transmittal, and publication.

Approximately one year ago, the Office began a major reconceptualization of its survey programs and is in the process of integrating the three surveys into one. With the development and testing of new survey procedures, the time seemed right to develop new systems for the new survey.

This paper is a report on the new system that was designed. The Integrated Data Capture System (IDC) is a Windows-based Power Builder application that accesses a SYBASE database located in Washington. The application can also access a similarly designed Watcom database loaded on the hard drive of a PC or laptop. The System development started in July 1995 and was used for the first time in a test survey in February 1996.

Structure of IDC Development Team

The Office-wide redesign was brought about by the exigency, due to downsizing, of combining the three surveys into one new survey. Collapsing the three surveys necessitated a review of all aspects of survey sampling, data collection methodology, review, and processing. Deciding which of these systems to revise and/or develop first presented the initial challenge. Given the publication criteria already in place within OCWC, it was determined that the sampling methodology would be similar to one of our current surveys, and thus, we could use one of the existing survey sampling systems with only slight modifications. In addition, it was felt that although we knew what new data we wanted to publish we were unsure just how much of the data collected from of our test surveys would meet the publication criteria. Given this uncertainty, we decided to take a wait and see attitude regarding making changes to the existing tabulation systems. It became quickly evident that the greatest change was within the data collection methodology -- we were combining three compensation surveys, each survey measuring a distinct aspect of compensation, and we wanted to retain, as much as possible, all the individual data

elements. Changes to data collection methodology implied that data entry and data editing system development had the highest priority. Furthermore, coordinating the redesign of these systems with the redesign of survey procedures would allow us to conduct a dual test of both the new survey procedures and the system at one time and one location.

To develop the IDC, we convened a small group of individuals, each of whom knew the content of the new survey and had worked on a variety of the Office's previous compensation systems and special projects. This small group served to coordinate IDC development. In addition, and perhaps more importantly, it was decided to split the systems staff who conducted the software development into specialties. In past efforts, the systems staff would work on the entire "System" starting at the beginning, and usually running short of time at the end. The systems staff patterned themselves after the System they were developing; they each became small "modules" developing discrete units of the System. From the beginning, people were placed within the larger framework and one person was identified to provide the interconnection between the small work units, and ensure that what was needed from each was available at the critical time. The person selected for this job was someone who was experienced on compensation systems projects and had been successful in coordination roles in the past.

IDC System Decisions

The first decision in designing the system was choosing the software in which to write the application. Traditionally, the Bureau of Labor Statistics used mainframe systems, even if the up-front data collection occurred on a PC. For the new system we wanted a system that would function solely in a PC/server environment. The data capture system is a windows-based Power Builder system with data being fed into a relational SYBASE database residing on a UNIX server. We chose Power Builder because Power Builder is a powerful and flexible graphical user interface (GUI) development tool. In addition, Power Builder has powerful database interface tools that work with both Watcom and SYBASE. This made it possible to develop the IDC without having to write separate data access routines for our similarly designed SYBASE and Watcom databases. SYBASE was chosen because it is a sophisticated relational database management system that meets our requirements for relational integrity, security, performance, and reliability. Finally, because we have staff collecting data all over the country, we needed a database tool that would work well on laptops. Watcom was chosen for this because it runs in DOS rather than UNIX and therefore is a better match for the PC/laptop environment.

IDC is a system that is used for data collection, data editing, and data transmission. It is not a sampling or data processing/publication system. These systems will be developed later. Four separate modules comprise the system:

- The data capture module which allows for the entry of establishment and occupation data.
- The transmission module lets the user move schedules to and from the main server.
- A utilities function is used to backup and restore data, and
- An assignment module is used to make data collector assignments. Each of these modules accesses either a local PC Watcom database or the central SYBASE database.

What Does IDC Do?

IDC is a windows based system, and thus, has recognizable features such as the ability to open multiple windows, maximize and minimize windows, and cut and paste portions of text. In addition, it is a visual package that makes use of many icons and pull-down menus that let the user navigate from screen to screen.

After the sample is selected, some basic information about each establishment is loaded into IDC. This includes a unique identification number for each unit in the survey, the name of the company, and any information from the sample that will help during data collection or data processing. The remaining data are entered by the data collector during or after the interview.

The data capture module of IDC is composed of 5 main sections: establishment information; an occupation selection calculator; occupation information; leveling information; and wage information. Each section is briefly described below.

- The establishment data section of the system is designed for the capture of establishment data such as the usability of the unit, Standard Industrial Classification (SIC), and employment. In addition, the administrative aspects of the unit -- such as date of collection, interview time, and method of collection -- are captured on this screen.

- Probability selection of occupations is a disaggregation technique for selecting the occupations within each establishment for which we will be collecting data. Because this is a random process, the data collector must perform a calculation to determine which occupations to select based on the establishment employment and the number of occupations that are required (statistically) for an establishment of its size. In the past, the data collector performed the calculation on paper using a hand-held calculator. In IDC, the system will perform the calculation. In most cases the only required entry is the total employment count for the establishment.

- The occupation section of the system captures the list of occupations that have been selected, along with certain occupational characteristics -- such as occupational employment, the census occupation code (a code that the Census Bureau uses to categorize the occupations of individuals as recorded by the decennial Census), whether the occupation is full-time or part-time, and whether it is union or non-union.

- One objective of the OCWC compensation survey is to determine the level of duties and responsibilities of each occupation. The office has developed a set of leveling criteria patterned after the Federal Point Factor Evaluation System. The technique, called Generic Leveling, is a way of easily leveling any job that is (randomly) selected in each establishment. To level a job, there are ten "factors" that must be measured for each job. Knowledge, complexity, supervisory controls, and physical demands are examples of factors. Each factor is further subdivided into a certain number of levels and corresponding number of points. The point total defines a particular generic level. The IDC generates a separate screen on which to record the level of every job. This screen has ten tabs that correspond to each of the ten factors. When a folder tab is chosen, a list of levels is presented. When the user makes a selection, the System generates the corresponding number of points. When all ten tabs are complete, the System totals the points to determine the corresponding Federal General Schedule (GS) level of the job and it displays the GS level on the screen.

- IDC is designed to accept wages in any form provided, including hourly, weekly, monthly, or annual wages. The System then changes the data entered into the form needed for publication -- typically hourly rates. To perform the transformation, the user enters hours and earnings. The System will calculate an hourly rate for each worker. In addition, the system will calculate an average hourly rate for each occupation. In many cases a worker receives a base rate of pay as well as some addition to the wage such as a commission. The system can also handle this type of wage calculation and not only calculates an hourly rate for the worker, but also produces a base hourly rate for each worker.

Unique Features of IDC

The final section of this paper highlights certain features of the Integrated Data Capture system that we in OCWC are particularly proud of. We feel these features make IDC a unique Federal government data capture system.

In the past all of the Office of Compensation and Working Conditions systems, even if PC based, were designed in such a way that data entry was difficult to key during an interview. The systems were designed to edit data at the same time as data entry was occurring necessitating that all fields be complete as the user progressed from screen to screen. Because the surveys often dealt with complex compensation data that the respondents did not typically have at their fingertips, the systems were difficult to use during actual interviews. IDC was built to overcome this problem. We wanted a system that could be used during the interview. While the Integrated Data Capture system is not an Expert System in that it does not lead nor help the data collector through the interview, it is designed for data capture during an interview. IDC permits the user to enable the System edits when convenient, rather than operate them automatically.

Because each occupation selected must be assigned a census occupation code, an electronic copy of the census codes and census titles is included in the system. By clicking on the field where the user enters the census occupation code, the system brings you to the census code list. Selection of a code simultaneously enters the code into the correct field and brings the user back into the census field. In the future the hope is to not only include census codes and titles on this list but to also include complete census definitions.

As a final example of unique aspects of IDC, in order to assist the user in leveling the chosen occupation, the system contains a prototype screen for generic leveling. These prototypes were developed using data from the typical factors for a given Federal General Schedule level. Once the user completes leveling all ten factors, they select a prototype tab. The system generates a graphical representation of what the user choices were versus what the prototype expected. Edits are generated when the user choices do not match the expected prototype.

Conclusion

While the Integrated Data Capture system was just a first step in the long process of developing all new and interconnected systems for OCWC, it went a long way in teaching us about successful systems development. From all indications the development process and design of the system were a success. OCWC was able to produce a system that functions well in terms of accuracy, usability, and speed for the data collection, editing, and transmission systems and were able to produce it on time. For this reason, we plan to continue to use the same development process, that is, splitting staff based on specialty, for our new sampling and processing systems as well as for the many additional data collection features still to be added to IDC. ■

Chapter 7

CATI-CAPI Technical

Chair: Fred Wensing, Australian Bureau of Statistics

Kevin Dooley

Timothy Triplett ◆ Beth Webb

Stanley E. Legum

Chapter 7

Questionnaire Programming Language (QPL)

Kevin Dooley, U.S. General Accounting Office

Abstract

The Questionnaire Programming Language (QPL) consists of a set of IBM/PC programs that automate many of the activities involved in gathering and preparing survey data for analysis. Using this software, complex computer-aided telephone interview (CATI) or data-entry programs can be written that are easy to use and provide a high degree of control over what information may be entered. Interviewers can be trained to use the CATI software in only minutes, and completed interview records can be edited quickly and accurately. Once a questionnaire program has been created, other QPL system programs can be used to automatically generate formatted questionnaire documents, Awk, QBasic, SPSS or SAS analysis programs; Lotus, dBase, or comma or tab delimited data files; or askSam text-based data files. ■

Chapter 7

Using a Parallel "CASES" Instrument to Edit Call Record Information and Remove Incorrect Interview Data

Timothy Triplett and Beth Webb
University of Maryland at College Park

Abstract

The Survey Research Center's "fix-it" program is a CATI (Computer-Assisted Telephone Interviewing) instrument that reads the same data files that are read by the main CATI questionnaire instrument. The Survey Research Center uses the Berkeley "CASES" software, which allows the CATI instrument writer to design multiple instruments that read the same data files. The term parallel instrument is the term used in the CASES documentation. The fix-it program works by correcting the following three types of errors. First, the fix-it program can change the recorded status of any call attempt or the interviewer ID number associated with any call attempt. For example, often interviewers will record an incorrect call disposition on one of their call attempts. With the fix-it program the person with access rights to the program can change any call attempt's call disposition. They can also change the interviewer ID number if the incorrect ID number is recorded. Interviewers sometimes use call back codes when they should be recording the call result as a refusal and also sometimes incorrectly code eligible households as ineligible.

In addition, sometimes interviewers simply make a data entry mistake. The fix-it program easily corrects these problems without having to manually edit the data file and fix-it also updates' the current status of the case by re-evaluating the entire history of call attempts.

Second, fix-it easily removes the information from the last call attempt. This is necessary when interviewers record call record information in the wrong case. After removing the last call attempt, the fix-it program re-evaluates the

Abstract (Cont'd)

case history to determine the current status of a case. While the removal of the last call attempt occurs less often than changing call attempt status, this feature is often combined with the final feature of fix-it, the removal of data. When an incorrect respondent is interviewed not only does the data need to be removed, but the last call attempt disposition must also be removed.

Third the fix-it program allows the supervisor to remove invalid interview data without having to edit the data file. Using the fix-it program to remove data prevents accidental deletion of both valid call record and respondent selection data. In addition, using the fix-it program ensures that all the invalid data is removed from the case.

There are a number advantages of using the fix-it program to edit sampling information CATI data files. Perhaps the most important is that it is easy and thus can get done in a timely fashion. Though just as important is that the fix-it program require no manual updating of a data file, thus safely updates and edits only those fields where a change is requested. Other advantages include the automatic re-evaluation and update of a case's call status. This automatic update is important for both keeping sample reports accurate and helping autoschedule programs accurately choose a sample to call.

Two other advantages of the fix-it program are first, it is easy to undo any changes, since the CASES software records key information in a history file. Second, the fix-it program is a generic program custom designed to work with the Survey Research Center's front end. The front end is the part of the CATI questionnaire instrument that stays the same from study to study. Thus, fix-it easily works for most SRC projects.

Using a Parallel "CASES" Instrument to Edit Call Record Information and Remove Incorrect Interview Data

Timothy Triplett and Beth Webb
University of Maryland at College Park

Introduction

The Survey Research Center's "fix-it" program was written using the Berkeley CASES (Computer-Assisted Survey Execution System) software. The fix-it instrument is used to edit data for CATI (Computer-Assisted Telephone Interviewing) studies. The three main functions of the fix-it program are editing incorrect sample disposition information, editing incorrect interviewer information, and removing survey data from interviews conducted with ineligible respondents.

The CASES software allows the instrument writer to design multiple instruments that read the same data files. These instruments are referred to as "parallel instruments" in the CASES documentation. The main questionnaire instrument is written and executed in the e-inst directory of a study. The fix-it instrument is usually written and executed in the e-inst2 directory. If there are multiple parallel versions of an instrument (e.g., different language versions), the fix-it instrument can be located in a different e-inst directory such as e-inst3 or e-inst4).

One advantage of writing the editing features into a parallel directory rather than in the main instrument, is that there is more control as to who is allowed to edit sample dispositions. Access can be limited, through the use of a password in the instrument, to those who are given responsibility for sampling issues. This reduces the likelihood of improper usage. When the fix-it program is executed the first screen asks for the user's identification and the second screen asks for that person's password. If the person does not have access rights to the fix-it instrument or an incorrect password is entered, execution is terminated.

Data-Editing Features

The first screen after the password screen offers the following:

- Option #1.--Change the status of a call attempt
- Option #2.--Remove the last call attempt
- Option #3.--Change the interviewer ID# for a particular call attempt
- Option #4.--Exit program, no change made.

Changing Disposition Codes

The first option allows the fixit program to change the disposition code for any call attempt. After a call attempt is completed the interviewer records the outcome status of that attempt using one of the outcome codes listed in Table 1. Supervisors review the interviewer's recording of call attempt disposition codes and sometimes feel the need to change the disposition code chosen by the interviewer. A common example is interviewers trying to hide refusals under the code of call back. Correctly coding a refusal call attempt is extremely important, since the call disposition code is the most important factor in determining how future call attempts are handled. In addition, one of the criteria of interviewer evaluations is their ratio of completed interviews to refusals (see Table 2); thus, it is important that the supervisor has the ability to assign the initial refusal to the correct interviewer.

Table 1.--Call Status Report

Code	Call Status	N	Percent
1	Completed Interviews	1,000	39.9
2	Partially completed interviews (Terms)	30	1.2
3	**Refusals**	**74**	**3.0**
4	**Call Backs (appointments)**	**65**	**2.6**
5	Home Recorders	10	.4
6	No Answers	21	.8
20	Language/Age/Health Problems	84	3.4
21	Finalized Partial completes	3	.1
22	2nd Refusal	174	6.9
23	Call Backs (finalized after 25 attempts)	35	1.4
24	Household no longer available	22	.9
25	Home Recorders (finalized after 25 attempts)	22	.9
26	Refusal (finalized after 25 attempts)	11	.4
30	No answer (after 20 attempts)	141	5.6
31	Non Working phone number	516	20.6
32	Non Household	297	11.9
98	**Possible ineligible respondents**	**1**	**.0**
		2,506	100.0

There are many other less common reasons why supervisors decide to change the call disposition code assigned by an interviewer. Some of these occur because of an interviewer keying error. Other sources of mistakes stem from interviewers misunderstanding how to properly code a call disposition. For example, an interviewer may code a call attempt as a respondent problem because the respondent is away for a week, whereas the proper code would have been to code the attempt as a call back for next week. In any event all mistakes by the interviewer in coding call dispositions need to be corrected so that future call attempts are handled correctly. In addition, corrections to the data file are needed so that the sample progress reports, interviewer progress reports and the final sample status report are accurate.

Table 2.--Cooperation Rate by Interviewer			
ID#	Completes	Refusals	Cooperation Rate Completes / Completes + Initial Refusals
1,009	33	8	80.5%
1,120	46	10	82.1%
1,121	27	10	73.0%
1,156	86	12	87.8%
1,160	45	22	67.2%
Column total	237	62	79.3%

The fix-it program not only allows easy editing of call disposition codes, it also updates all the sample status variables and call attempt counters affected by the change.
To change the call disposition code:

- Select Option #1, "change status of a call attempt."
- Enter the call attempt that is in error.
  (You will be told total number of attempts.)
- Enter new call attempt disposition.
  (The old disposition is shown at the top of the screen,
  and a list of possible disposition codes is displayed.)
- Verify the change.
  (The caseid #, the call attempt #, the old disposition,
  and the new disposition will appear in a confirmation screen).

There are safeguards built into this editing feature. For example, the status of a case can not be changed to a complete or partial complete when there are no interview data in the case. When there are data in the case and the status of the case is being changed to something other than a complete or a partial, a screen appears with the warning that the data will be deleted. The change must be confirmed at this point, in order for it to be made.

Removing Last Call Attempts

The second option, "remove the last call attempt", is a special case of editing an incorrect call attempt. This option is used when an interviewer either misdials a number or the interviewer dials correctly but then records the information in the wrong case-id. Often the interviewer will realize the error and report it immediately to the supervisor; other times it is caught when a supervisor attempts to verify the status of the case.

When the last call attempt is removed, the disposition of the case is changed to what it was on the call attempt previous to the last call attempt. To make this change:

- ❒ Select Option #2, "remove the last call attempt."

- ❒ Verify the change.
  (The caseid #, the call attempt #, and the disposition of the call attempt being removed will be displayed in a confirmation screen.)

Editing Interviewer IDs

The third option allows the editing of the interviewer identification number associated with any call attempt. The interviewer must record his or her identification number for each call attempt. Occasionally the identification number is miskeyed. Either the interviewer realizes the error immediately and reports it to the supervisor, or the error is discovered when the "Cooperation Rate by Interviewer" table (see Table 2) is reviewed.

The following steps are taken to change the interviewer ID number:

- ❒ Choose option #3, "Change the Interviewer ID# for a particular call attempt"

- ❒ Enter the call attempt which is in error
  (The total number of attempts will be displayed)

- ❒ Enter the correct interviewer ID number.
  (The current interviewer ID # will be displayed.)

- ❒ Verify the change
  (The caseid #, the current interviewer ID number, and the correct Interviewer ID number will appear in a confirmation screen.)

There is a disposition code for complete or partial interviews with possible ineligible respondents(see Table 1). This code is used when during or after an interview, the interviewer is given information that leads him or her to believe that the respondent is not eligible for the study (e.g,. the population is households and we've reached a respondent in a nursing home; or adults are being interviewed and the demographics section of the questionnaire indicates that the respondent is 17). The interviewer then codes the case as an interview (or partial interview) with a possible ineligible respondent. A case is automatically assigned this code if the respondent does not verify the phone number called at the end of the interview.

When the fix-it program is executed for a case with this code, the following options appear:

- ❒ ***No, Don't Clear Data -- But Return This Case to the Status It was Before It Became a Problem.***

- ❒ ***Exit the Program, No Change Made.***

- ❒ ***Yes, Clear the Data.***

If it is decided that the interviewer was incorrect and the respondent **is** eligible to be interviewed for the particular study, Option #1 is chosen and the case becomes a completed interview.

If the interviewer was correct and the respondent is **not** eligible to be interviewed for the study, Option #3 is chosen. This clears the data and returns the case to the disposition it had on the previous call attempt. Before the data are cleared, a confirmation screen appears displaying the caseid number and giving the same options as above.

Summary

The old method of editing sample disposition information data involved manually editing the physical data files, which was a very time-consuming and error-prone process. This was especially true when an early call attempt had to be changed, necessitating the editing of the data for each subsequent call attempt. The fix-it program always writes changes to the correct variables and all status and counter variables are automatically updated. This greatly reduces the chance of keying error and the amount of time needed to fix sample problems. When the status and counter variables are accurately and frequently updated, this improves the accuracy of programs which use these variables such as sample reports and auto-schedulers.

Another important feature is the portability of the program from study to study. The shell of the fix-it program can be written in as little as a week by someone who is familiar with the front-end. As long as the front end stores the sampling variables in the same location, it takes less than 10 minutes to update and prepare the fix-it program to be used for individual studies. ■

Chapter 7

A Computer-Assisted Coding and Editing System for Non-Numeric Educational Transcript Data

Stanley E. Legum, Westat, Inc.

Abstract

Westat has completed a number of large studies in which the basic data have come from thousands of high school transcripts produced by hundreds of schools. These data needed to be coded and combined into a common database. Some of the challenges presented by these studies have been:

- Transcripts from different schools have different formats and present different information;
- Courses with the same titles in different schools often have different content;
- Courses with similar content may have different names in different schools;
- Within a school, remedial and honors courses may be distinguished from regular courses of the same name by codes on the transcript;
- The number of student contact hours represented by one credit differs from school to school;
- Schools use different grading systems.

We will demonstrate two software tools developed for the transcript studies:

- The Computer Assisted Data Entry (CADE) system; and
- The Computer Assisted Coding and Editing (CACE) system.

Abstract (Cont'd)

The Computer-Assisted Data Entry System

The Computer Assisted Data Entry system was designed for use by clerks who enter data directly from the transcript onto forms appearing on their computer screens. The CADE system has the advantage of naturally guiding the clerk to look for the needed information wherever it may appear on the transcript and of providing a consistent means for entering it. The CADE system, which is written in Clipper, includes range and logic checks which are active during data entry. It also includes a provision for double keying by a second clerk. When the second keyer enters information different from the first keyer, the system gives the second keyer a message and provides an opportunity to change the entry or confirm that the second entry is correct.

The Computer-Assisted Coding and Editing System

The Computer-Assisted Coding and Editing system is designed for use by subject matter experts (in this case, curriculum specialists) in the process of mapping objects (high school courses) to a predefined classification system (in this case, the Classification of Secondary School Courses). Since subject matter experts are selected for their knowledge rather that their data entry skills, the CACE system minimizes the amount of keying that needs to be performed.

Clerks pre-key all the titles from a high school catalog into a file which is read by the CACE system. The system, which is written in Paradox, presents the coder with one course title at a time from a school, a suggestion list of classification codes which might match the course, and a window displaying the full text of the classification manual that applies to the currently highlighted suggestion. Coders are free to ignore the suggestions and browse the entire classification manual and select any applicable code. A system of "flags" lets the coder record specialized information about a course such as it being the first course in a sequence or a later course in a sequence or that it is taught off campus. Only valid codes and flags are accepted by the system.

A subsystem of the CACE system is used by coders to match course titles in course catalogs to the corresponding course titles on the transcripts. This process cannot be fully automated because of the sometimes idiosyncratic ways in which course titles are abbreviated on transcripts.

Summary

In developing solutions to the challenges of entering and coding complex educational transcript data, Westat has developed generalizable software that may be applicable to other content domains such as medical record abstracting or document cataloging. ■

Chapter 8

Statistical Techniques -- I

Chair: Alan Estes, Federal Reserve Board

Michael Scrim

Linda Simpson ◆ Henry Chiang ◆ Cathy Tomczak

Adeline J. Wilcox

Rethinking the Editing Algorithm for the Survey of Employment Payrolls and Hours

Michael Scrim, Statistics Canada

Chapter 8

Abstract

The original editing systems and methodology of 11 years was costly, rigid, and required much human intervention. As a part of the redesign of the Survey of Employment Payrolls and Hours (SEPH), a revamped editing system was created. It has two major components, the first is that it takes into account the factors of industrial detail, firm size and seasonality. How it does this is with "curved bounds" (Hidiroglou-Bertholot Bounds), allowing more variation in data for larger firms. The second part of the change, equally as important, was the use of a Score function, a tool used to rank all records with errors to allow resources to focus on the most severe cases. The end results have been a cost saving of around $260,000 annually, and the number of human interventions reduced to around 1,500 records a month versus 30,000 with the old system while maintaining data quality. ■

A Statistical Edit for Livestock Slaughter Data

Linda Simpson, Henry Chiang, and Cathy Tomczak,
National Agricultural Statistics Service

Chapter 8

Abstract

For the past four years, the National Agricultural Statistics Service has edited data from its weekly survey of livestock slaughter plants using a PC-based statistical edit system. The data consist of daily numbers of cattle, hogs, calves and sheep inspected by U. S. Department of Agriculture meat inspectors, as well as weekly live-weight and dressed-weight totals.

This interactive edit is based on a robust estimator, called Tukey's Biweight. Each plant's historical data are used to flag outliers, determine which species are normally inspected, determine if a pattern is typically followed (i.e., slaughter only on Saturdays), and impute for missing data. This allows a "custom" edit for specialty (i.e., veal) or very large plants, and frees up time to reconcile data problems not possible with the previous mainframe edit.

The previous system, which was keyed on a PC, but edited on a mainframe in batch mode, used a generalized edit system. The edit, available several hours later, flagged values if they differed more than a given percent from the plant's previous three week average or outside some predetermined range; and it did not impute for missing data. ■

Chapter 8

A CSFII Data User's Principal Components Analysis for Outlier Detection

*Adeline J. Wilcox, Beltsville Agricultural Research Center**

Abstract

In my analysis of Continuing Survey of Food Intakes by Individuals 1989-1991 respondents whose dietary folate intake exceeded the safe upper limit recommended by the Centers for Disease Control and the Food and Drug Administration, I found seven of 277 high intakes which appeared to me to be due to coding errors. These apparent errors are of two types; one intake where 217 grams of *Tang*-powdered concentrate was reported, evidently 1 cup beverage prepared with water miscoded as 1 cup powdered concentrate, and six intakes where *Kool-Aid* appears to have been miscoded as *Tang*. *Tang* is a folate-fortified food.

I investigated the usefulness of principal components analysis for detecting these outliers. I ran the SAS procedure PROC PRINCOMP on two sets of variables. First, intake of energy, cholesterol, carbohydrate, vitamins A and C, folate and iron, and age. Second, intake of energy, carbohydrate, vitamin C and folate and age.

All seven of these outliers could have been discovered by examining the top percentile of folate intake. Using the second, smaller set of variables, six of the seven outliers could have been found in the 0.16 percent of the data with the largest positive first principal components. The only female among these outliers cannot be detected in this principal components analysis combining data from both sexes.

*Present affiliation: U.S. Bureau of the Census

A CSFII Data User's Principal Components Analysis for Outlier Detection

*Adeline J. Wilcox, Beltsville Agricultural Research Center**

Introduction

I investigated the usefulness of principal components analysis (PCA) for detecting suspect observations in nutrient intake data from the U.S. Department of Agriculture's Continuing Survey of Food Intakes by Individuals (CSFII) 1989-91. This work is related to my effort to estimate the proportion of the U. S. population whose intake of dietary folates on any given day exceeded the safe upper limit recommended by the Centers for Disease Control (CDC) and the Food and Drug Administration (FDA). Among those with extreme intake of dietary folates, there are a few whose high values may be due to coding error. When the proper variables are selected, PCA can identify observations known to be suspect.

In the analysis I plan, the weighted numerator of the proportion will comprise those with dietary folate intake of at least 1000 μg. The denominator will be a weighted total of all survey respondents. For a conservative estimate of this proportion, I will remove observations with apparent coding errors from the numerator.

Dietary Folate Intake Assessed by USDA Food Consumption Surveys

The CSFII 89-91 endeavored to collect three consecutive days of food consumption data from persons sampled in the 48 coterminous states. USDA obtained daily nutrient intake totals for each respondent by using food composition data to summarize the information collected on the kinds and quantities of food individuals consumed. Nutrient intake from dietary supplements such as multivitamin tablets was not included in these totals.

I used final USDA in-house CSFII data. Public use data are available (U. S. Department of Agriculture, 1996a, 1996b). After I did this work I discovered the in-house CSFII 89-91 nutrient intake data have more digits right of the decimal than the public use data.

*Present affiliation: U. S. Bureau of the Census

In CSFII 89-91, 15,398 respondents provided a total of 39,696 days of dietary intake data over the three days they were surveyed. Of the 15,398 respondents, 232 reported dietary folate intakes of at least 1000 µg on at least one day. Both the CDC (Public Health Service, 1992) and the FDA (Department of Health and Human Services, 1993) have advised limiting total folate intake to less than 1000 µg per day. Because some of these 232 respondents exceeded this safe upper limit on more than one day they were surveyed, 277 of the 39,696 dietary intakes exceeded the safe upper limit for folate. See Tables 1 and 2.

Table 1.--Intakes Judged Complete by Survey and Day Surveyed

		Intakes Judged Complete		
		Dietary Folates		
Survey	Day	>= 1000 µg	< 1000 µg	Total
1989-91	One	116	15,076	15,192
	Two	89	12,281	12,370
	Three	72	12,062	12,134
	Total	27	39,419	39,696
1994	One	53	5,536	5,589
	Two	38	5,273	5,311
	Total	91	10,809	10,900

Table 2.--Days Excess Dietary Folate Intake Consumed by Survey

	Number of Days Respondents Reported Dietary Folate Intake of at Least 1000 µg				
Survey	Zero	One	Two	Three	Any
1989-91	15,166	195	29	8	232
1994	5,502	83	4	*	87

*Only two days surveyed in 1994.

The histogram in Figure 1 shows an estimate of the distribution of dietary folates intake among the U.S. population for 1989-91. Day One refers to the first of the three days individuals were surveyed. Observations from individual respondents have been weighted up to the population total. The dotted line at 1000 μg of dietary folate intake marks the safe upper limit recommended by CDC and FDA. The dietary folate intake distribution is right-skewed with a long tail extending well past the 1000μg mark. The maximum value came from a respondent who enjoyed fried chicken livers. In my estimate of the proportion of survey respondents with dietary folate intakes of at least 1000 μg, the numerator consists entirely of extreme values. Dietary folate intake exceeding 1000 μg is also a feature of the CSFII 94 data displayed in Figure 2.

Two Types of Apparent Coding Errors

To study the source of this excess dietary folate, I listed the food codes which contributed the most folate to each diet and caused total daily folate intake to exceed 1000 μg. Among these, I found two types of apparent coding errors.

The first food code, 925-4200 FRT FLVRD DRNK, HI VIT C, FROM DRY MIX, caused six of the 277 dietary intakes to exceed 1000 μg folate. Although the description for this food code indicates the fruit-flavored drink mix contains added vitamin C, it does not reveal that it is also fortified with folic acid. See box below. *Tang* is a folic acid-fortified food. It appears that fruit-flavored drink mix without added folate (*Kool-Aid*) may have been coded as fruit-flavored drink mix fortified with folic acid (*Tang*) by coding some beverages as 925-4200 instead of 925-4101.

925-4101	Fruit-flavored drink, made from powdered mix, with sugar and vitamin C added (Include Kool-Aid, Wylers, NS as to sweetner)
925-4200	Fruit-flavored drink, made from powdered mix, mainly sugar, with high vitamin C added (Include Borden's Instant Breakfast Drink, Keen, Tang Instant Breakfast Juice Drink)

For one intake, 217 grams of food coded as 292-0010 TANG, DRY CONCENTRATE, was reported. One cup of dry Tang powdered concentrate weighs 217 grams. While it is humanly possible to consume this quantity of Tang powder, it seems likely that 1 cup beverage prepared with water was miscoded as 1 cup powdered concentrate.

In all, seven intakes appear affected by these two types of apparent coding errors. The records with apparent coding errors are listed in Table 3. If I had access to the original data collection forms, I would have been able to verify whether these items were coded correctly or not.

Figure 1.--USDA CSF11 89-91 Day One -- Weighted Sample of 15,192 Males and Females, All Ages

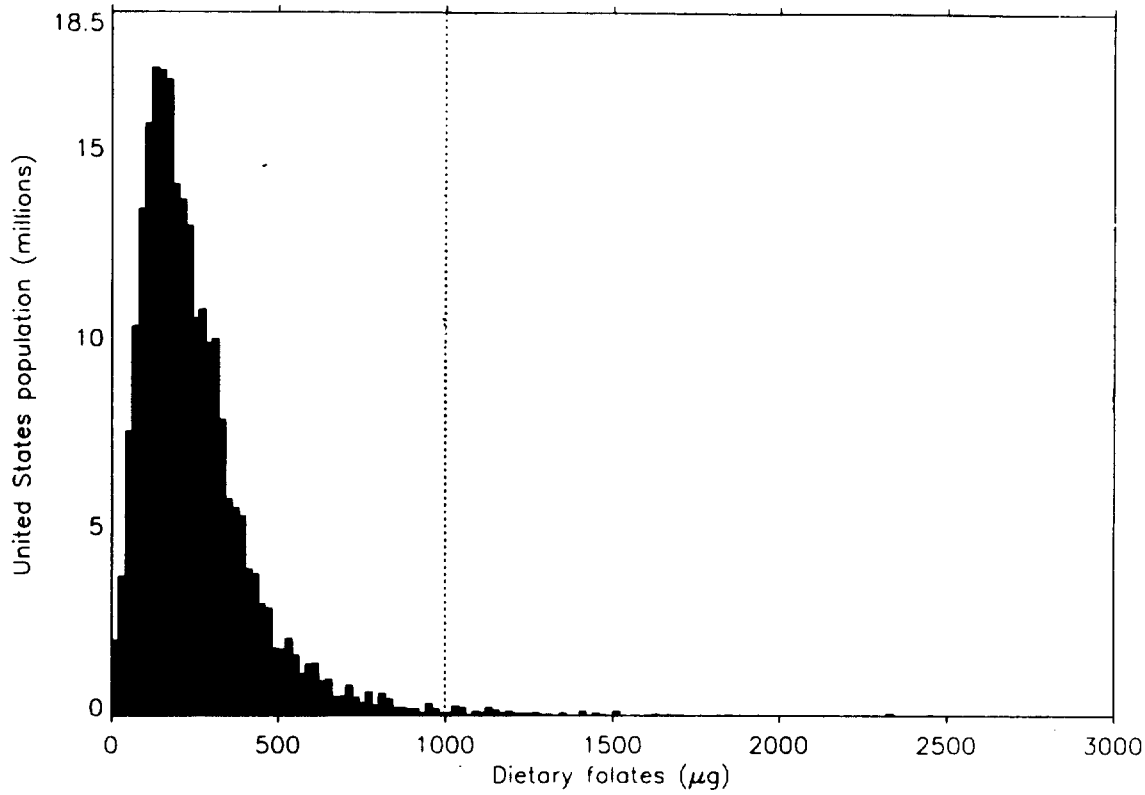


Figure 2.--USDA CSF11 94 Day One -- Weighted Sample of 5,589 Males and Females, All Ages

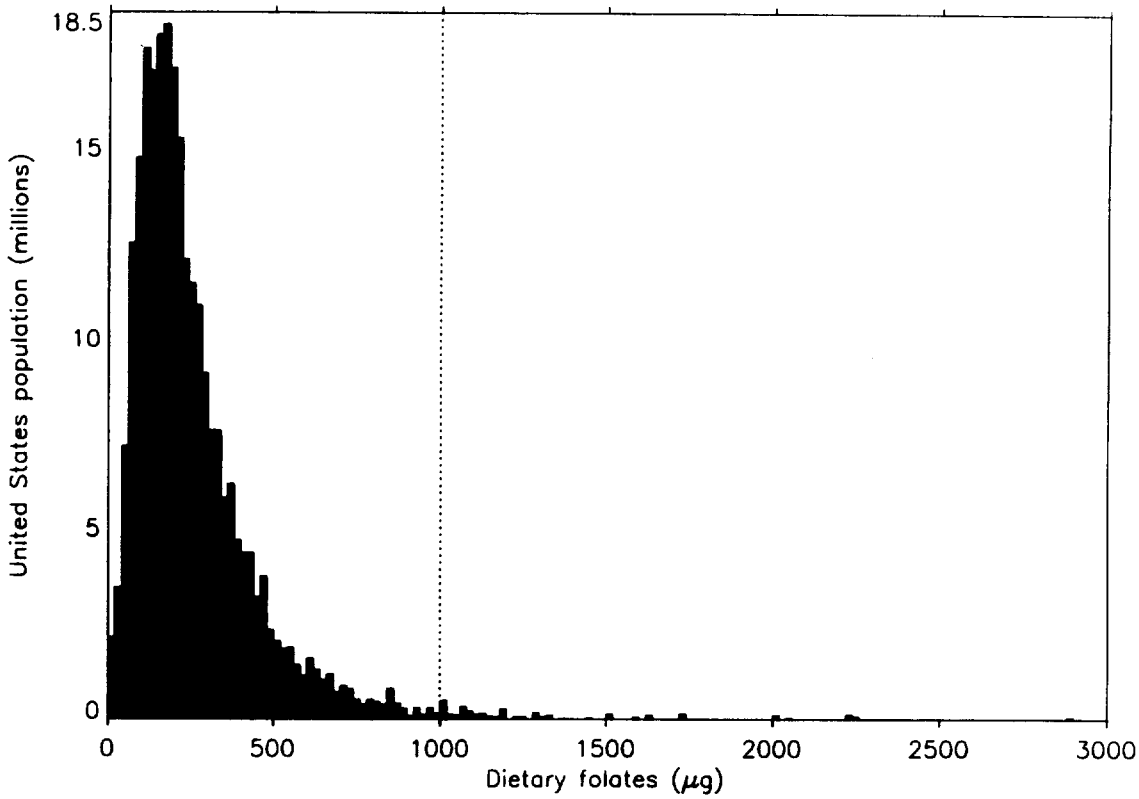


Table 3.--Suspect Observations, Their Principal Components and Ranks

ID	RESPNUM	DAYCODE	AGE	SEX	SOURCE	PRIN1OF8	RKPRN1_8	PRIN1OF5	RKPRN1_5	PRIN1OF4	RKPRN1_4	RANKFOLA
1116100	1	1	26	M	koolaid?	4.85739	229	6.9355	18	6.9360	18	33
1226079	4	3	13	M	powder	2.83278	576	4.7775	85	4.7728	85	137
2260877	1	1	65	M	koolaid?	4.00732	313	7.1605	17	7.1752	17	27
2260877	1	2	65	M	koolaid?	3.98032	317	7.4110	13	7.4260	12	22
2260877	1	3	65	M	koolaid?	7.21401	106	10.7103	2	10.7278	2	2
3117346	2	2	28	F	koolaid?	2.75879	606	3.3352	354	3.3348	355	262
3117346	3	1	11	M	koolaid?	2.65234	672	6.1076	28	6.1018	28	79

Principal Components Analysis

I investigated the usefulness of principal components analysis for detecting these seven outliers with PCA, a known statistical method for detecting outliers. In the Current Index to Statistics 1975-1993, I didn't find any entries on the use of PCA for editing nutrient intake data. Since the nutrient data base used to convert food consumption data to nutrient intake values contains no missing values (imputed values are used where data are not available), there are no missing values for respondents' daily nutrient intake totals, a great convenience for PCA. Nutrient intake data are particularly suitable for PCA because they are really collinear. I will explain collinearity in nutrient intake data later.

Selection of Variables for PCA

Inspecting my list of food codes contributing excess folate to respondents' diets, I learned that liver, folic acid-fortified cereal, and *Tang* were among the foods contributing to dietary folate intake in excess of 1000 µg. In my first attempt to search for outliers with PCA, I decided, based on my subject-matter knowledge, to use nutrients and food components for which liver and *Tang* are rich sources as variables. I didn't consider cereal because many brands are fortified with several vitamins and minerals. Liver is a rich source of iron, vitamin A, folate, and cholesterol, and a good source of vitamin C. Besides contributing carbohydrate to the diet, *Tang* is fortified with vitamin C and folic acid. Because nutrient intake is related to energy intake and age is related to energy intake, I also used energy measured in kilocalories and age measured in years as variables in my first look at the data with PCA. I selected eight variables for PCA: age, energy intake (KCAL), cholesterol (CHOL), carbohydrate (CHO), vitamin A (VITA), vitamin C (VITC), folates (FOLA), and iron.

Collinearity

The correlation matrix of variables selected for PCA reveals values for correlation coefficients greater than 0.5 between cholesterol and energy, carbohydrate and energy, iron and energy, folate and carbohydrate, iron and carbohydrate, and folate and iron. No relation between age and nutrient intake is apparent from the correlation matrix below.

	AGE	KCAL	CHOL	CHO	VITA	VITC	FOLA	IRON
AGE	1.00	-.03	0.02	-.05	0.09	0.02	0.04	0.02
KCAL	-.03	1.00	0.55	0.89	0.23	0.29	0.47	0.56
CHOL	0.02	0.55	1.00	0.32	0.24	0.09	0.21	0.26
CHO	-.05	0.89	0.32	1.00	0.23	0.35	0.51	0.56
VITA	0.09	0.23	0.24	0.23	1.00	0.24	0.45	0.36
VITC	0.02	0.29	0.09	0.35	0.24	1.00	0.49	0.28
FOLA	0.04	0.47	0.21	0.51	0.45	0.49	1.00	0.71
IRON	0.02	0.56	0.26	0.56	0.36	0.28	0.71	1.00

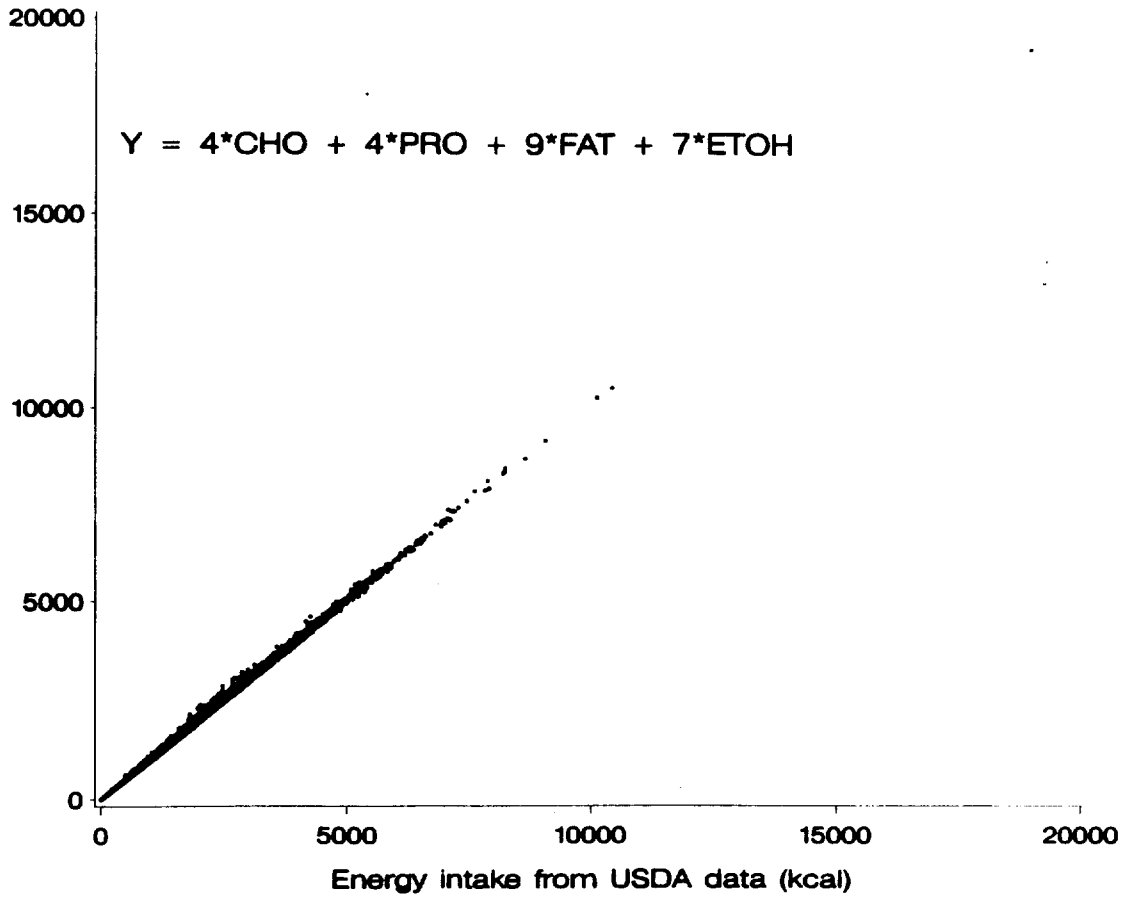
Nearly exact collinearity exists between energy intake and a linear combination of carbohydrate, protein, fat and ethanol intake. Figure 3 shows a linear combination of the macronutrient intake variables plotted against energy intake measured in kilocalories. This linear combination, Y, is energy intake computed from macronutrient intake, measured in grams. The units of the coefficients of this linear combination are kilocalories per gram. Most of the 39,696 points plotted in Figure 3 fall along a straight line at a 45 degree angle to the abscissa, indicating nearly exact collinearity. Note the outlying energy intake value of 18,955 kilocalories.

Principal components are orthogonal and rid me of collinearity.

Choice of Principal Components

I used the first principal component (PC) for detecting outliers, following the method I learned from Robert M. Hamer in a short course sponsored by SAS® Institute Inc. (Hamer, 1995). Johnson and Wichern recommend using the last few PCs for outlier detection (Johnson and Wichern, 1992). They illustrate use of scatter plots and Q-Q plots for finding an outlier in a small data set. This proved useless for my large data set. Plotting the seventh PCs against the eighth PCs left all seven suspect observations

Figure 3.--Evidence of True and Nearly Exact Collinearity



buried deep in the cloud of 39,696 points. With PCA on five variables, described later, I plotted the fourth PCs against the fifth PCs, but none of the seven suspect observations fell outside the dense scatter of points.

PCA Versus Checking the Top Percentile

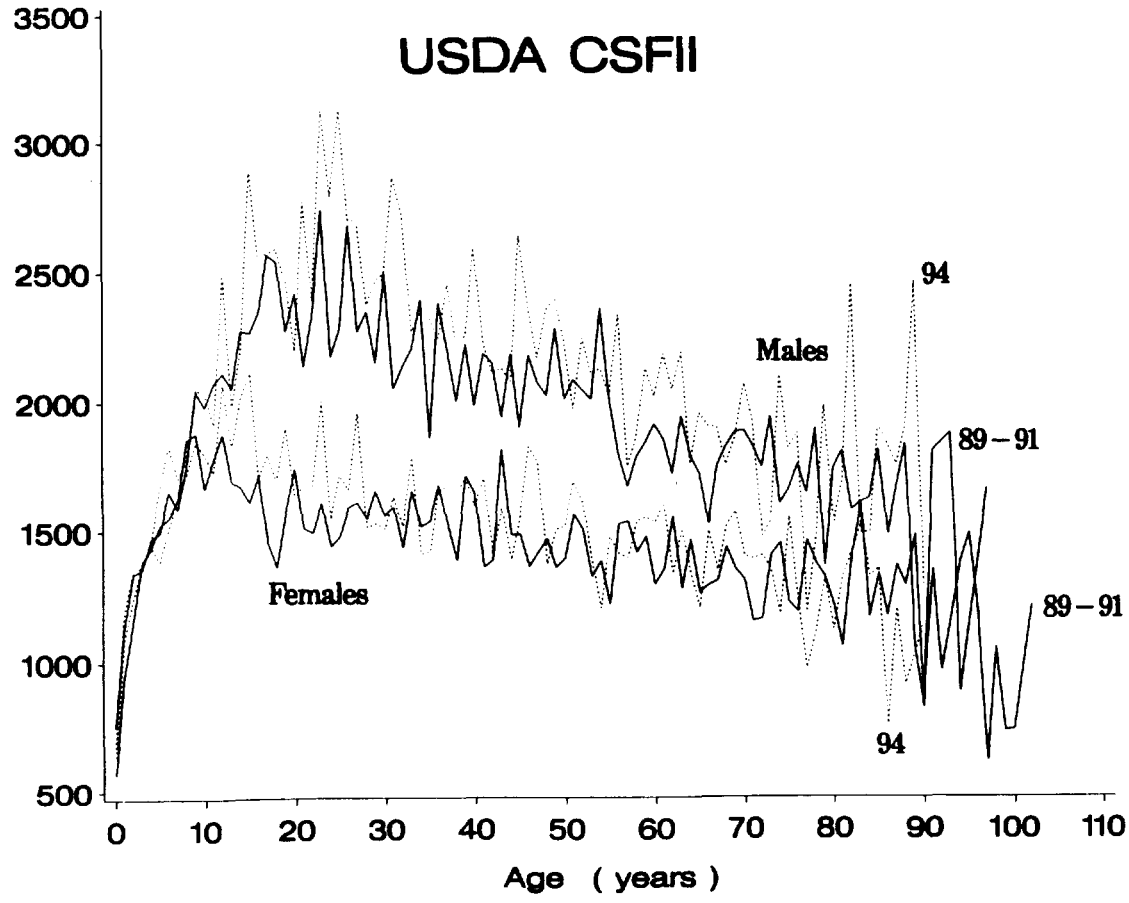
To be more useful than simply examining the top percentile of dietary folate intake, PCA should rank the seven suspect observations above the ranks obtained by sorting the data by descending dietary folate intake. In a column headed RANKFOLA, Table 3 shows the ranks assigned when the 39,696 records, sorted by descending dietary folate intake, are ranked. All seven of these outliers could have been discovered by examining the top percentile of folate intake.

Using the SAS procedure PROC PRINCOMP with the eight variables listed above, I got the first PCs listed in Table 3 under the heading PRIN1OF8. I ranked these by descending value as shown in the column headed RKPRN1_8. If one started checking data with the largest positive first PC and continued until all seven suspect observations had been examined, one would have looked at more than the top percentile of the data. Thus, PCA with these eight variables is not more efficient than checking the top percentile of dietary folate intake.

Since I was trying to identify outliers involving *Tang*, not liver, I eliminated nutrients or food components which liver, but note *Tang* is a rich source of, that is, iron, cholesterol and vitamin A. PCA with five variables results in PCs, PRIN1OF5, ranking the suspect observations in the 100 largest PCs except for the only female among them. Her first PC is ranked 354th. See the column headed RKPRN1_5 in Table 3. That this female outlier doesn't rise to the top of the data as well as the male outliers do when PCA is applied, suggests performing PCA separately on males and females.

At this point, I took another look at the correlation matrix and realized that age was not correlated with intake of any nutrient or cholesterol. However, I knew from experience with nutrient intake data that age and energy intake are related. Figure 4 shows median energy intake plotted against age for each age in years for both the CSFII 89-91 and CSFII 94. The median values for each gender are connected by solid lines for 1989-91. Dotted lines connect the median values for 1994. There is a relationship between age and energy intake but it is not a linear one. This plot shows the data used for detecting outliers, not the data used for assessing dietary intake. The difference between these data sets is the inclusion of the known energy intake of breast-fed infants. The medians for infants on this plot should not be used for nutritional assessment.

Figure 4.--USDA CSFII Median Day One Energy Intake by Age and Sex



Given the lack of a linear relationship between age and energy intake, I performed another PCA without the age variable. As shown by the columns headed PRIN1OF4 and RKPRIN1_4 in Table 3, neither the PCs nor their ranks changed much, indicating that age is not an important variable for detecting folate outliers with PCA.

Specificity of PCA

Laura Gillis inquired if I had looked at the specificity of PCA for outlier detection. I did not. In this analysis I studied only sensitivity. However, it is reassuring to note that, in the PCA using four variables, the largest PC belonged to the individual who reported consuming 18,955 kilocalories.

Conclusion

PCA shows promise for editing nutrient intake and food consumption data.

Acknowledgments

I wish to thank Will Potts and Gordon Marten for giving me the opportunity to learn more about multivariate methods of statistical analysis, Robert M. Hamer and Julie A. Smith for looking over this paper and making helpful comments, and my statistician friends for their nontechnical support.

References

Department of Health and Human Services, Public Health Service (1992). Recommendations for the Use of Folic Acid to Reduce the Number of Cases of Spina Bifida and Other Neural Tube Defects, *Morbidity and Mortality Weekly Report*, 41/No. RR-14.

Department of Health and Human Services, Food and Drug Administration (1993). Food Labeling: Health Claims and Label Statements: Folate and Neural Tube Defects, 58 *Federal Register* 53254-75.

Hamer, R. M. (1995). *Multivariate Statistical Methods: Practical Applications Course Notes*, Cary, NC: SAS Institute, Inc.

Johnson, R. A., and Wichern, D. W. (1992). *Applied Multivariate Statistical Analysis*, 3rd ed., Englewood Cliffs, NJ: Prentice-Hall, Inc.

U. S. Department of Agriculture (1996a). *Continuing Survey of Food Intakes by Individuals and the Diet and Health Knowledge Survey 1989-91* (on CD-ROM), Accession No. PB96-501747. Springfield, VA: National Technical Information Service.

U. S. Department of Agriculture (1996b). *Continuing Survey of Food Intakes by Individuals 1994* (on CD-ROM), Accession No. PB96-501010. Springfield, VA: National Technical Information Service. ■

Chapter 9

Case Studies -- II

Chair: Ann Hardy, Centers for Disease Control

Clifford Adelman

Jimmy Hwang ◆ Bo Kolody ◆ William A. Vega

Susan S. Jack

The Thin Yellow Line: Editing and Imputation in a World of Third-Party Artifacts

Clifford Adelman, U.S. Department of Education

Chapter 9

Abstract

Each of the longitudinal studies of the National Center for Education Statistics includes a file of transcripts from colleges, commmunity colleges and trade schools attended by survey participants. The transcripts are gathered at about age 30, coded by a contractor, and delivered to NCES. Given the idiosyncratic record-keeping practices of 2,500 institutions (in the most recent collection) and inconsistencies in coding of graduate students who usually do not know what they are looking at, the delivered files are a tangle of contradictions.

The editorial process takes 12-15 months to complete, and is carried out with interagency support from the National Science Foundation.

This paper both reports and demonstrates what has been learned from the editing of two such samples, the development of decision rules, and the feedback of the decision rules into the initial coding process. More importantly for data quality and standards, the paper demonstrates where the line between editing and imputation lies in such an archive, and how the survey data guides the editor in making (rare) decisions to impute with respect to key variables.

The case in point is the most important variable for student records, namely, the credential/degree earned.

The Thin Yellow Line: Editing and Imputation in a World of Third-Party Artifacts

Clifford Adelman, U.S. Department of Education

Introduction

The tasks of editing and the occasions of imputation in data sets using third-party documents raise an ontological and epistemological issue in the same breath. When one imputes a phenomenon, one asserts its existence; and it is legitimate to ask for measures of confidence that the phenomenon, in fact, exists. The very process of imputation implies that the phenomenon did not emerge *ex nihilo*, rather is dependent or derived. The epistemological challenge lies in the identification of qualities of other, known phenomena on which the imputation depends or from which it is derived. The strength of those qualities and the logic of derivation determine how you know the phenomenon. The greater the strength of those qualities and the longer and more stable the histories of relationships among the variables described in the data set, the less likely you are guessing when information is missing. The less likely the guessing, the more the task is dominated by "editing" and the less by "imputation."

On the other hand, as the strength of these qualities diminishes to near zero, the greater the leaps of faith. At some point, the chasm between a phenomenon and its potential representation is so wide that only imagination can cross it. As much as we value imagination in the history of civilization, its place in data set construction is rather limited.

Editing is always involved in imputation, since the process of data editing identifies the missing. But it is difficult to describe the point on a continuum at which the balance of editing and imputation tips toward the latter. The task is somewhat akin to establishing a passing score on a "high stakes" test such as a licensure exam. A great deal of empirical evidence is assembled, and replication of results with different populations at different times is a necessary procedure. Both test publishers and test users wish to minimize the cases of false positives and false negatives. Minimize. A passing score is a guide, not an absolute. The passing score is like that thin yellow line down the middle of a country road: you want to keep the right traffic going in the right direction on either side.

In some data sets, e.g., those that produce the CPI, there are high stakes consequences of false imputations. Even though individuals are not being judged with the same consequences as a licensure examination, individuals are directly affected by the CPI. In other national data sets used in the course of policy-setting by states, institutions, and organizations, the stakes are not as high, and individuals are more indirectly affected. While these data sets are constructed and used to estimate aggregates, the direct and indirect effects on individuals argue that whenever we approach the mixture of editing and imputation, confidence levels are critical standards.

Third-Party Artifacts: the Historian's Stuff

The mass of data editing is conducted on the second-party level. That is, information is collected directly from subjects or their agents. The mechanism is a survey or an unobtrusive measure that is unmediated. I have a questionnaire, you are interviewed, and your responses are directly transcribed or transcribed by an interviewer who follows very standard and tight protocols. In the language of data editing, these protocols function at the capture stage. Or you engage in a discrete activity, such as filing for unemployment benefits, that leaves a direct, unmediated trace.

A third-party artifact involves a different order of evidence. Archaeologists, anthropologists and historians know it well. You engage in activities that are recorded for *sui generis* purposes. They are recorded in formats and symbolisms that can best be described, in Geertz's phrase, as "local knowledge." They are inscribed on documents we can call "artifacts." At some future time, these artifacts are discovered, collected from many sources, and re-recorded in a standardized format by a third-party. The original artifact is thus twice removed from the form in which it appears in a database. Depending on the bureaucratic protocols of the collection, the data can be edited in either a coding or post-coding phase.

Take, for example, the debarkation lists of boats arriving on the eastern seaboard in the National Period of our history. In the 1820s and 1830s, customs agents in Charleston, Baltimore, Philadelphia, New York, and Boston recorded information on the nature and destination of arriving immigrants. In no two ports was there a standardized form for doing so. Sometimes we got full names; sometimes not; sometimes gender, age, occupation, relationships and ethnicity; sometimes not -- or, in the case of the Irish, negative ethnic stereotypes and a sorting based on skin color. Data for key variables are always missing. There is no one port for which they are complete.

There is one exception to the missing variables: the name of the ship and its arrival date.

If we are building a modern database from these lists, we have a phenomenological choice: we can accept the classifications made by the customs agents as reflecting the views/perceptions of the customs agents -- in which case, we'd be writing a database that is more about the customs agents than the immigrants; or we can look for ways to fill in the information. We cannot imagine information of this type. And it is very difficult to impute.

But if we examine the port records on the other side of the Atlantic, we find that the embarkation lists are often more detailed than the debarkation lists. The level of detail was particularly high in ports such as Hamburg and Rostok. The Library of Congress possesses some of this material. For a complete examination, the investigator takes the name of the ship and checks the registries until the port of origin can be determined; goes to the port of origin, rummages around the archives, and finds the manifest. Sweat, toil, tedium. Nice travel, but hardly suited to an instant electronic environment, and not in the habits of data teams that have to release the monthly *Consumer Price Index* at 8:30 on a Thursday morning, hot-decking the price of laundry detergent in Seattle on the basis of analogous products.

It took Fernand Braudel and his associates 20 years to write *The History of the Mediterranean in the Age of Philip II*, meticulously hunting down meteorological data for the entire basin covering a period of a millenium, let alone records of caravans and harvests. We could save Braudel a lot of time today. The question is whether we want to and how. For our task, in many respects, would be just as tedious. Third-party documents are often like that.

The Case of College Records

We all generate unobtrusive records in our lives, records that become part of the grist for analyses of economic, social, and public health issues and trends. The case discussed here is that of college transcripts, a type of record now generated for over half the adult population between the ages of 19 and 35. One of our jobs in the research and statistics division of the U. S. Department of Education is figuring out what goes on in that vast and sometime amorphous enterprise called U. S. higher education, and whether, where, and how our individual and collective investments in higher education have convincingly measurable impacts on our worklives, citizenship, and adult development.

To determine what goes on, we could always perform a content analysis of college catalogues. Unfortunately, these documents tend to be higher education's contribution to American fiction and ought to be placed on appropriate shelves in the library. Or we could ask individuals, in the course Computer-Assisted Telephone Interviews (CATI) or paper/pencil interviews, what types of education they pursued after high school, in what kinds of institutions. That strategy, as we've discovered, leads to what can euphemistically be described as exaggeration--but, as we shall see, not always, at least in the matter of degrees earned.

In terms of what students study, we can examine enrollment surveys conducted by learned and professional societies. What we discover quickly, though, is that enrollments are not students. Rather, they represent the same students cycling themselves through many courses within the field(s) covered by the enrollment survey. No learned society will admit that fact because each is concerned with getting the maximum share of what we ex-deans call "enrollment mix." Having eliminated enrollment surveys, we can try course schedules. Like the catalogues, these at least show what was really offered. Unfortunately, the evidence of course schedules does not indicate whether enough students registered for a course to make it a "go." I made the George Washington University course schedule look fairly interesting a couple of semesters in a row with a course on quantitative historical methods that used those immigration lists, focused on women's roles, and wound up with a collective class project using the first *Women's Who's Who* (1916). Exactly one person signed up both times and the course was cancelled.

So we turn to transcripts. They don't lie, they don't exaggerate, they don't forget. But they are a mess. And they are even more a mess because, to arrive in a national database, they are re-coded in a standardized form by graduate students working for a contractor. Graduate students are supposed to be smart; but in the matter of the documents at issue, they are not fluent in the histories of the variables nor experienced in translating the oftimes idiosyncratic formats and signs used on those documents. Put more simply, they have little idea of what they are looking at. The editor's job is to spot and fix their errors.

For certain tasks, such as coding courses into over 1,000 course categories from an empirically-derived taxonomy, we gave the graduate student coders the assistance of "search strings." Given the existence of certain words in the title of the course, the search string presents the coder with a range of possibilities for coding. The coders choose. But the problem with search strings is that they do not provide decision rules for context. The coders either choose incorrectly or resort to residual categories for "unknowns" on 20 percent of the entries.

For example, if they are presented with a title such as "Composition and Conversation" in the junior year of a student at a selective college and code it as English Composition, I will wager they are wrong. In that situation, I look for context and derivation, and will immediately scan for the foreign language courses on the transcript that may set the context. If the title were merely "Composition," I would look for music or studio art as a guide to correct coding.

This example illustrates the editorial process, post-coding. If the context determines that "Composition and Conversation" most likely applies to a Russian language course and I recode it according, I am not changing the reality, rather making the mark of reality "fit" or "represent" the reality more accurately than the form in which the mark was delivered. But course titles such as "TEN BADTAB TEN" or GREEN BOX WORKSHOP or RAGS TO RICHES or THE GOOD LIFE or (yes) GOOD BOOKS, I would rather leave alone.

National Samples, Unique Institutions

We have taken two national samples of college transcripts in the course of longitudinal studies. I have edited both of them, and, in each case, the editing process took two years. The samples are very robust. The first (known as the NLS-72) involved 12,600 students, 19,500 transcripts, and 485,000 courses. The second ("High School and Beyond/Sophomore Cohort") was smaller, but has taken no less time: 8,400 students, 13,300 transcripts, 320,000 courses. For each course, there are 18 variables to which I must pay attention. For each transcript record, another 10. Into this mix, I can import other variables from CATI interviews, paper and pencil surveys, and high school records. The purpose of importing is to guide the decision-making process in determining the accuracy of data coding and entry.

While much common-sense guides decision rules, specialized knowledge is absolutely necessary. When the coders read, on an MIT transcript, "Math 1," and code it as a remedial course, you wonder how much common sense can be impaired. But when they read a sequence from an engineering student at, let us say, Wisconsin, who has Calc 1, 2, 3, 4 and they code all of those courses as Calculus, you can forgive their ignorance. There is a big difference between elementary functions and infinite series, and that difference is important for understanding the careers of engineering students. A lot of engineering deans and advisers want to know. The data editor cannot be a copy editor, rather someone who has to know a great deal about how specific colleges, community colleges, and trade schools work. This knowledge sets up a web of dependencies on which the editing decisions rest.

In Search of Accuracy: Consulting the Source

Unlike researching debarkation lists in the early 19th century, we have another choice with contemporary data bases: we can call the source. Given the uniqueness of institutions, accuracy in editing requires contacting their registrars to assist in interpretation. Third-party data from similar contemporary sources allow for such a procedure. Surveys of the current or recent recorded activities of individuals in health facilities using different record-keeping systems would be a good analogue.

Following the contractor's delivery of the tape for the High School & Beyond college transcript file, we made a list of schools where there appeared to be a great deal of inconsistency and contradiction in matters of credits, grades, dates, and course titles. There were 700 schools out of 2,500 where these problems were rampant. We telephoned them. It took three months to get the guidelines. Sometimes, the registrars didn't know the answers to our questions. After all, colleges are in a market, and try to

grab a niche, finding any way they can to be unique. The evidence of such dubious niches include complex credit value systems or academic calendars of such a nature that students probably need pacemakers to tell them when to go to class.

Why is accuracy important and how much should one sacrifice accuracy for timeliness? We get bad legislation if we are not accurate. The Student Right-to-Know Act (1990) is a premier example. Under this act, colleges are required to report rates of graduation (or, in the case of community colleges, persistence). We understood the problem with this legislation: over half the undergraduates in this U.S. attend more than one institution, and (as it turns out), more than 20% change institutions across state lines (rendering it impossible for any state higher education authority to track them). The student may start in a college in South Carolina but graduate from a college in North Carolina. The first school is penalized by the propaganda of published graduation rates under Student Right-to-Know. And the cost to both institutions to produce information that few parents or students actually use exceeds the benefits.

But we were not there in time to testify on this legislation because the data sets were riddled with errors. Time to degree; average credits to degree. State legislators deserve--at the least--an accurate national tapestry that provides some norms. If they don't get it, or if the data are sloppy, somebody will suffer. Our accuracy is also critical to interpretation of labor market data. I have asked field interviewers from the Census Bureau and the Bureau of Labor Statistics how they know when the person who--given a reasonable demographic profile--says he/she is a doctor is, in fact, a physician? A data editing system within our framework would inquire whether, according to the evidence of transcripts, the person possessed the requisite educational credentials and history. If these credentials are missing, the person is probably either a physician's assistant or some other kind of "doctor," and their occupation code should be edited accordingly.

The Mediator and Code 590

Again, unlike the 19th century debarkation list case, we can also go back to the third-party when the same reality is represented in both literal (unmediated) and symbolic (mediated) form. The third-party is responsible for the mediated form that is delivered on tape as a database. When we don't understand the symbol, we can ask for the literal.

For example, in looking at the occupations of individuals in the NLS-72 database when these people were 32/33 years old (in 1986), I saw that a substantial number received the occupation code "590." The coding manual, a collection of symbols used in the mediating process, did not list "590." I telephoned the contractor and asked them what "590" signified. After a pause, the response came back: "Craftsmen in the Military." This was a strange response, particularly as 65% of the people in the "Code 590" bin held bachelor's degrees (the evidence came from the transcripts). I then asked the contractor to send "the literals," the direct transcriptions of what the respondent wrote (or said, if the data collection method was CATI) on the questionnaire. It turned out that one-third of the respondents assigned to Code 590 were active-duty military, one-third were civilian employees of the Department of Defence, and one-third belonged in occupation/industry categories that had nothing to do with the military.

The correct way to represent the occupation/industry of an accountant at Bolling Air Force Base is occupation=accountant and industry=U.S. military. Unfortunately, the coding scheme used by the mediator did not distinguish the military from the most aggregate notion of government employer. The

results of the Code 590 inquiry thus included not only the corrections for hundreds of miscoded cases, but also a recasting of the "industry" variable to disaggregate the military from civilian government agencies.

Errors and Imputation

In these brief accounts of contexts overriding search strings, consulting the sources, and sending the "literals" all I have described is editing. Imputation did not enter these transactions.

What is a true imputation in this business, and where does the limit lie? We are invited to impute, I contend, only in those cases that cannot be labelled "errors" or the result of errors, and that involve missing information.

Let us illustrate with variations on the most critical issue in the review of national college transcript samples: whether a student received a degree. Figure 1 is an actual page from the computerized records of a student as we created those records from the tape delivered by the mediator. The page does not represent the entire record for this particular student, but contains enough information to illustrate the case.

Figure 1.--Excerpt From Sample Student Transcript Records

```
IS System

NT ID=237
nued)

 CLASS  TRANSNUM  FINDISP  STARTYR  TERMONTH  TERMTYPE  CRSECODE  CRSENAME                CRSEID  GENCRED  CRSGRADE  CRSGRADN  DEGREE1  DEGYEAR1  DEGMON1  DEGREE2  DEGYEAR2  DEGMON2  MAJCODA1
    32        1        1       85         1         1    230101  MD LIT TRANSLATION        58       0         H        -9        1        -1       -1       -1       -1       -1       999
    32        1        1       85         1         1    420101  INTR TO PSYCHOLOGY        59       3         C         2        1        -1       -1       -1       -1       -1       999
    32        1        1       85         1         1    500903  JAZZ ENS STG BAND         60       1         C         4        1        -1       -1       -1       -1       -1       999
    32        1        1       85         1         1    500903  CHORALE                   61       2         C         2        1        -1       -1       -1       -1       -1       999
    32        1        1       85         1         1    300401  HUMANITIES                62       3         B         3        1        -1       -1       -1       -1       -1       999
    32        1        1       85         1         1    500903  UNIV BAND                 63       1         A         4        1        -1       -1       -1       -1       -1       999
    32        1        1       85         1         1    131203  LOWER SECONDARY           64       1         B         3        1        -1       -1       -1       -1       -1       999
    32        1        1       85         1         1    500903  UPP SEC SAXOPHONE         65       2         A         4        1        -1       -1       -1       -1       -1       999
    32        1        1       85         6         1    500906  RENAISSANCE MUSIC         80       2         A         4        1        -1       -1       -1       -1       -1       999
    32        1        1       85         6         1    230601  ENG LANG DESCR STD        81       3         C         2        1        -1       -1       -1       -1       -1       999
    32        1        1       85         6         1    230801  BRITISH LIT OF IMAG       82       3         B         3        1        -1       -1       -1       -1       -1       999
    32        1        1       85         6         1    230801  REN THEMES IN VICT        83       1         A         4        1        -1       -1       -1       -1       -1       999
    32        1        1       85         9         1    400101  FOUND PHYS SCI            27       3         C         2        1        -1       -1       -1       -1       -1       999
    32        1        1       85         9         1    450802  CULTURE TO 500 AD         28       3         B         3        1        -1       -1       -1       -1       -1       999
    32        1        1       85         9         1    230701  AMER LIT I                29       3         A         4        1        -1       -1       -1       -1       -1       999
    32        1        1       85         9         1    230802  SHAKESPEARE               30       3         B         3        1        -1       -1       -1       -1       -1       999
    32        1        1       85         9         1    500903  CHOIR                     31       1         B         3        1        -1       -1       -1       -1       -1       999
    32        1        1       85         9         1    500903  JAZZ ENSEMBLE             32       1         A         4        1        -1       -1       -1       -1       -1       999
    32        1        1       86         1         1    450801  CULTURE-500-1650          66       3         B         3        1        -1       -1       -1       -1       -1       999
    32        1        1       86         1         1    272001  MATH IN MOD WRL           67       3         F         0        1        -1       -1       -1       -1       -1       999
    32        1        1       86         1         1    110101  INTRO COMPUTER SCI        68       0         H        -9        1        -1       -1       -1       -1       -1       999
    32        1        1       86         1         1    230701  AMER LIT II               69       3         A         4        1        -1       -1       -1       -1       -1       999
    32        1        1       86         1         1    380201  SIGN/SYMBOL/MYTH          78       3         A         4        1        -1       -1       -1       -1       -1       999
    32        1        1       86         1         1    230301  CONT LITERATURE           71       3         B         3        1        -1       -1       -1       -1       -1       999
    32        1        1       86         1         1    500903  CONCERT CHOIR             72       1         C         2        1        -1       -1       -1       -1       -1       999
    32        1        1       86         6         1    270201  MATH IN MOD WRL           84       3         C         2        1        -1       -1       -1       -1       -1       999
    32        1        1       86         6         1    450818  AM CV SINCE 1865          85       3         A         4        1        -1       -1       -1       -1       -1       999
    32        1        1       86         9         1    240103  LOGIC-CR THINK            33       2         A         4        1        -1       -1       -1       -1       -1       999
    32        1        1       86         9         1    160905  ELEM SPANISH I            34       4         B         3        1        -1       -1       -1       -1       -1       999
    32        1        1       86         9         1    500903  JAZZ ENSEMBLE             35       1         A         4        1        -1       -1       -1       -1       -1       999
    32        1        1       86         9         1    110101  INTRO COMPUTER SCI        56       0         H        -9        1        -1       -1       -1       -1       -1       999
    32        1        1       86         9         1    500601  A STUDY OF FILM           57       3         A         4        1        -1       -1       -1       -1       -1       999
    32        1        1       87         1         1    231001  RHET/PUB. SPEAK           73       2         A         4        1        -1       -1       -1       -1       -1       999
    32        1        1       87         1         1    450101  POL/ECON SYS-THE          74       3         C         2        1        -1       -1       -1       -1       -1       999
    32        1        1       87         1         1    904000  SENIOR SEMINAR            75       2         A         4        1        -1       -1       -1       -1       -1       999
    32        1        1       87         1         1    239999  PRACT. IN ENGL            76       3         A         4        1        -1       -1       -1       -1       -1       999
    32        1        1       87         1         1    160905  ELEM SPANISH II           77       4         B         3        1        -1       -1       -1       -1       -1       999
    12        2        1       87         9         1    131206  TCHNG FR COMP           1081       3         B         3        1        -1       -1       -1       -1       -1       999
    12        2        1       87         9         1    230801  SMNR-19 C BR AESTHET    1082       3         B         3        1        -1       -1       -1       -1       -1       999
    12        2        1       88         1         1    231301  INTRO TO GRAD STUDIES   1083       3         C         2        1        -1       -1       -1       -1       -1       999
    12        2        1       88         1         1    230801  SEM BRIT LIT 19TH C     1084       3         B         3        1        -1       -1       -1       -1       -1       999
    21        3        1       82         6         1    900000  TRANSFER COURSES         139      -1         TR       -8        1        -1       -1       -1       -1       -1       999
    21        3        1       82         9         1    500903  INTRO TO MUS STUDY        86       4         B         3        1        -1       -1       -1       -1       -1       999
    21        3        1       82         9         1    500903  UNIV BAND(PERMREQ)        87       1         B         3        1        -1       -1       -1       -1       -1       999
    21        3        1       82         9         1    500903  JAZZ ENS STAGE BND        88       1         A         4        1        -1       -1       -1       -1       -1       999
    21        3        1       82         9         1    230402  GRAMMAR & COMP            89       3         B         3        1        -1       -1       -1       -1       -1       999
```

The contractor interviewed the student (the process is carried out prior to and separate from the gathering of the transcript). The student said she had a bachelor's degree. She listed the three schools she attended. We requested transcripts from the schools and received all three of them. The student has 135 credits on Transcript #1 from a liberal arts college (CCLASS=32). Neither a degree nor degree date are indicated, despite what appears to be a decent academic record and a course entitled, "SENIOR SEMINAR" in the year one would expect a 1982 high school graduate to be receiving a college degree. There is also a graduate school transcript (#2). The editorial process spots all these characteristics, and considers the student's claim to a degree against the missing information about the degree in the third-party presentation.

The evidence allows us to impute a bachelor's degree, a major in English, and a degree date of May, 1987 (her last term of undergraduate attendance was a semester that began in January of 1987, she entered graduate school in September of that year, and schools with semester systems hold commencements in May). There are virtually no degrees of freedom in this imputation. Our confidence level is very high.

On the continuum of balance between editing and imputation, there is more of the former than the latter in this case. Why? There are three historical relationships between the evidence and the receipt of a bachelor's degree that are strong enough to say that what appears to be "missing" is more the result of oversight (a form of "error"): entry to graduate school in an academic discipline; numbers of credits earned; and senior seminars in a pattern of courses with a dominant field (English literature) and an ostensible GPA comfortably at or above the norm.

How do I know that similar cases will produce similar results? The editor of a data set based on third-party artifacts has an ethical obligation to examine a sample of original artifacts in cases where such critical signs are missing. For the High School & Beyond college transcript sample, I looked at over 100 records where degrees were in question. In 70 percent of these cases, the degree was, in fact, indicated on a transcript, usually on the back side of a page the data entry person never turned over. That evidence was sufficient to justify similar imputations.

Consider a second variation in the case of our 1987 English Lit major: what if there was no graduate school transcript, and all we had were two fragmentary records with poor grades, remedial work and lots of withdrawals? No matter what degree the student claimed she had earned, there would be no clues, nothing from which the claim could be validated with confidence. The decision to leave the third-party presentation of the record alone is wholly an editorial decision: there are no errors.

And a third variation: what if the degree-bearing transcript was submitted by the college as a "blocked transcript"? A blocked transcript indicates only the degree, degree date, and major. No course work, credits, or grades are included. The "blocked transcript" is the student's choice, and we must respect that choice. At the same time, our data standards indicate that it is irresponsible to present degree data without course and credit data. So these data, which are missing but not as a result of error, are imputed.

In this imputation, we employ what lawyers call "custom and usage" guidelines, that is, special knowledge of organizational behavior and rules in colleges and universities. We know the student earned at least 120 semester credits (the accepted -- and empirical -- minimum for a bachelor's degree), a degree in English (with, by custom, at least 30 credits in the major), that the degree was awarded in May, 1987 and that the student graduated from high school in the spring of 1982. Furthermore, the

student has told us, in the surveys, what she was doing (school, work, other) for every month since 1982. We can thus enter blocks of courses and credits, by term, e.g.,

IMPUTED UNDERGRAD COURSEWORK 1985 SPRING 15 CREDITS

IMPUTED MAJOR COURSEWORK 1985 FALL 15 CREDITS.

The former entry receives a special course code for imputed courses of unknown discipline. The latter entry receives the course code, within a field, that covers unknown, missing, and residual cases. As for grades, only "CR" is entered, along with a flag that tells the software to add the credits but not to include the course in the computation of GPA. These entries are put in a flat file which, when completed, is appended to the master data set.

In entries such as these, imputation outweighs editing, but must be limited to variables and values that can be asserted with confidence. How do we know that the balance has shifted? The phenomena did not exist previously in the universe of representations we call a database and could not be created algorithmically by reference to existing signs in that data base. At the same time, they are not imaginary phenomena: they are dependent and derived, and hence are clearly on the imputation side of that thin yellow line.

Invitation

Exploratory papers usually invite readers to further research and reflection. This occasion is no exception. To develop theory and guidelines for imputation in data sets built from third-party artifacts requires investigations in a variety of fields. Recent economic history, public health, and mass communications are all inviting areas. Of these, only public health has been compelling enough to produce systematic data collection of national scope. It is obviously the next terrain for charting the thin yellow line. ■

Sampling Design and Estimation Properties of a Study of Perinatal Substance Exposure in California

Jimmy Hwang, University of California (San Diego),
Bo Kolody, San Diego State University, and
William A. Vega, University of California (Berkeley)

Chapter 9

Abstract

During the period of 1992-93, a group of researchers from UC Berkeley, San Diego State University, the State Department of Alcohol and Drug Programs, and the University of California at San Diego conducted a comprehensive survey on substance abuse problems among pregnant women in California. A fully probabilistic stratified cluster sample was used to estimate the prevalence of perinatal drug exposure for the state of California. Included in the sampling plan were 29,494 pregnant women presenting for delivery in 202 hospitals, which were sampled from 602 hospitals throughout the state of California. Urine specimens were taken from women presenting for delivery and later linked by code number to demographic variables, tobacco use and prescribed drug data gathered at intake. Urine specimens were then shipped and tested at a NIDA certified lab. Based on the survey results, the study further projected that there were about 67,361 perinatal exposures to one or more non-prescribed drug, including alcohol, and 52,346 exposures to tobacco in California.

The findings of the study have many significant clinical and public health implications. The purpose of the presentation is to offer several practical experiences in data editing and exposition from the study. The discussion will be useful in the area of applied statistics and the implications of sampling survey. Since the study was the first of this kind in substance abuse programs and in sampling targeted subjects, the presentation will illustrate its innovative and unique sampling design of the study. The presentation will also present the problems and solutions, advantages and disadvantages of employing certain weighting procedures in adjusting the estimates. A detailed description is provided about the sampling process, its rationale, sampling factors, statistical estimation, and computational procedures.

Sampling Design and Estimation Properties of a Study of Perinatal Substance Exposure in California

Jimmy Hwang, University of California (San Diego),
Bo Kolody, San Diego State University, and
William A. Vega, University of California (Berkeley)

Introduction

A study was conducted according to a multistage probability sampling design to estimate the prevalence of perinatal substance exposure in California in 1992. The study used coded urine samples from 29,494 women presenting for delivery in 202 hospitals, screened for toxins; and later linked the results of toxicology by code number to the subjects' demographic variables and their reported use of tobacco and prescribed drugs. The study reported the survey results by age, marital status, county of residence, ethnicity and prenatal care history for state-wide and regional estimates. The findings have many significant clinical and public health implications (Vega et al., 1993a and 1993b). This paper presents a general discussion of the sampling process and statistical design of the study. The presentation is useful in the area of applied statistics and the implications of sampling survey.

Sampling Process and Its Rationale

The most important considerations governing the choice of the sampling design in the study involved several factors.

The sampling frame be representative of all births taking place in maternity hospitals in California. The practical constraint limited the study to pregnant women admitted to maternity hospitals at time of delivery. To ensure sampling efficiency and study feasibility, the study included only hospitals that had more than 10 births annually, and excluded federal hospitals, hospitals on military bases, hospitals that delivered babies on an emergency room only basis, and birthing centers. Births in these hospitals account for a proportion of about 2 percent of statewide births. Their exclusion would not bias the estimates.

The sampling procedure be fully probabilistic and thus yield population estimates with known sampling errors. Sampling of study subjects within a hospital was not based upon subject characteristics (e.g., race/ethnicity). The study defined all admissions within a specified time frame of March through October in 1992 as study subjects. This course of action necessitated a large sample size in order to have sufficient women from all ethnic groups enter the sample through a process of natural selection.

The study attempted to estimate regional differences in California. Since approximately 80% of births are in the ten counties with the highest number of births, a key objective of the study was to derive separate prevalence estimates for each of these ten counties as well as for the remaining forty-eight counties as aggregated into sampling strata. The strata design that was used conformed to geo-

administrative county clusters previously established as Health Service Areas (HSAs). Although the possibility of using HSAs was considered, they were too numerous (N=26) for acceptable confidence intervals of subgroups. The county and county clusters (i.e., sampling strata), based on fourteen HSAs and the ten largest counties by birth population in California, constitute meaningful geographic divisions, for which separate estimates are both feasible and desirable. As a result, there were 11 counties (out of 58) that did not have a hospital in the study. Two of them had a hospital that did not want to participate. The others either did not have a hospital, or their one hospital was not randomly chosen. The stratum design allows for the absence of any one hospital from any one county in a stratum by allowing for grouping of results from hospitals in the other counties in the stratum. Based on the stratum design used in the study, the results obtained for any stratum can be generalized for any county in the stratum.

The sampling frame for the study was 583,487 births (98 percent of births statewide) in approximately 305 maternity hospitals in California. There were 21 sampling strata in the study, one for each HSA and large county and combining these 21 sub-samples into a single statewide sample. The sampling strata vary considerably in terms of their number of births.

For example, Los Angeles county recorded 206,457 births while Imperial county recorded 2,777 births. Given these wide disparities, the objective was to draw a sample sufficiently large for reasonably precise estimates in the small sampling strata while allowing for large n's in large sampling strata. A strictly proportionate to size allocation would render the larger strata too large or the smaller strata too small.

For Stratum #1.00, for example, the sampling fraction is n=1107 while in the largest sampling stratum (#11.00) n=4879. This strategy yielded an overall, statewide sample size n=29,200. Given the substantially larger sample size in Stratum #11.00 (Los Angeles county) the prevalence estimates were approximately twice as precise as in Stratum #1.00. This higher precision is desirable inasmuch as Los Angeles county accounts for about one third of statewide births. When combining stratum prevalence rates into a statewide rate, the higher precision for Los Angeles county, as well as for other large strata, minimized the impact of weighting, which was used to adjust for stratum size. Table 1 gives the actual numbers for each stratum.

Two-Stage Probability Sampling

Within each stratum hospitals form the clusters, the first stage sampling units. The method of systematic sampling was used to select hospitals within the stratum. A separate prevalence estimate could be made for each of the 21 strata. The size of each of these samples was proportionate to the number of births in the stratum.

The selection procedure began by listing hospitals ordered on ownership type. Ownership was used in order to assure that a representative proportion of women entering every type of hospitals was included in adequate numbers for the sampling design. Within type, hospitals were ordered on the annual number of births. A systematic sampling procedure was used to delete every third hospital from this ordered hospital list. Across the 21 strata, this procedure yielded a sample of 202 hospitals that were included in the study.

Table 1.--Target and Actual Samples in the Study

Stratum #	Number of Births	n=Target Samples to be Collected	Actual Number of Usable Samples	Counties
1.00	10,135	1,107	1,173	Northern CA
2.00	9,013	512	546	Golden Empire
2.34	20,969	1,358	1,429	Sacramento
3.00	12,710	1,253	1,337	North Bay
4.00	8,174	510	540	West Bay
4.38	14,589	1,198	1,198	San Francisco
5.00	11,401	1,333	1,267	Contra Costa
5.01	22,757	1,676	1,639	Alameda County
6.00	21,701	1,178	1,172	N.San Joaquin
7.00	31,380	1,575	1,657	Santa Clara
8.00	13,188	1,129	1,194	Mid Coast
9.00	20,036	1,154	1,158	Central CA
9.10	15,681	1,141	1,063	Fresno County
10.00	17,458	1,144	1,097	Ventura/S.B.
11.00	206,457	4,879	4,918	L.A. County
12.00	4,200	101	0	Inyo/Mono
12.33	22,813	1,241	1,278	Riverside County
12.36	28,434	1,547	1,530	San Bernardino Co.
13.00	53,678	2,682	2,714	Orange County
14.00	2,777	183	203	Imperial County
14.37	45,936	2,299	2,381	San Diego County
Totals	593,487	29,200	29,494	

The number of subjects sampled within selected hospitals was set to be directly proportionate to the number of births, specifically, the proportion of stratum births during the 1990-1991 fiscal year. To adjust for any disproportion due to slight over or under sampling by hospital size or ethnicity, weights were applied to conform the outcomes to the 1991-92 parameters on ethnic distributions in each hospital. To achieve the statewide estimates, appropriate weights were also applied to adjust for disproportion by stratum to conform the total sample to the statewide 1991-92 distributions on race/ethnicity by hospital and stratum.

Anonymity and Urine Testing

Subjects were selected in a manner designed to minimize selection bias and to ensure the anonymity of those from whom urine specimens were obtained. Starting on a given day, nurses were instructed to test all admitted patients until the sampling fraction for the hospital was met. Nurses collected urine specimens and basic descriptive information from each subject and recorded this information on a code sheet that contained no personally identifying information. The same code number was used on the code sheet and the urine-specimen label.

In accordance with national standards of nursing care, all patients in California hospitals are asked at the time of admission whether they currently smoke. There is no standard phrasing for this question. The answer does not indicate the frequency of extent of smoking. Information on patients' smoking was recorded on the code sheets; it was missing for 6.4 percent of the subjects in the sample. The direction of bias, if any, due to the missing data could not be judged.

Procedures and safeguards were established to ensure that no personally identifying information could be linked to the results of urine testing. The research protocol was reviewed and approved by the Committee for the Protection of Human Subjects, Health and Welfare Agency, State of California and by the Human Subjects Review Committee of the University of California, Berkeley, as well as by institutional review boards at the individual hosptials. The urine specimens were sent to a laboratory certified by the National Institute of Drug Abuse (PharmChem, Menlo Park, California). Table 2 lists the tests performed and the detection periods for each substance.

Each specimen was assayed by personnel who did not know its origin or the subject's demographic characteristics. They used enzyme-multiplied immunoassay techniques for all drugs and an enzymatic assay for alcohol (Test materials for these assays were manufactured by Syva, a subsidiary of Syntex, Palo Alto, Califorina.). If a screening test was **negative, i.e., failed to detect the presence of drugs or alcohol, the test results were reported to be negative and no further testing was conducted. If an enzyme-multiplied immunoassay was positive for any drug except a cannabinoid (marijuana), the result was confirmed by gas chromatography; this technique is commonly used by analytical toxicologists to confirm the presence of drugs or their metabolites in urine or other biologic fluids. If an assay was positive for a cannabinoid, the result was confirmed by high-performance thin-layer chromatography; the result of this technique correlates highly with that of gas chromatography, and the sensitivity of the test is comparable. When an immunoassay is combined with an appropriate confirmation assay that is chemically independent, the likelihood that a drug will be correctly identified is greater than 99 percent.

The pharmacological characteristics of alcohol differentiate it from other drugs. Its concentration in blood, breath, and urine can be estimated according to body weight if the amount ingested and the time elapsed are known; alcohol is rapidly absorbed into the bloodstream and distributed throughout body water. If an average-sized person in good health drinks 6 oz of beer (170 ml), 2 oz of wine (60 ml), or 0.5 oz of distilled spirits (15 ml), urine collected 1 to 1.5 hours later should contain at least 10 mg of alcohol per deciliter, a level used as the cutoff value for the study. A person who has one or two drinks at night and urinates after awakening in the morning would have a lower urinary alcohol level and not test positive. A person who consumes at least 1 oz of distilled spirits 2 to 2.5 hours before urine collection would test positive. Alcohol cannot be measured accurately in urine specimens containing glucose; to avoid confounding, the study considered such specimens to be negative for alcohol.

Statistical Estimation

A separate prevalence and errors estimate based on two-stage sampling design was made for each of the 21 strata. Calculation of standard errors took the sampling design into account by weighting values to conform the data to the distribution of births for the period 1991-1992 within hospitals, within regions, and statewide. The 95 percent confidence interval for each subgroup reflected the observed percentage of positive tests and the standard error of the estimate. To facilitate subgroup comparison, differences between proportions were tested, with the Student-Newman-Keuls ranges adjustment (a two-tailed test) for multiple comprisons, in analyses of prevalence according to racial or ethnic group and the duration of prenatal care. The comparison of the prevalence of tobacco use included the exact t-test value and probability for each substance.

Table 2.--Drug and Alcohol Testing Procedures

Drug	Screening Method	Cutoff	Method	Confirmation Cutoff	Detection Period
Alcohol	EA	10 mg/dl			
Ethanol			GC	10 mg/dl	Very short*
Glucose			GC	1 mg/dl	
Amphetamine	EMIT	1000ng/ml	GC	300 ng/ml	
Amphetamine			GC	300 ng/ml	12-72 hr
Methamphetamine			GC	300 ng/ml	12-72 hr
Barbiturates	EMIT	200ng/ml	GC	200 ng/ml	
Amobartital			GC	200 ng/ml	2-4 days
Butalbital			GC	200 ng/ml	2-4 days
Pentobarbital			GC	200 ng/ml	2-4 days
Phenobarbital			GC	500 ng/ml	up to 30 days
Secobarbital			GC	200 ng/ml	2-4 days
Benzodiazepines	EMIT	200ng/ml	GC	200 ng/ml	
ACB			GC	500 ng/ml	up to 30 days
MACB			GC	500 ng/ml	up to 30 days
Cannabinoid	EMIT	50ng/ml	HPTLC	50 ng/ml	
THC metabolite			HPTLC	50 ng/ml	up to 14 days**
Cocaine metabolite	EMIT	300ng/ml	GC	500 ng/ml	
Benzoylecgonine			GC	500 ng/ml	12-72 hr
Methadone	EMIT	300ng/ml	GC	300 ng/ml	
Methadone			GC	300 ng/ml	1-4 days
Opiates	EMIT	300ng/ml	GC	20 ng/ml	
Codeine#			GC	500 ng/ml	2-4 days
Hydromorphone			GC	1000 ng/ml	2-4 days
Morphine			GC	200 ng/ml	2-4 days
Phencyclidine	EMIT	25ng/ml	GC	200 ng/ml	
Phencyclidine			GC	200 ng/ml	up to 14 days**

Note: ACB denotes acetylbenzophenone, MACB Methylacetylbenzophenone, THC tetrahydrocanabinol; EA denotes Enzymatic assay, EMIT enzyme-multiplied immunoassay technique; GC denotes gas chromatography, and HPTLC high-performance thin-layer Chromatography.

The advantage of two-stage sampling is obvious in this study. We have the opportunity of obtaining some smaller number of observations that apear more efficient and hence produce more precise estimates. In our study, we first used systematic elimination to select n hospitals in each stratum. Then, for each selected unit, a random method was given for selecting the specified numbers of subjects from it. In finding the mean and variance of the prevalence estimate, averages were taken over all samples that were generated by the process. To calculate the average, we first averaged the estimate over all second-

stage selections that were drawn from the n hospitals. Then we averaged over all possible selections of n hospitals by the study.

For an estimate, p, the method can be expressed as

$$E(p) = E_1[E_2(p)],$$

where E denotes expected or average value over all samples, E_2 denotes averaging over all possible second-stage selections from a fixed set of hospitals, and E_1 denotes averaging over all first-stage selections.

For the variance V(p), it is readily shown the following result (Cochran, 1977),

$$V(p) = V_1[E_2(p)] + E_1[V_2(p)],$$

where $V_2(p)$ is the variance over all possible subsample selections for a given set of units.

To illustrate the computational implications, for any subject in a hospital H, say y_{iH}, let y_{iH} be 1 if the subject possesses substance of our interest and 0 otherwise. Let n_H be the number of sampled subjects in the hospital H and N_H be the number of births in the hospital H. Then, the prevalence of a given substance for the hospital H, p_H, is defined as:

$$p_H = \frac{\sum_i y_{iH}}{n_H}$$

The prevalence of a given substance for the stratum A, p_A, is defined as:

$$p_A = \frac{\sum_H \sum_i W_A y_{iH}}{\sum_H n_H}$$

where W_A is the weighting value based on probability of sampling allocations of the hospitals for each stratum.

Alternatively, the prevalence for the stratum A may be calculated as:

$$p_A = \frac{\sum_H W_A n_H p_H}{\sum_H n_H}$$

This two-stage probability sampling involves two sources of the variance for the prevalence of the stratum: variance between hospitals and variance within hospitals. Although the weighting value is omitted from the following derivation of the variance, the weighted prevalence and weighted standard errors were used in the study.

For each stratum, the variance between hospitals may be defined as:

$$s_1^2 = \frac{\sum_H (p_H - P_A)^{\wedge}2}{(n_{sA} - 1)} * \frac{(1 - \frac{n_{sA}}{n_{tA}})}{n_{sA}} ,$$

where n_{sA} is the number of sampled hospitals and n_{tA} is the number of the total hospitals within the stratum.

The variance within hospitals may be defined as:

$$s_2^2 = \frac{\sum_H n_H(p_H q_H)}{n_{sA}(n_H - 1)} \frac{(\frac{n_{sA}}{n_{tA}})(1 - \frac{n_H}{N_H})}{N_H n_{sA}} .$$

Therefore, the variance of the prevalence is the summation of the variance between hospitals and the variance within hospitals, i.e., $s_1^2+s_2^2$.

The standard error of the prevalence is bounded (at 95% level) by $1.96*(s_1^2+s_2^2)^{0.5}$.

A Computational Example

As a computational example, the Table 3 gives the prevalence and error estimate of using alcohol for African-American women at time of delivery in stratum #11.00 (Los Angeles county). To derive the error estimate, we need to have the following information for each selected hospital: prevalence (p_H), sampled subjects (n_H), and total births (N_H) (see column 2-4). We also need the following stratum information: the number of selected hospitals (n_{sA}), the number of total hospitals (n_{tA}), and stratum prevalence (p_A) (see column 5-7).

Computational Steps

- Calculate the squared differences between sample means and population mean (i.e., p_H and p_A). (Column 8)
- Divide Column 8 by (n_{sA}-1) to calculate the sampling variances. (Column 9)
- Adjust the sampling variances by probability factor, $(1-n_{sA}/n_{tA})/n_{sA}$, to yield the variances between hospitals. (Column 10)
- Calculate the product of $n_H(p_H)(1-p_H)$. (Column 11)
- Divide Column 11 by n_{sA}(nH-1) to yield the sample variances for each hospital. (Column 12)
- Adjust the sample variances by probability factor, i.e., $\frac{(\frac{n_{sA}}{n_{tA}})(1 - \frac{n_H}{N_H})}{N_H n_{sA}}$. (Column 13)

Table 3.--A Computational Example: Alcohol Use of African-American Women in Los Angeles County

HID (C1)	PS1 (C2)	NSAMP (C3)	BTOTAL (C4)	NSHOSP (C5)	NTHOSP (C6)	PO1 (C7)	(C8)	(C9)	(C10)	(C11)	(C12)	(C13)
133	28.571	12	3330	50	77	11.689	2.85	0.06	0.00	244.90	0.45	0.00
134	0.000	3	1254	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
136	50.000	2	860	50	77	11.689	14.68	0.30	0.00	50.00	1.00	0.00
146	12.698	47	5330	50	77	11.689	0.01	0.00	0.00	521.04	0.23	0.00
151	8.947	63	2947	50	77	11.689	0.08	0.00	0.00	513.25	0.17	0.00
164	0.000	4	1048	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
168	0.000	1	1102	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
174	4.546	9	1300	50	77	11.689	0.51	0.01	0.00	39.05	0.10	0.00
177	0.000	1	4329	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
180	50.000	3	1939	50	77	11.689	14.68	0.30	0.00	75.00	0.75	0.00
182	0.000	1	1473	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
183	0.000	1	1763	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
187	0.000	1	1391	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
191	7.143	31	8309	50	77	11.689	0.21	0.00	0.00	205.61	0.14	0.00
195	66.667	5	1708	50	77	11.689	30.23	0.62	0.00	111.11	0.56	0.00
196	0.000	22	5534	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
204	6.849	27	2249	50	77	11.689	0.23	0.00	0.00	172.26	0.13	0.00
205	4.546	20	4229	50	77	11.689	0.51	0.01	0.00	86.78	0.09	0.00
207	0.000	1	973	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
213	50.000	3	2584	50	77	11.689	14.68	0.30	0.00	75.00	0.75	0.00
217	10.526	10	5475	50	77	11.689	0.01	0.00	0.00	94.18	0.21	0.00
219	19.231	12	14578	50	77	11.689	0.57	0.01	0.00	186.39	0.34	0.00
223	18.605	36	8225	50	77	11.689	0.48	0.01	0.00	545.16	0.31	0.00
227	13.636	24	5248	50	77	11.689	0.04	0.00	0.00	282.65	0.25	0.00
228	23.684	8	554	50	77	11.689	1.44	0.03	0.00	144.60	0.41	0.00
229	0.000	1	710	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
233	0.000	1	955	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
237	0.000	2	1500	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
240	0.000	1	2275	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
254	0.000	7	2269	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
261	0.000	1	1889	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
263	0.000	1	2457	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
265	0.000	5	2966	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
267	0.000	6	1322	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
275	0.000	1	662	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
277	0.000	1	2677	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
278	16.667	6	2430	50	77	11.689	0.25	0.01	0.00	83.33	0.33	0.00
279	0.000	1	589	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
280	0.000	2	1843	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
287	15.625	70	4922	50	77	11.689	0.15	0.00	0.00	922.85	0.27	0.00
289	0.000	1	1698	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
290	0.000	4	2609	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
297	0.000	15	4120	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
303	0.000	4	2266	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
305	0.000	1	1198	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
309	0.000	2	1600	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
310	0.000	1	1174	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
315	9.091	5	5412	50	77	11.689	0.07	0.00	0.00	41.32	0.21	0.00
702	0.000	1	952	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00
729	0.000	5	2263	50	77	11.689	1.37	0.03	0.00	0.00	0.00	0.00

Variance between hospitals = 1.7750 Variance between hospitals = 0.0667

Standard Error Estimate = 1.7750 + 0.0667 =1.8417

- The variance of the two-stage sampling is the summation of Column 10 and Column 13. The square root of the sum produces the standard error.

- The error estimate at the 95 percent confidence level is multiplying 1.96 by the standard error.

Results and Discussion

The total prevalence for any category of drug takes into account whether the women presenting for delivery had any drug administered prior to urine collection. The nurse filling out the data form was asked to check any category of drug that was administered. Those categories were Opiates, Benzodiazepines, Barbiturates, Other, and None. If a urine tested positive for any drug in any of these categories and if the nurse checked any of these categories as a previously administered drug, then the urine was not counted in the prevalence total for that category.

The Illicit Drug prevalence total was for any drug positive for the following substances: THC, Cocaine, Methamphetamine, Phencyclidine, and Heroin. The Non-illicit Drug prevalence total was for any drug positive for the following substances only if the nurse did not indicate that a drug in a drug category was administered prior to collection of the urine sample: Amphetamine, all Barbiturates, all Benzodiazepines, Methadone, Codeine, Hydromorphone, and non-illicit Morphine use. In other words, if the nurse indicated that a barbiturate was administrated prior to urine collection, then whatever urine positive PharmChem Laboratories detected (e.g., amobarbital) was neither counted in the Barbiturates total, nor in the Non-Illicit Drug Total. The prevalence total for Non-Illicit Drugs is the total of all non-illicit drugs not administered prior to urine collection.

Finally, totals for all drug categories as well as the total drug and alcohol prevalence will always be lower than the sum of the individual drugs for two reasons: first, multiple drug use, though not common, was still generally one-half of 1 percent; and second, PharmChem Laboratories reported all positives as they tested them. In reporting totals by drug category, urine samples from patients who had drugs administered prior to urine collection were not considered positive. PharmChem had no way of knowing which positive urine samples were due to prescribed drug use. This was determined only after matching the data forms with the urine samples.

Table 4 presents statewide prevalence rates overall and by race/ethnicity. From the table, the statewide prevalence rate of perinatal substance use among California women in 1992 was 11.35 percent. Illicit substance exposure was 3.49 percent, with subestimates of 1.11 percent for cocaine and 1.88 percent for marijuana. The sub-estimate for alcohol is 6.72 percent and the tobacco use 8.82 percent. Only about 0.48 percent had used more than one non-illicit or illicit drug. These results suggest that alcohol and tobacco use during pregnancy is quite common in California. About 1 in 20 pregnant women had used one or more non-illicit or illicit drug, not including alcohol and tobacco, 1 in 14 used alcohol, and 1 in 11 used tobacco in the hours or days before hospitalization for delivery.

From Table 4, the highest rate of alcohol, 11.5 percent, illicit drug use 11.9 percent, and tobacco use, 20.12 percent were found among African American women; contrasts between African Americans and other ethnic subgroups are statistically significant in every instance ($p < .05$). Cocaine prevalence was high at 7.79 percent, as was marijuana (THC metabolite) at 4.59 percent. One of 4 African American women tested positive for a licit or an illicit drug use at time of hospitalization for delivery.

Table 4.--Statewide Prevalence Rates Overall and by Race/Ethnicity

Sample Size n=	Overall 29494		Asian 1645		African 2280		Hispanic 13194		White 10615		Other 1142	
Substance	PP	ET	PP	ET	PP	ET	PP	ET	PP	ET	PP	ET
1. Alcohol	6.72(.30)		5.07(1.08)		11.58(1.34)		6.87(.44)		6.05(.23)		4.03(1.16)	
2. Amphetamines	.66(.10)		.06(.12)		.19(.18)		.35(.10)		1.32(.11)		.24(.30)	
3. Barbiturates	.26(.06)		.19(.22)		.23(.20)		.22(.08)		.32(.05)		.22(.28)	
4. Benzodiazepines	.09(.04)		.00(.00)		.30(.22)		.03(.04)		.15(.04)		.01(.04)	
5. THC Metabolite	1.88(.16)		.21(.22)		4.59(.88)		.61(.14)		3.25(.17)		1.21(.64)	
6. Cocaine	1.11(.12)		.06(.12)		7.79(1.12)		.55(.12)		.60(.07)		.20(.26)	
7. Methadone	.15(.04)		.00(.00)		.25(.22)		.16(.06)		.16(.04)		.06(.14)	
8. Opiates	1.47(.14)		1.34(.56)		2.54(.66)		1.06(.18)		1.59(.12)		1.11(.62)	
9. Phencyclidine	.04(.02)		.00(.00)		.16(.16)		.06(.04)		.01(.01)		.00(.00)	
10. Total Drug Positives (Any illicit or non-illicit drug from 2-9)	5.16(.26)		1.82(.66)		14.22(1.46)		2.75(.28)		6.79(.24)		2.88(1.00)	
11. Polydrug Use (Positive for more than one drug)	.48(.08)		.04(.10)		1.75(.54)		.27(.08)		.57(.07)		.17(.24)	
12. Polydrug Use (Positive for both Alcohol and any drug)	.52(.08)		.05(.10)		1.77(.56)		.25(.08)		.56(.07)		.15(.24)	
13. Total Any Positive (Alcohol or drugs)	11.35(.36)		6.84(1.24)		24.02(1.78)		9.37(.50)		12.28(.32)		6.76(1.48)	
14. Illicit Drugs	3.49(.22)		.39(.30)		11.90(1.36)		1.51(.22)		4.92(.21)		1.57(.74)	
15. Non-illicit Drugs	1.71(.16)		1.49(.60)		2.38(.64)		1.26(.20)		1.96(.13)		1.31(.68)	
16. Tobacco	8.82(.34)		1.73(.66)		20.12(1.76)		3.29(.32)		14.82(.35)		4.81(1.30)	

White non-Hispanic women had the second highest rates of licit and illicit drugs with the exception of alcohol. White non-Hispanic women had much higher rates of smoking, 14.92 percent, than all other ethnic sub-groups except for African Americans. One in eight White non-Hispanic women tested positive for a licit or an illicit drug.

Hispanic women had the second highest rates of alcohol use, 6.87 percent. However, they had lower prevalence rates for all other drugs. About one in ten Hispanic women tested positive for a licit or an illicit drug.

Asian and Pacific Islander women had negligible prevalence for all drugs except for alcohol, 5.07 percent. One in fourteen Asian and Pacific Islander women tested positive for a licit or an illicit drug.

Table 5 presents regional prevalence estimates of total drug and/or alcohol use in descending order. The table is designed to facilitate direct comparisons of regions statewide, and to illustrate how some regions have lower prevalence for some substances, and have the highest prevalence for other substances. No one or two regions consistently have the highest or the lowest prevalence for all the substances. In

Table 5.--Regional Prevalence Rates in Descending Order by Overall and Drug Use

Overall		Alcohol		Illcit Drugs		Non-illicit		Tobacco	
Sampling Region	%	Sampling Region	%	Sampling Region	%	Sampling Region	%	Sampling Region	%
Golden Empire (2.00)	17.54	Golden Empire (2.00)	10.67	North Bay (3.00)	8.03	Alameda Co. (5.01)	3.72	Northern CA (1.00)	21.04
Alameda Co. (5.01)	16.92	Alameda Co. (5.01)	10.05	Contra Costa Co. (5.00)	6.82	San Bern. Co. (12.36)	2.78	Sacramento Co.(2.34)	15.23
Contra Costa Co. (5.00)	16.44	Contra Costa Co. (5.00)	9.03	Sacramento Co.(2.34)	6.7	N. San Joaquin (6.00)	2.7	North Bay (3.00)	15.03
Sacramento Co.(2.34)	15.21	Sacramento Co.(2.34)	7.7	Alameda Co. (5.01)	6.13	Northern CA (1.00)	2.5	Contra Costa Co. (5.00)	14.82
San Bern. Co. (12.36)	14.62	Mid Coast (8.00)	7.7	Northern CA (1.00)	6.12	San Frans. Co. (4.38)	2.35	Golden Empire (2.00)	14.79
Northern CA (1.00)	14.04	Imperial Co. (14.00)	7.42	Golden Empire (2.00)	5.88	Sacramento Co.(2.34)	2.1	San Bern. Co. (12.36)	12.81
North Bay (3.00)	14.04	Fresno Co. (9.10)	7.29	San Bern. Co. (12.36)	5.36	Fresno Co. (9.10)	1.89	Central CA (9.00)	12.18
N. San Joaquin (6.00)	12.65	LA Co. (11.00)	6.94	N. San Joaquin (6.00)	4.06	Golden Empire (2.00)	1.77	N. San Joaquin (6.00)	11.97
Fresno Co. (9.10)	12.61	Marin/San Mat (4.00)	6.9	Central CA (9.00)	3.96	LA Co. (11.00)	1.75	Alameda Co. (5.01)	11.42
Mid Coast (8.00)	11.86	San Bern. Co. (12.36)	6.85	Fresno Co. (9.10)	3.72	Santa Clara Co. (7.00)	1.73	Riverside Co. (12.33)	10.63
Riverside Co. (12.33)	11.36	San Diego Co. (14.37)	6.68	Riverside Co. (12.33)	3.53	Contra Costa Co. (5.00)	1.67	Mid Coast (8.00)	9.67
San Frans. Co. (4.38)	11.12	Riverside Co. (12.33)	6.58	Sn Frans. Co. (4.38)	3.47	Riverside Co. (12.33)	1.46	San Diego Co. (14.37)	9.6
LA Co. (11.00)	10.77	Ven/S.B. (10.00)	6.49	Mid Coast (8.00)	3.37	Ven/S.B. (10.00)	1.44	Fresno Co. (9.10)	9.52
Ven/S.B. (10.00)	10.04	North Bay (3.00)	6.12	Santa Clara Co. (7.00)	3.21	Orange Co. (13.00)	1.36	Ven/S.B. (10.00)	7.81
Santa Clara Co. (7.00)	9.76	N. San Joaquin (6.00)	6.1	LA Co. (11.00)	2.6	Mid Coast (8.00)	1.01	Santa Clara Co. (7.00)	7.3
Central CA (9.00)	9.51	San Frans. Co. (4.38)	5.99	Marin/San Mat (4.00)	2.42	Imperial Co. (14.00)	0.98	San Frans. Co. (4.38)	6.6
Imperial Co. (14.00)	9.44	Northern CA (1.00)	5.96	San Diego Co. (14.37)	2.35	Central CA (9.00)	0.95	Orange Co. (13.00)	5.88
San Diego Co. (14.37)	9.44	Santa Clara Co. (7.00)	5.5	Ven/S.B. (10.00)	2.21	North Bay (3.00)	0.79	LA Co. (11.00)	5.84
Marin/San Mat (4.00)	9.36	Central CA (9.00)	4.78	Orange Co. (13.00)	2.06	San Diego Co. (14.37)	0.61	Marin/San Mat (4.00)	5.73
Orange Co. (13.00)	7.49	Orange Co. (13.00)	4.49	Imperial Co. (14.00)	1.04	Marin/San Mat (4.00)	0.53	Imperial Co. (14.00)	4.7
Inyo/Mono (12.00)		Did not participate in the study							

Marijuana		Amphetamines		Opiates		Cocaine	
Sampling Region	%	Sampling Region	%	Sampling Region	%	Sampling Region	%
North Bay (3.00)	6.36	San Bern. Co. (12.36)	2.8	Alameda Co. (5.01)	3.51	Alameda Co. (5.01)	3.21
Golden Empire (2.00)	4.48	Golden Empire (2.00)	1.73	San Bern. Co. (12.36)	2.6	San Frans. Co. (4.38)	2.15
Contra Costa Co. (5.00)	4.41	Riverside Co. (12.33)	1.45	Northern CA (1.00)	2.59	Contra Costa Co. (5.00)	2.04
Northern CA (1.00)	4.4	Sacramento Co.(2.34)	1.45	N. San Joaquin (6.00)	2.45	Fresno Co. (9.10)	1.96
Sacramento Co.(2.34)	3.89	Northern CA (1.00)	1.44	San Frans. Co. (4.38)	2.41	Sacramento Co.(2.34)	1.67
N. San Joaquin (6.00)	2.58	N. San Joaquin (6.00)	1.42	Sacramento Co.(2.34)	2.03	LA Co. (11.00)	1.4
Central CA (9.00)	2.55	Contra Costa Co. (5.00)	1.4	Golden Empire (2.00)	1.77	San Bern. Co. (12.36)	0.95
San Bern. Co. (12.36)	2.55	Central CA (9.00)	1.09	Contra Costa Co. (5.00)	1.64	Santa Clara Co. (7.00)	0.9
Mid Coast (8.00)	2.43	North Bay (3.00)	0.85	Fresno Co. (9.10)	1.55	Mid Coast (8.00)	0.87
Alameda Co. (5.01)	2.35	Imperial Co. (14.00)	0.52	Santa Clara Co. (7.00)	1.5	North Bay (3.00)	0.78
Santa Clara Co. (7.00)	2	Alameda Co. (5.01)	0.47	LA Co. (11.00)	1.3	Central CA (9.00)	0.76
Riverside Co. (12.33)	1.9	San Diego Co. (14.37)	0.45	Riverside Co. (12.33)	1.29	Orange Co. (13.00)	0.7
Ven/S.B. (10.00)	1.48	Ven/S.B. (10.00)	0.45	Orange Co. (13.00)	1.11	Marin/San Mat (4.00)	0.59
Fresno Co. (9.10)	1.46	Fresno Co. (9.10)	0.43	Mid Coast (8.00)	1.08	Riverside Co. (12.33)	0.56
San Diego Co. (14.37)	1.42	Marin/San Mat (4.00)	0.34	North Bay (3.00)	1.04	San Diego Co. (14.37)	0.53
Marin/San Mat (4.00)	1.22	Orange Co. (13.00)	0.32	Ven/S.B. (10.00)	0.99	Northern CA (1.00)	0.41
Orange Co. (13.00)	1.15	Santa Clara Co. (7.00)	0.31	Central CA (9.00)	0.72	Ven/S.B. (10.00)	0.38
San Frans. Co. (4.38)	1.15	Mid Coast (8.00)	0.25	Marin/San Mat (4.00)	0.7	N. San Joaquin (6.00)	0.33
LA Co. (11.00)	1.03	LA Co. (11.00)	0.22	San Diego Co. (14.37)	0.52	Golden Empire (2.00)	0.1
Imperial Co. (14.00)	0.52	San Frans. Co. (4.38)	0.18	Imperial Co. (14.00)	0.49	Imperial Co. (14.00)	0
Inyo/Mono (12.00)		Did not participate in the study					

sum, there is one outstanding feature of the comparisons: prevalence estimates are not related to rural or urban strata per se, but both urban and rural regions in the northern part of California are more likely to have higher total prevalence rates than those in the southern part of California.

For example, the four regions with the highest total prevalence rates (for either drugs and/or alcohol) are in the northern part of California, and nine of the first ten are in the central and northern part of California. The only exception is San Bernardino county. While some predominantly rural regions have high total drug and/or alcohol use, other have relatively low use, such as Imperical county. These results reveal one other interesting finding. There seem to be large variations in prevalence rates among counties within the same geographic area. Such is the case in the region surrounding San Francisco Bay. The relatively low prevalence levels of Marin-San Mateo (Stratum #4.00), 9.36 percent, contrast with those of Contra Costa (Stratum #5.00) and Alameda (Stratum #5.01) counties, 16.44 percent and 16.92 percent, respectively. The rates for the North Bay (Stratum #3.00), Santa Clara (Stratum #7.00) and San Francisco (Stratum #4.38) counties range at all points in between. In the southern part of California, similar results are found when comparing San Bernardino (Stratum #12.26) and Orange (Stratum #13.00) counties. These variations underscore the complexity of interpreting these findings, and the importance of conducting a more detailed analysis to understand the reasons for this distribution.

While it is evident that regions in the northern part of California generally have higher prevalence rates than regions in the southern part of California, this does not seem attributable to income differences among counties/regions. However, there probably is a direct link to sociodemographic composition of the regions. Regions in the southern part of California with proportionally large Hispanic populations had lower prevalence on illicit drugs and tobacco use because Hispanic women had a low statewide prevalence for these substances generally. However, it remains enigmatic why White non-Hispanic maternity patients or Asian and Pacific Islander maternity patients, for example, in the northern part of California seemed so much more likely to test positive for various substances during the final term of their pregnancy than women in the same race/ethnic group in the southern part of California.

It appears that there is a strong cultural basis for perinatal substance use in California that operates to minimize use, or influence the decision to use particular types of substances. There also appeared to be "hot spots" of concentrated substance use, such as cocaine in Alameda county, amphetamines in San Bernardino county, and marijuana in the North Bay HSA (Napa, Solano, and Sonoma counties). These important differences in regional characteristics underlying the estimates suggest that they have critical public health implications for targeting services.

References

Vega, William A.; Kolody, Bohdan; and Hwang, Jimmy (1993a). Prevalence and Magnitude of Perinatal Substance Exposures in California, *The New England Journal of Medicine*, 329:850-854.

Vega, William A.; Kolody, Bohdan; and Hwang, Jimmy (1993b). *Profile of Alcohol and Drug Use During Pregnancy in California, 1992*, California Health and Welfare Agency.

Cochran, William G. (1977). *Sampling Techniques*, John Wiley & Sons, 276. ■

The Processing and Editing System of the National Health Interview Survey: The Old and New

Susan S. Jack, National Center for Health Statistics

9
Chapter

Abstract

The National Health Interview Survey (NHIS) has been fielded continuously since 1958. The procedures for processing, editing, producing, and documenting clean data with precise documentation have evolved over time. The data "products" have also changed somewhat over time, consistent with evolving technology.

The forthcoming redesigned NHIS, which has been converted to Computer-assisted Personal Interviewing format using the CASES authoring system, is a natural outcome of this evolving technology. The concurrent challenge is to redesign a processing and editing system that retains the "spirit" and the positive aspects of the old system while benefiting from technologic advances, minimizing the amount of labor intensive editing, and producing new forms of documentation appropriate to the NHIS's widely varying audience.

This presentation will give an overview of the current data processing/editing procedures, which reflect a paper-and-pencil NHIS, potential human error in responding, recording, and keying of data using a mainframe system, according to detailed specifications prepared by subject matter and editing experts. It will also describe the process NHIS is using to design a new multi-dimensional system from raw data to clean, documented data release. Using personnel with

Abstract (Cont'd)

a wide variety of experience and expertise, interrelated work groups will set policies and priorities which should ultimately

- ❒ reduce the amount of time involved in edit specification and programming,
- ❒ allow for minor modifications as well as major changes involving cyclical modules,
- ❒ automate and simplify the documentation process, and
- ❒ decrease the turn-around time for clean data release.

The goal is to synthesize the best of the old NHIS policies with the experience of others undergoing the CAPIization process to produce an essentially "generic" editing/ processing system which can be used by other large complex surveys, particularly those within NCHS. ■

Chapter 10

Neural Networks

Chair: Ray Board, Federal Reserve Board

L. H. Roddick

L. H. Roddick

Svein Nordbotten

Chapter 10

Data Editing Using Neural Networks

L. H. Roddick, Statistics Canada

Abstract

Computerized data editing is normally done using rule based systems or programs. Neural Networks provide the ability to create an edit system directly from the data and to allow the edits to evolve over time through periodic retraining. This is an especially useful facility for monthly surveys, where the edits can adapt over time as the external conditions affecting the survey change.

Data Editing Using Neural Networks

L. H. Roddick, Statistics Canada

Introduction

Data editing is a major activity at Statistics Canada and at other statistical agencies. Computerized data editing is normally accomplished either by producing a customized edit program or by using a generalized editing system. A customized approach involves specifying the edits, programming the specifications, testing the program's correctness and then implementing the edit program. A generalized editing system usually requires the user to specify the edits in terms of some complete set of acceptance or rejection rules (Fellegi and Holt, 1976). In both techniques the edits are implemented in terms of rules.

The Neural Network technique described herein addresses editing data based on the actual data, not on rules about the data. Only one facet of statistical data editing is described, but it is thought that non-statistical areas might also make use of this type of Neural Network.

Two types of data are normally edited, discrete valued variables and continuous valued variables. Discrete variables have a limited number of values that are either alphabetically or numerically coded. Continuous variables are numeric variables such as integers.

The Neural Network model presented here addresses discrete variables, however it can be used for continuous variables if they are grouped by value into classes. For example, an income variable can be grouped into a number of income classes such as $0 -- $20,000, $20,001 -- $40,000 and greater than $40,000, thus making the resulting coded income variable discrete.

The Process

Editing of discrete variables can be accomplished using Back-Propagation (B-P) Neural Networks without having to go through the specification, programming and testing phases. This is accomplished by providing the set of correct and incorrect values for the variables being edited to a data generation program that will generate the training set for the Neural Network. The Neural Network is trained on this generated data and the resulting Neural Network module is incorporated into a capture or edit system.

This technique only addresses the inter-field edits within a record. It is not applicable to inter-record edits where a model different from the one proposed would be required.

Although a record may have many fields (numbering even into the hundreds), normally only a few fields (fewer than 15) interact at a time. Thus the inter-field edits can be grouped according to sets of fields to be edited together. The edits for each of these groups is distinct and separable and thus a

separate Neural Network can be trained for each group. This does not mean the proliferation of many programs to edit a record. Once each edit Neural Network is trained, it can become a module of the overall edit system or can be introduced as a module of an interactive capture system to edit the records as they are being keyed.

When this technique was originally tested, it was envisaged that data generation of the complete set of possible values would be the most convenient mechanism for the user to train the Neural Network. It is with this in mind that the example for the paper was generated, however, after discussion of the mechanism it became apparent that there are better ways to construct the training set. Alternative methods of constituting the training set are discussed after the Generated Data example.

Important to data generation is the fact that the set of correct answers for discrete variables, in comparison to the set of all answers, is normally quite small. For example, if three variables interact and each variable has twenty values, there are 8,000 possible answers, although normally only a small number of these is correct. This makes it expedient to specify only correct answers to the training data generation program and allow it to generate the incorrect set. It is however perfectly feasible to specify the incorrect set and generate the correct set.

There are two important points to make about B-P Neural Networks. First, a B-P Neural Network is much better at distinguishing between the values of separate variables than the individual values within a variable (Lawrence, 1991); i.e., the network will learn far more from nine binary variables than it will from one coded variable using a scale of 1 to 9. This is the principle that the data preparation program uses when it generates the network training data. This technique is also used to transform the data for presentation to the network for editing.

Second, a B-P Neural Network must be presented with approximately equal numbers of each output case during the learning phase (Klimasauskas, 1993). This is to ensure that the network learns as well about the rare cases as it does about the more frequent cases. Since the correct cases are normally less numerous that the incorrect cases, the correct cases must be replicated enough times to create equal representation in the training set. Once the network is trained, it's weights are frozen and no longer change, therefore there is no need for this equal representation in the data being edited.

An Example

The B-P Neural Network data editor is presented by way of an example. Assume that there are three variables that interact, Var1, Var2 and Var3, and that these variables have 5, 9 and 9 possible values respectively. The three variables in combination can produce 405 different answers. An arbitrary set of correct answers for the three variables was chosen and are specified in Table 1. These combinations of values produce 17 correct answers and 388 incorrect answers.

Table 1.--Correct Variable/Value Combinations

Var1	Var2	Var3
1	3 or 7	6 or 9
2	1 or 2	2 or 7
3	4 or 5	1 or 3
4	6 or 8	4 or 5
5	9	8

In order to balance the training set the data generation program created 20 copies of the correct answers and 1 copy of each incorrect answer. This produced a training set of 728 records consisting of 340 correct and 388 incorrect records.

The data generation program transformed the codes for each variable into a set of binary variables with one binary variable for each possible data variable answer. This transformation produced 5 binary variables for Var1, 9 binary variables for Var2 and 9 binary variables for Var3. For each value that occurred in Var1, Var2 or Var3, the corresponding binary variable for the value was set to 1, all the other binary variables were set to 0.

For each record, the data generation program created two binary result fields, Correct and Incorrect. When a correct record was to be generated the Correct field was set to 1 and the Incorrect field was set to 0. When an incorrect record was to be generated Correct was set to 0 and Incorrect was set to 1.

The data generation program generated 728 training records, each with 25 binary variables. The list of binary variables, the data variable from which they were generated and the data variable value that produced a 1 in the binary variable are specified in Table 2.

Table 2.--Network Training Record Layout and Inputs

Binary Variable	Generated From
Var1b1	Var1 = 1
Var1b2	Var1 = 2
Var1b3	Var1 = 1
Var1b4	Var1 = 1
Var1b5	Var1 = 2
Var2b1	Var2 = 1
Var2b2	Var2 = 2
Var2b3	Var2 = 3
Var2b4	Var2 = 4
Var2b5	Var2 = 5
Var2b6	Var2 = 6
Var2b7	Var2 = 7
Var2b8	Var2 = 8
Var2b9	Var2 = 9
Var3b1	Var3 = 1
Var3b2	Var3 = 2
Var3b3	Var3 = 3
Var3b4	Var3 = 4
Var3b5	Var3 = 5
Var3b6	Var3 = 6
Var3b7	Var3 = 7
Var3b8	Var3 = 8
Var3b9	Var3 = 9
Correct	Assigned
Incorrect	Assigned

The data generation program also produced a test data set containing 405 records, one for each possible case. These records were generated in the same way as the training data, including the answers (correct or incorrect), except that there was no need to expand the number of correct cases. During the testing phase, the Neural Network did not use the answers, however it wrote these answers to the output file along with its predictions. These answers were compared to the predictions to do the analysis of the Neural Network results.

Several Back-Propagation Neural Networks were constructed and tested using the NeuralWare Inc. NWorks commercial Neural Network to find a reasonable working version. A Neural Network as specified in Table 3 was able to predict the test set answers correctly all of the time. Since the test set had all possible cases, this Neural Network could be used as an editor for the three variables as they were described.

Table 3.--Back-Propagation Neural Network Parameters

Control Strategy:	Back-Propagation
Type:	Hetero-Associative
Input PEs:	23
Hidden Layer PEs:	23 in 1 hidden layer
Output PEs:	2
Input Ranges:	0 -- 1, using a min./max. table
Output Ranges:	0 -- 1, using a min./max. table
Transfer Function:	Input & Bias layers -- Linear Hidden & Output layers -- TanH
Learning Rule:	Input & Bias layers -- None Hidden & Output layers -- Norm-Cum_Delta
1st Learning Coefficient:	Hidden layer -- 0.3 Output Layer -- 0.15
2nd Learning Coefficient:	Hidden layer -- 0.4 Output layer -- 0.4
Learn Data Presentation:	Shuffle & Deal Randomization without replacement 50,000 presentations

Discussion

A number of points should be made about this data editing technique. First, the specification of the correct data required to generate the training and test sets need not be long and cumbersome. The data generation program input could be in a tabular format with the user specifying only the axis points of interest and allowing the program to generate the correct combinations. This would simplify and minimize the data specification process. It should be noted that if the edit were done conventionally, the user would have to specify all of the correct cases anyway.

Second, the outcome of the Neural Network, i.e., the set of edits is not limited to a correct/incorrect scenario. The outputs could be numerous, reflecting a much more complex relationship in the data being edited. As an example, the desired result could be the stratification of the data based on its inputs, in which case the outputs could be as many as there are strata.

Third, the implementation of these Neural Networks has been made much simpler by the appearance of commercial Neural Network packages on the market. While more complex in nature than a spreadsheet, Neural Networks can be generated and modified with this same spreadsheet type of ease using the new commercial packages.

Although a Neural Network can be generated quickly using commercial packages, this should not be interpreted to mean that Neural Networks are easy to use. The Neural Network concepts and packages have a long learning curve to master and every implementation requires its own analysis and tuning. However, once the expertise is there, the time to produce a Neural Network is very fast. The Neural Network presented in this paper underwent three iterations where it was specified, run and tested each time, this took two person-days to complete.

Finally, the advantage of a data generation technique from the Neural Network point of view is that it is not vulnerable to memorization, a major problem with Neural Networks. When memorization occurs instead of learning, it causes poor generalization of the Neural Network. However since the data generation program has generated all possible cases, it does not matter whether the Neural Network has memorized them or learned them as long as it always gets its prediction right.

Other Applications

An effective use of the Neural Network's ability to learn would be its application to monthly surveys. The conditions under which a monthly survey occurs change over time, for example the edit limits for an economic survey might be much looser during good economic times and much more severe during a recession. It would be desirable for the edits to modify themselves over time as the conditions changed.

The Neural Network could be trained initially on historical data and run on the first month's data. After each month has been run the training set could be augmented with selected data from the previous month and the network retrained. This new set of data could include data that the network found in error, but that follow-up procedures found to be correct due to changing conditions. Depending on the extent of the survey follow-up, data could also be added for records that the network predicted as correct, but that conditions now invalidate.

Another technique to assist in the building of monthly survey training sets is the thermometric encoding of the outputs (Guiver and Klimasauskas, 1991). The output of the training and the test set are predictions of the correctness of the records. The training set answers could be provided as a set of variables which represent the "hotness" (correctness) of the records. The more correct the record, the hotter the answer, thus the network would learn to predict correctness on a gradient scale. This could be very useful in surveys with critical and non-critical populations, where the critical population must all be edited to the "hottest" (most correct) level, but the non-critical population could be edited to a lower temperature. This is one of a number of output encoding techniques used to facilitate different types of learning in Neural Networks.

Acknowledgment

The author would like to acknowledge the Research Fellowship Program at Statistics Canada, without which none of this work would have been accomplished.

References

Fellegi, I. P. and Holt, D. (1976). A Systematic Approach to Automatic Edit and Imputation, *Journal of the American Statistical Association*, 71, 353, Applications Section, 17-35.

Guiver, J. P. and Klimasauskas, C. C. (1991). Applying Neural Networks Part IV: Improving Performance, PC AI, July/August, 34-41.

Klimasauskas, C. C. (1993). Neural Network Development, A Methodology, Course Notes, NeuralWare Inc., Revision 4.2, 173-197.

Lawrence, J. (1991). Data Preparation for a Neural Network, *AI Expert*, November, 34-41. ■

Editing Monthly Survey Data Using Neural Networks

L. H. Roddick, Statistics Canada

10 Chapter

Abstract

Systems which are used to edit survey data should be able to adjust automatically for the changing conditions under which the survey is taken. For example, a set of rules developed to edit respondent data from an economic survey may need to be adjusted as the underlying economy moves from a period of recession to a period of economic expansion. Since the progression from recession to expansion is generally a gradual process, the system should be capable of evolving a set of edits, as changes are observed in the characteristics of the input respondent data set. This requirement is especially relevant in the case of monthly surveys, where time constraints do not normally permit the re-specification, re-programming, and re-testing of the edits. An edit program, based on Neural Network technology, can be developed which is able to evolve its set of edits automatically, as the characteristics of the input survey data set change. Unlike a conventional edit system, it is able to produce automatically a new instance of itself, tailored to the changed survey environment, each month. The paper presents an implementation of an edit system using an evolving artificial neural network.

Editing Monthly Survey Data Using Neural Networks

L. H. Roddick, Statistics Canada

Introduction

An artificial neural network (ANN) can be created to perform an edit of several related variables (Nordbotten, 1993); (Roddick, 1994); and a number of ANN's can be connected together to create an edit system. This editing system can be implemented as a stand alone edit system or as modules that are invoked as part of a data capture system.

An editing ANN must be able to recognize three types of records: correct (CR), incorrect (IC) and don't know (DK) records. An edit will normally involve the interaction of several variables, and when these are discrete variables, the cross product of these variables can be listed empirically. For example, an edit of the variables Mother Tongue versus Ethnicity will produce a large set of possible combinations; however, relatively few of the combinations are CR, a few are known to be IC, and the rest are initially DK. An edit system can be built that will allow the subject matter expert to analyze DK cases as they occur and assign them to either the CR or the IC set. If an occasional unique DK record appears, then there is a likelihood that it is IC, however if a number of records of the same DK appear in a survey occasion, then maybe a new, valid Mother Tongue/Ethnicity combination has been identified. This approach uses data analysis to produce the edit program rather than the traditional specification, programming and testing approach.

Continuous variables can be edited with this form of ANN by creating classes of values for each continuous variable to be edited. For example, an income variable can be grouped into 3 classes: 0-20,000; 20,001-50,000; and > 50,001, and thus the continuous variable can be treated as a 3-value discrete variable.

There are three issues involved in creating an ANN which can edit survey responses and in which the edits can evolve. The first issue is the creation and ongoing maintenance of the ANN training set. The second issue is the training of the ANN, both initially and the monthly retraining as conditions change. The third issue is the assignment of the ANN DK cases to either the CR or IC sets and the periodic re-analysis of the whole training set.

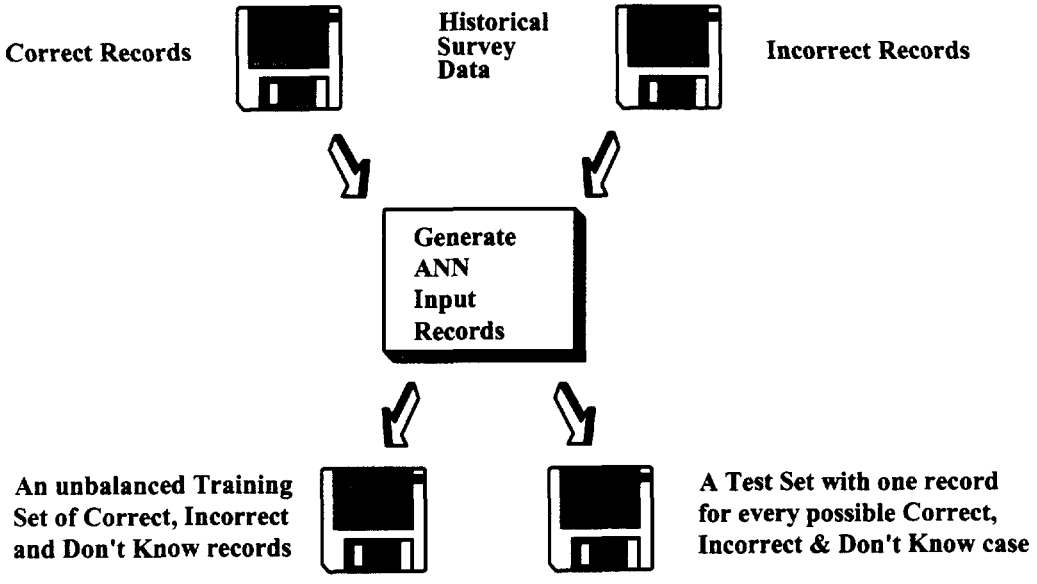
ANN Training Set Creation and Maintenance

The ANN training set can be created initially by either data analysis or by specifying all of the CR and IC cases. For an ongoing survey, the CR cases can be generated from historical edited data, as all the responses on the edited data set should be correct. The IC cases can be generated by comparing the unedited historical data with the edited data and assigning all the cases that exist in the unedited data, but not in the edited data, to the IC set. The DK cases are all of the remaining possibilities and can be

generated automatically. For a new survey, the expert can specify all of the known CR and IC cases, allow the system to assign the remainder to DK and let the first survey occasion generate DK records, which the analyst will then assign to either CR or IC. This may produce a large workload of DK cases on the first survey occasion, but the volume of DK cases should diminish quickly thereafter.

The ANN training sets are maintained on a relational database, with a table for each edit, containing all of the training records for that edit. The edit tables contain a column for each variable used in the edit and one column containing the answer for each case: CR, IC, or DK. These edit tables are uniquely keyed by all of the edited variables to ensure that there are no duplicate cases. To ensure that no invalid values are introduced, each variable in the edit table has a codeset of the acceptable values for the variable. This database approach provides access to all of the most current tools for the analysis and maintenance of the edit sets and ensures the completeness and the correctness of the training data.

Figure 1.--Initial ANN Training Set Creation



The question that arises from this approach is: If all of the edit answers are contained in a database table, why not just look up the answers directly instead of using an ANN? The answer is a performance issue. If the edit system is only applying a few edits to a limited number of records and has only a few possible cases, then a lookup table is very feasible. Often however there are many edits and hundreds of possible cases for each edit. An ANN reduces the edit to the execution of a series of mathematical equations with no I/O involved, and is therefore orders of magnitude faster than a lookup approach.

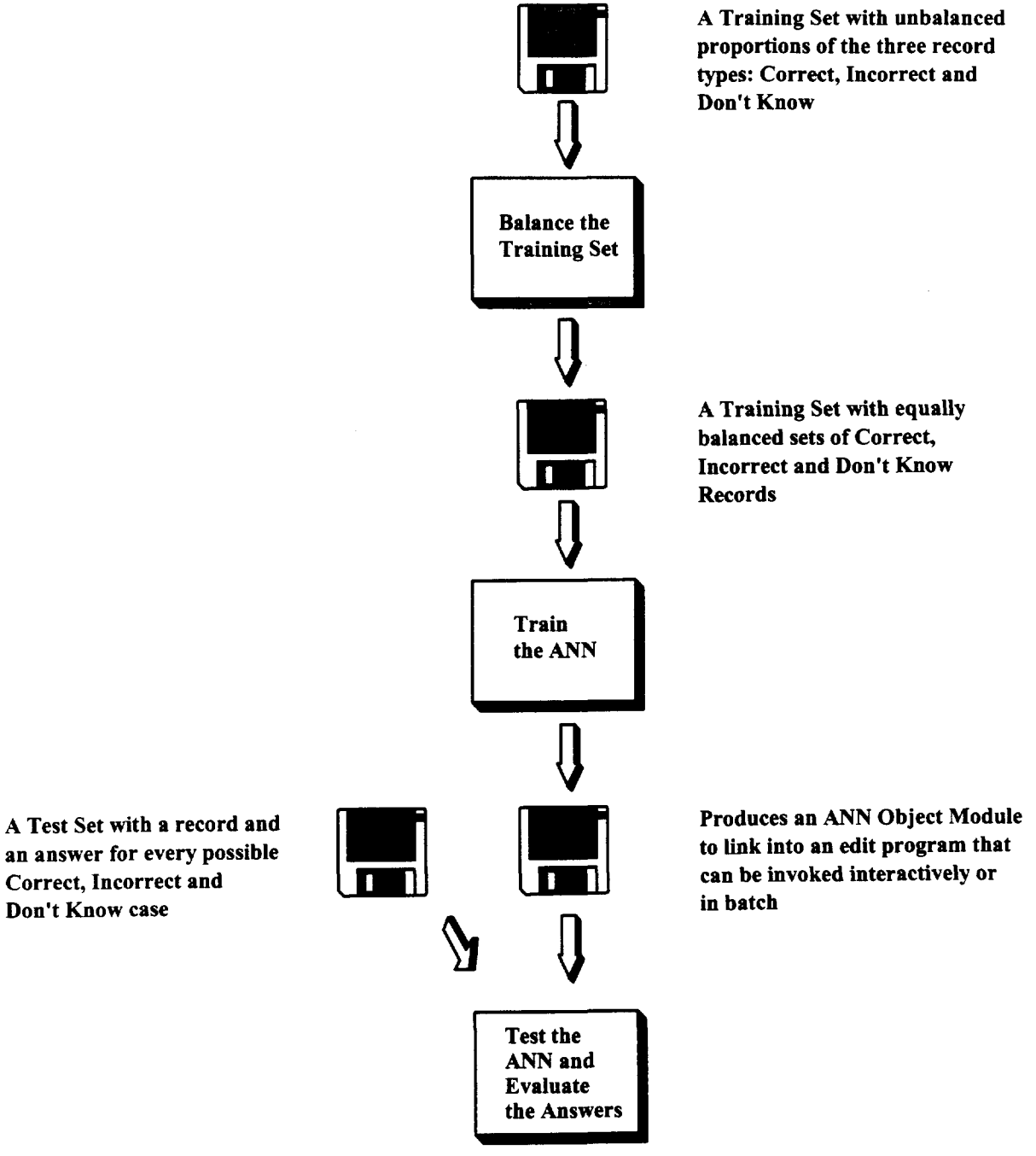
Training and Deploying the ANN

The ANN must be trained initially and then re-trained every time the training set is modified. The training set is generated from the database edit table by transforming the records into a form that the ANN can use. Two functions must be performed on the records; the distance must be established between the values of each variable (Roddick, 1994) and the training set must be balanced to have equal representation of all record types in the training set (Roddick, 1994).

Distance is achieved by mapping each of the variable values to a binary variable. The ANN does not distinguish well between a value of 5 versus a value of 6 in a single variable. However if that single variable is converted to a set of binary variables, with one for each possible value, the ANN distinguishes very well between the binary variables Var5 = 1, Var6 = 0, versus Var5 = 0, Var6 = 1.

Balancing the training set forces the ANN to learn about all of the cases. It involves replicating the rare cases enough times to ensure that the network sees them as often during training as the frequent cases. If 95 percent of the training set were DK cases, the ANN would learn very quickly to always choose DK, and it would be right 95 percent of the time. By making the CR and IC cases occur as often as DK, this forces the ANN to learn what distinguishes CR, IC, and DK.

Figure 2.--ANN Training and Deployment

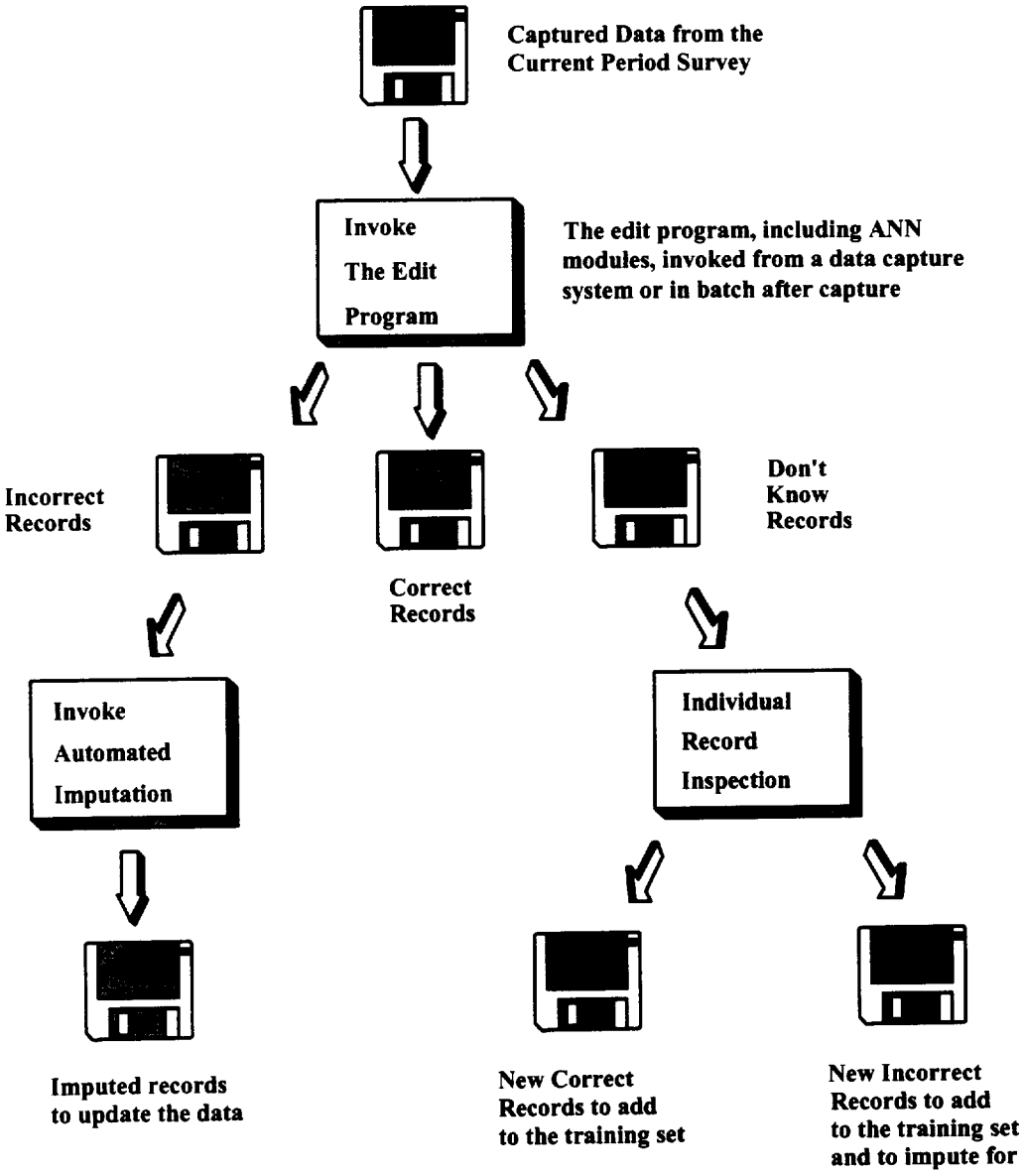


When all of the possible cases are available to train the ANN, it can be trained to be 100 percent correct, if the training set is presented properly. Having all of the possible cases for training produces a classifier ANN, whereas having a subset of the cases for training produces a predictor ANN.

The result of the ANN training is an executable subroutine that can be called from a data capture system or an edit system. The ANN is tested to ensure that it is correctly trained by passing all of the cases on the database edit table through the ANN. The ANN is successfully trained if it correctly classifies all of the cases.

The ANN is invoked by a capture system (or an edit system) via a control subroutine that the capture system calls. The capture system calls the control subroutine with the name of the edit being invoked and the data from the fields being edited. The control subroutine converts the discrete variable values to

Figure 3.--Edit ANN Invocation



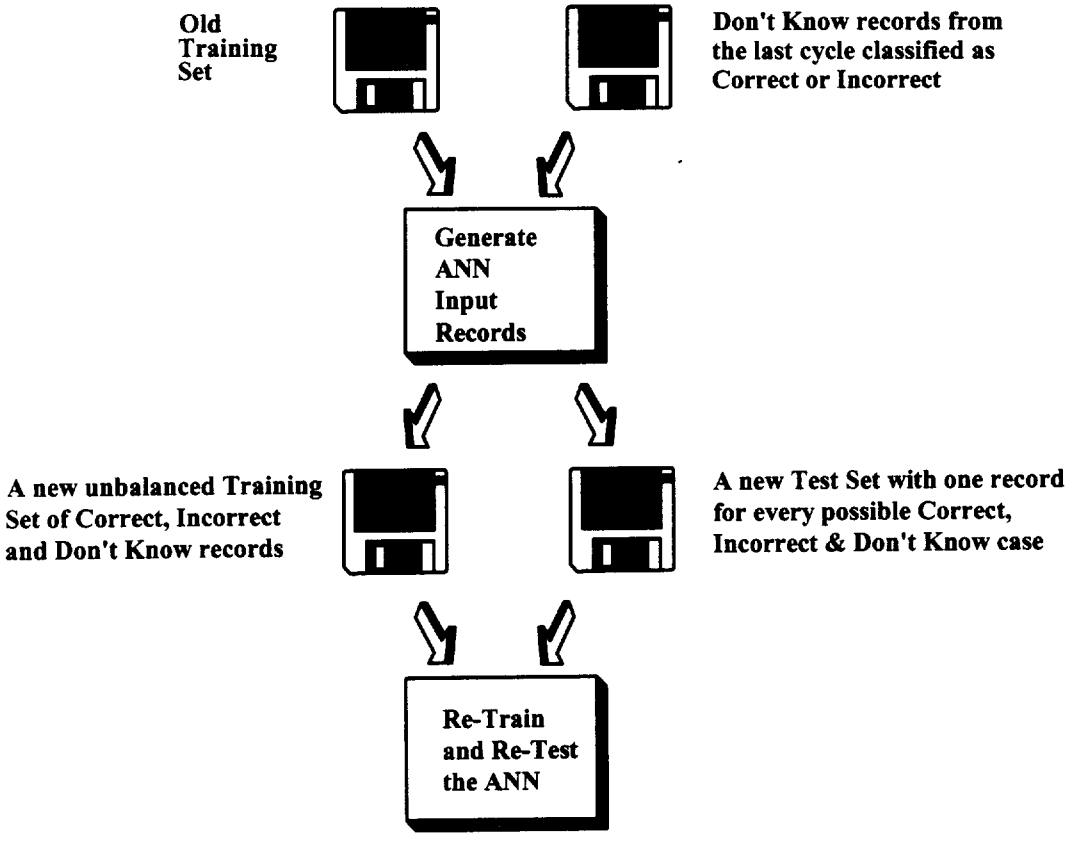
binary variables and calls the edit ANN that is being requested. The ANN edits the values and passes its answer, CR, IC, or DK, back to the control subroutine. The control subroutine converts the answer into the form that the capture system can use and passes the answer back to the capture system. This control layer of insulation is used to ensure that any system using the ANN does not have to know about the peculiarities of the ANN inputs or outputs. It also allows for the regular re-training of the ANN with no impact on the capture system. A generalized control subroutine can be written that can action any of the edit ANNs in the system and interface with the capture system to pass the answers back.

The Production Cycle

As the ANN is run each month it classifies records into one of the three types. The CR records pass through untouched, the IC records go on for correction, either by an imputation ANN or by a conventional imputation routine, and the DK records are output to a database table of DK's for that specific edit. The subject matter specialist analyzes the table and assigns each record to either CR or IC and the database edit table is updated. Once the DK records for the current month have been classified, the ANN is re-trained to produce a new program for the next month. Using a database table for each edit's DK records provides an analytical history of what decisions have been made about that edit.

Assuming that there are no radical shifts in response patterns, the set of DK records output each month should be small and manageable. If the original specification of the CR and IC records is well done, there should only be few DK records in any one month and these should represent new knowledge about the subject matter that should be individually examined.

Figure 4.-- Periodic ANN Regeneration



Implementation

The edit ANN, as described, has been implemented for a data capture system on a PC using Microsoft Access® for the GUI and the DBMS. The Access capture screen called the control subroutine library, written in C, which in turn called the ANN subroutine. The ANN subroutine was generated using a commercial ANN product, NeuralWorks Professional II/Plus® from NeuralWare Inc. The analytical tools to assign the DK cases were built in Access.

Conclusions

ANN edit systems can be constructed that provide a viable alternative to traditional approaches to data editing. These ANN edit systems shift the emphasis to data analysis rather that specification for the generation of the system and therefore the system development methodology better emulates the modus operandi of the subject matter expert/analyst.

The ability of the edit to evolve over successive survey occasions without re-programming and re-testing should result in substantial development and maintenance savings for the survey. The fact that the ANN can interface with a capture system, an edit system, or both, should increase the flexibility of where the edits are applied.

A classifier ANN system using this method of training, with all of the cases provided, should also be effective for the construction of stratification and auto-coding systems.

References

Nordbotten, S. (1993). *Editing Statistical Records by Neural Networks*, Department of Information Science, University of Bergen, Norway.

Roddick, L. H. (1994). *Data Editing Using Neural Networks*, EXPERSYS '94, 655 - 659. ■

Chapter 10

Editing and Imputation by Means of Neural Networks

Svein Nordbotten, University of Bergen

Abstract

Editing and imputation of statistical data are possible because we take advantage of some prior knowledge about the type of statistical objects we investigate. The processes of editing and imputation are considered expensive parts of survey and census costs. This presentation discusses the use of neural network methodology to improve the efficiency of these processes. Two applications are discussed as demonstrations of the approach.

Editing and Imputation by Means of Neural Networks

Svein Nordbotten, University of Bergen

Models of Statistical Objects

Knowledge of Statistical Objects

In statistical agencies, a huge amount of knowledge exists about the different classes of the objects of which a modern society is considered composed. When preparing surveys for updating this knowledge, the already existing knowledge can and is extensively used. When screening data to identify suspect records, the data are checked to see if it is corresponding to the prior knowledge structure. When suspect records are detected, the structure is again used for imputing or modifying the data.

The situation may be compared to the recipient who is reading a received, handwritten letter. When he arrives to a string of symbols he does not recognize, he uses his knowledge of vocabulary, grammar and the context to decide if the word may be a valid, but for him unknown word, or a misspelled word. In case he decides that the word is a valid, but for him an unknown word, he will probably add it to his vocabulary, while in the case of a misspelled word, he may decide to correct the spelling.

In previous times, the statistical agencies hired specialists on demographic aspects to edit the data collected about individuals and their families. After the appearance of programmed computing equipment, much effort has been spent trying to get the computer systems to simulate the specialists or even to behave as specialists ideally should.

Before automatic editing of statistical records, the most typical and stable patterns among the features of objects are identified, embedded in programs as editing criteria. The individual records could be classified in groups as "acceptable," "suspicious," "defect," etc., and adequate procedures for handling the different groups be devised.

In other word, we can think of the development of the editing process as building editing models reflecting the typical patterns of properties for the object belonging to the class investigated. These models may be further extended to include the imputation, i.e., predict the most probable properties for object that are only partly known. The problem is of course how to extract from specialists the knowledge needed for specifying operative editing and imputation models (Fellegi and Holt, 1976).

Estimating Model Structure by Neural Networks

The models known as neural networks can be considered from a number of different angels. The name, "neural network," indicates an origin in neuro-physiology. Some researchers point out that "parallel processing" is an important feature of the models (Rumelhart and McClelland, 1986). Statisticians can claim that some of important algorithms used, as f.ex. the back propagation learning algorithm, were originally developed as extensions to already existing statistical methods (Werbos, 1974; Cheng and Titterington, 1994). Finally, mathematicians emphasize that the methods used can be considered as universal approximators.

In context of editing and imputation, it may be convenient to regard neural networks as sets of non-linear, multi-variate regression models used for prediction of values of independent variables for individual objects subject to existence of values for some known dependent variables of the same objects. The neural network models can be designed for prediction of a single dependent variable as well as for simultaneous prediction of a number of variables. In contrast to the multi-variate, statistical models, the neural network models are non-parametric and non-linear. This can be a serious drawback because it does not permit any direct inference about the quality of predictions based on statistical theory. The type of neural network most likely to be useful for editing, have, on the other hand, the great advantage that the networks can "learn" from a set of records representing the prior knowledge. The term "learn" means in this context that algorithms exist for estimating the huge number of parameters or internal connection weights of the networks.

Applications

Neural networks have many potential applications in statistical production besides editing and imputation. In this presentation, we limit the discussion to two applications demonstrating:

- simultaneous editing and imputation, and
- large scale imputation as a separate process.

Application of neural networks for editing and imputation in surveys and censuses, is probably just in the initial stage. One of the very few who has done research and promoted the use of neural networks in editing is Hugh Roddick (Roddick, 1993). It has also recently been reported that the National Statistical Office in UK is now evaluating neural networks for their 2001 Population Census (Clark and Street, 1996).

CASE 1: Simultaneous Editing and Imputation

Editing

The first application we shall discuss demonstrates simultaneous editing and imputation of data collected in an imaginary survey. Further details are discussed in Nordbotten (1996a). The scenario considered is a sample survey comprising 12.000 individuals, for whom values of the 9 categorical, socio-economic attributes -- sex, age group, marital status, geographic region, number of children, education, industry, employment status, and income group -- were collected. It was assumed that each record during collection and pre-processing, was subjected to a risk for partial non-response and erroneous data fields.

The questions studied in connection with this scenario were:

- Is it possible to design a neural network model to simulate the editing and imputation of a human?
- Can such a neural network be trained from a subsample of 2,000 objects for which raw as well as edited records are available?
- How well does a trained neural network edit and impute when applied to the remaining 10,000 raw records?

A Synthetic Population Model

Real data are not easily available for a researcher outside statistical agencies. In some situation, synthetic data can be used as a second best solution. Synthetic data have the great advantage that it renders a much closer experimental control.

For this study, synthetic data were used. A stochastic model which reflected the statistical properties of a typical individual was established, and used to generate the needed syntethic set of 12,000 data records for individuals.

The a random sample of 2,000 individual was drawn from the population of 12,000, and the true records were established in corresponding files for the 2,000 and the 10.000 individuals. Table 1 shows the 9 distributions by category for the true values for the 10,000 individuals.

Table 1.--Population Distributions Based on True Data

	Cat. 1	Cat. 2	Cat. 3	Cat.4	Cat. 5	Cat. 6	Total
Sex	5,031	4,969					10,000
Age	3,002	3,961	2,003	1,034			10,000
Marriage	6,954	3,046					10,000
Region	4,048	2,511	2,455	986			10,000
Children	7,161	1,344	945	360	190		10,000
Education	2,998	5,958	1,044				10,000
Industry	4,247	474	283	1,171	1,814	2,011	10,000
Employment	5,329	4,671					10,000
Income	5,087	3,297	1,522	94			10,000

An Error Model of Data Collection

Collection of the data about a population introduces errors of different types, i.e., interview errors, observation errors, non-response errors, transcription errors, etc. To simulate the effects of the data collection process, a stochastic error model was developed. The model introduced random errors in the records for each individual. The probabilities for errors were specified to be similar to observed error frequencies in real surveys.

When the true data records were processed by the error model, each record was transformed to an observed or "raw" record which might contain errors of two types. The first type was partial non-response errors resulting in blank fields in the records, the second error type was called interchange errors. An interchange error was created when a true category was substituted by another category of the same attribute, f.ex. when the response to the question about sex was changed from "male" to "female."

As for the true records, the raw records were also kept apart for the 2,000 sample and the remaining 10,000 individuals. Table 2 shows the 9 attribute distributions by category based on the raw records for the 10,000 individuals. About 3,400 attributes out of 90,000 have non-response while another 900 attributes have other categorization errors. Since the two error types were mutually excluding, there were approximately 4,300 erroneous records in the file of the 10,000 raw records.

Table 2.--Population Distributions Based on Raw Data

	Cat. 1	Cat. 2	Cat. 3	Cat.4	Cat. 5	Cat. 6	Total
Sex	4,819	4,656					9,475
Age	2,775	3,629	2,014	1,147			9,565
Marriage	6,418	2,964					9,382
Region	3,908	2,430	2,375	1,041			9,754
Children	6,942	1,290	923	371	273		9,799
Education	2,878	5,551	1,180				9,609
Industry	4,084	451	266	1,135	1,777	1,979	9,692
Employment	5,257	4,643					9,900
Income	4,585	3,136	1,499	192			9,412

Neural Network Training

A neural network was used in this study to investigate if errors in raw records for the 10,000 individuals could be identified and corrected. The neural network can here be considered as a set of non-linear, multi-variate regression equations in which the fields of each raw record correspond to a set of independent variables and the outputs of the regressions are a predicted set of true values for the same object.

Because of the non-linearity, the statistical estimation of the regression coefficients in neural networks is substituted by an iterative "learning" process which determines the values of a large set of weights in the network.

It was assumed that the 2,000 raw records were expertly edited and imputed with the result that the 2,000 true records also became available. In our case the paired set of 2,000 true and raw records was used as training set for a neural network. By representing each of the 9 attributes by 6 categories, each record could be expressed as a binary vector with 54 elements. Not all 6 categories were used for each attribute. F.ex. for sex, only two categories were used while the remaining 6 always were set equal to 0. The neural network used, was a so-called feedforward net with 54 input neurons, one layer of 300

hidden neurons, and 54 output neurons. The network used sigmoid transfer functions in hidden and output neurons.

Training of such a network requires the determination of the value of about 32,000 connection weights, and a large number of iterative cycle through a training set of records (Rumelhart and McClelland, 1986).

The twin sets of 2,000 true and raw records were used to train the network through in total 300 iterative cycles.

Model Verification

After training, evaluation of the network ability to simultaneously edit and impute could be carrieds out by means of the file of 10,000 raw records which had not been used for training. By using these records as input, the trained network produced edited records, or prediction for the true records. The closer the 10,000 edited records were to the corresponding 10,000 true records, the more successful would the neural network edit and imputation be evaluated.

Table 3.--Population Distributions Based on Edited and Imputed Data

	Cat. 1	Cat. 2	Cat. 3	Cat.4	Cat. 5	Cat. 6	Total
Sex	5,127	4,872					9,999
Age	3,002	3,878	2,024	989			9,893
Marriage	7,090	2,910					10,000
Region	4,046	2,488	2,443	946			9,923
Children	7,130	1,364	918	336	146		9,894
Education	3,088	5,870	974				9,932
Industry	4,243	460	265	1,173	1,784	1,985	9,910
Employment	5,371	4,621					9,992
Income	5,029	3,352	1,427	67			9,875

Table 3 shows the attribute distributions based on the 10,000 edited records. Comparison with Table 1 shows that 580 out of the 3,400 non-responses are still unresolved while there are still 460 other errors in the edited records compared with 900 in the raw records.

We have demonstrated by means of this example how we simultaneously can edit and impute a set of raw records if we have access to, or can produce, a training set of paired raw and true records. As pointed out in the beginning of this paper, neural networks can also successfully to the editing step only (Roddick, 1993).

Using synthetic data have the advantage that the discussion about what are the true values, and how to obtain true data for comparison, can be disregarded. We can therefore make statements about this particular approach to simultaneous editing and imputation. A more interesting comparison with alternative editing and imputation procedures, would require that these procedures were applied to the same 10,000 raw records and their results compared with the neural network results.

CASE 2: Large Scale Imputation

A Population Census Case

In the previous section, it was pointed out that neural networks could be used for editing as well as for simultaneous editing and imputation. In the second case to be presented, we shall describe how neural networks can be used for imputation only. A more complete discussion is available in Nordbotten (1996b).

The 1990 Population Census in Norway was based on data from administrative registers and supplemented by a special population sample survey. In Norway, there are about 435 municipalities, and those are further subdivided into census tracts, some with only 100-200 inhabitants. The census tracts have been convenient geographical subdivisions for a number of users of statistics. In municipalities with less than 6,000 inhabitants, the population survey was extended to all, while the survey size varied with the population for those with more than 6,000 inhabitants. A few municipalities also paid themselves for a complete survey.

The 1990 Population Census was evaluated by an independent group in 1993. One point emphasized by the group and others was that, for smaller areas, many estimates could not be released for publication because of the high uncertainty associated with estimates.

The purpose of this second case to be discussed below was to investigate if improved estimates of proportions for small subpopulations could be produced by means of neural networks.

Design of the Population Census Experiment

This experiment was made possible by a cooperation contract with the Central Bureau of Statistics of Norway. The population records from a municipality with 17,326 inhabitants which paid for a complete survey, were selected for the experiment. The census record in the Population Census of 1990 was composed by one set of fields with information transferred from administrative registers and a second set of fields from information collected by the special population census survey. For the investigation carried out, 97 administrative field and 49 fields with data from the survey for each inhabitant were extracted from the population record and adjusted to the form required by the experiment. The survey fields selected, represented 10 attribute groups with from 2 to 9 categories. The 49 fields were representing observations as binary variables, while the register fields represented data as both binary, categorical and continuous variables. Other experiments have also included continuous dependent variables.

A random sample of 2,007 inhabitants was drawn to represent what the situation could have been if this selected municipality had not requested and paid itself for a complete survey. In the situation simulated, both survey and register data, therefore, existed for each of the inhabitants of this municipality. From this sample, 1,845 records were randomly extracted and named Sample 1; the remaining 162 records of the original sample were named Sample 2.

Ten neural imputation networks were defined, one for each attribute group of survey variables. The structure of all models corresponds to a feedforward neural network with 97 input neurons, one hidden layer of 25 neurons and from 2 to 9 output neurons depending on thge number of categories in the

particular group represented by the model. Each neuron in the hidden and the output layers, produced its output by converting its input by means of a sigmoid transfer function. A neural network of this type is specified by the weights used in connecting all input and hidden neurons, and all hidden and output neurons. The number of weights in the networks varied from 2,502 for the nets designed to predict 2 survey variables to 2,684 for the net which predicted 9 survey variables simultaneously.

Training the Imputation Models

The 10 imputation networks were estimated or "trained" to produce estimates of the survey variables for individual records based on the register variables of the corresponding records. Sample 1, which contained 1,845, was used as a training set. The training was done by means of a standard backpropagation algorithm (Rumelhart and McClelland, 1986). Since training is an iterative adjustment process, some of the models required several thousand iterations and some hours of computer time.

Imputing Survey Variable Values

The 10 trained imputation models were then used to impute individual survey variable values for each individual in municipality. In total, these add up to almost 850,000 imputed values which required only minutes to compute. The models did not produce binary values, but values in the interval 0 to 1. The real values were converted to binary values according to usual rules the values less than 0.5 became 0, and the remaining 1.

At this stage of the experiment, we had both observed values and imputed values for the complete population. Comparison showed that the imputed values for individuals in the training sample were much closer to the observations than for individuals in Sample 2 and Sample 3. This is a phenomenon called overfitting, and a well known problem in applications of neural network. The cause is that the network during training adjusts too well to the training set.

Comparison of Estimates

The observed values from Sample 1 and Sample 2 were added and divided with the number of records in Sample 1 and Sample 2 to obtain simple, unbiased estimates for the proportions of the population. To obtain alternative, imputed estimates, the imputed values for Sample 2 and Sample 3, and observed values from Sample 1, were added and divided with the size of the population. The relative precision, defined as the standard error of an estimate divided by the estimate, is usually used as an indicator of accuracy for sample based estimates. Statistics Norway restricted f.ex. general publication of estimates from the Population Census 1990 to those with a relative precision corresponding to 0.3 or less. With the sizes of the population of the selected municipality and Sample 1 + Sample 2, we can by means of standard theory predict that only estimates of proportions 0.008 or less, or 0.992 or larger, will have an unacceptable, relative precision.

The estimates based on a sum of imputed and observed values, have no similar sampling error. If we repeated the drawing of the training sample, the training and the imputation, we would get a basis for estimating an assumed structural sampling errors of the neural network. This has not been done in this experiment. The imputed values may, however, be biased and create an inaccuracy in the aggregated estimates. The exact biases can not be observed in a real life situation, but they can be estimated. The purpose of Sample 2 was to obtain estimates for the biases. We used ratios of estimated proportion biases from Sample 2 divided by the estimatd proportions as relative accuracy indicators for the imputed

proportion estimates. In contrast to the relative precision for unbiased estimates, the values of the relative accuracy indicators for imputed estimates are not expected to vary with the size of a subsample. In other words, while the unbiased estimates frequently are useless for small subsamples, the imputed estimates should be expected to be useful also when computed for small subpopulations if the network used for imputation have been trained on a sufficiently large sample.

Comparing Estimates for the Population of a Municipality

The relative sampling errors for the unbiased estimates and the relative biases for the imputed estimates were computed for the municipality population. From these accuracy predictions, it was evident that the unbiased estimates would generally be better than the imputed.

The particular municipality used, permitted us to test empirically the predicted accuracies. The target proportions could be computed because we had access to the set of observations for the whole population. The unbiased estimates for the municipality computed from the observations from Sample 1+2, and the imputed estimates of the proportions which were arrived to by adding observations for the individuals in Sample 1+2 and the imputed values for individuals in Sample 3, could therefore be compared with the target proportions.

The comparison test confirmed the expectations -- 27 out of the 49 unbiased estimates were closer to the target proportions than the imputed estimates.

Comparing Estimates for the Population of a Small Area

The purpose of the experiment was, however, to test the estimates for the small census tract area. Calculating unbiased estimates from the 18 individuals of this area in Sample 1+2, and the relative sampling error for these estimates, 44 estimates were predicted not to satisfy the 0.3 requirement. The corresponding number for the relative bias for imputed estimates, which exceeded 0.3, were 22.

Also for the census tract, the estimates were computed and compared with the targets. Table 4 shows the results for the first attribute group proportions. While most of the unbiased estimates, as expected, deviate significantly from the target proportions, the imputed estimates were close to the targets. When

Table 4.--With Whom Are You Living? Target and Estimated Proportions, Census Tract of 162 Individuals

Attribute	Target	Unbiased estimate	Imputed estimate
1. Nobody	0.14	0.17	0.11
2. Spouse	0.48	0.33	0.47
3. Cohabitant	0.10	0.28	0.11
4. Children	0.30	0.28	0.32
5. Parents	0.13	0.28	0.14
6. Siblings	0.06	0.17	0.07
7. In-laws	0.04	0.11	0.03
8. Grandparent-child	0.04	0.11	0.03
9. Other	0.04	0.11	0.03

considering the relative deviation, Table 5 shows that 32 of the unbiased estimates deviated from the target with more than 0.3 requirement from the target proportion, while the corresponding figure for the imputed estimates was 16. Compared with the predictions, both estimators yielded better results than expected. However, both the predictions and the observed deviations clearly point to the imputed estimator as the better for areas or subgroups of this size (Nordbotten, 1996c).

Table 5.--Predicted and Observed Relative Acuracy of Estimates for Proportions, Census Tract of 162 Individuals

Unbiased estimates:

		Observation		
		<=0.3	>0.3	Total
Prediction	<=0.3	4	1	5
	>=0.3	13	31	44
		17	32	49

Imputed estimates:

		Observation		
		<=0.3	>0.3	Total
Prediction	<=0.3	22	5	27
	>=0.3	11	11	22
		33	16	49

A Third Estimator

The investigation showed clearly that the correct selection between unbiased and imputed estimators varies from variable to variable and by the size of then population. Following a proposal by Spjøtvold and Thomsen (1987), the optimal strategy can be to use a composite estimate in which the two estimators are weighted. Optimal weights can be determined by f.ex. minimizing the squared sampling error of the unbiased estimate and the squared bias of the imputed estimate.

Such a composite estimate was computed for the census tract case discussed. For 14 of the in total 49 proportions estimated, the composite estimator gave better results than both of the simpler estimators. Unfortunately, the composite estimator in some cases also produced results which had a significant lower accuracy than the better of the two estimators of which it was composed.

Final Remarks

The two case studies reported do not permit any final conclusions about the use of the neural networks for building editing and imputation models. There are many questions still which need to be investigated and answered before neural networks eventually can be a standard editing and imputation tools. Studies have taught us that this approach may have a substantial potential. My prediction is that neural networks will be important components in production of many future, statistical products.

The two case studies give an interesting lesson in research design. The first case discussed, supports the idea of simultaneous automatic editing and imputation, as well as the adequacy of experiments based on synthetic data. Simultaneous editing and imputation will save both time and resources. The use of synthetic data gives researchers the possibility to systematically focus attention to one by one of many error factors in the editing process and study their effects. The drawback of the synthetic approach is that we may easily overlook factors and complexities of the real world.

The second case discussed, indicates how we can provide statistics from sample surveys supported by administrative data; also for small groups and areas if we can learn the relationships between survey data and administrative data from a more general sample. It also demonstrates how we can, if we are lucky, use real data in research and still have the advantage of knowing the correct answers to an ideal process for evaluation purposes. We simply assume that some of the existing information is saved for the evaluation of the process to be investigated.

Acknowledgments

This paper is based on research in the SIS Statistical Information Systems project at the Department of Information Science, University of Bergen. Part of the work reported was carried out under a cooperation contract with Statistics Norway.

References

Cheng, B. and Titterington, D.M. (1994). Neural Networks: A Review from a Statistical Perspective, *Statistical Science*, 9, 1, 3-54.

Clark, A. and Street, L. (1996). Planning for the 2001 Census of the United Kingdom, Presentation for the U. S. Bureau of the Census, Annual Research Conference 1996, Washington DC.

Fellegi, I. P. and Holt, D. (1976). A Systematic Approach to Automatic Editing and Imputation, *Journal of the American Statistical Association*, 71, 17-35.

Nordbotten, S. (1996a). Editing Statistical Records by Neural Networks, *Journal of Official Statistics*, 3, Stockholm.

Nordbotten, S. (1996b). Neural Network Imputation Applied on Norwegian 1990 Population Census Data Utilizing Administrative Registers, Department of Information Science, University of Bergen, Bergen.

Nordbotten, S. (1996c). Predicting the Precision of Imputed Proportions, Department of Information Science, University of Bergen, Bergen.

Roddick, L. H. (1993). Data Editing Using Neural Networks, Statistics Canada, Ottawa.

Rumelhart, D. E. and McClelland, J.L. (1986). Parallel Distributed Processing -- Explorations in Microstructure of Cognition, 1, *Foundation*, MIT Press, Cambridge, Mass.

Spjøtvold, E. and Thomsen, I. (1987). Application of Some Empirical Bayes Methods to Small Area Samples, *Proceedings of the 46th Session of the International Statistical Institute.*

Werbos, P. (1974). Beyond Regression: New Tools for Prediction and Analysis in Behaviour Sciences, Ph.D. dissertation, Harvard University. ■

Chapter 11

CATI-CAPI Conceptual

Chair: M. Denice McCormick Myers, National Agricultural Statistics Service

R. Jamieson

David O'Connell

Mark Pierzchala

Statistics Canada's Experience in Moving to CAI from Paper and Pencil

R. Jamieson, Statistics Canada

Chapter 11

Abstract

In Survey Operations we have been using Computer-Assisted Interviewing (CAI) to improve quality and reduce cost for several years.

In the past 3 years we have converted 14 economic surveys from paper and pencil mode to CAI. By the end of January 90% of our surveys will be using this mode of collection for annual, quarterly and monthly data.

These interacting applications have been developed to handle both telephone and the mail components for each survey. The edits include the standard interfield, historical, tolerances, balancing, and consistency.

The presentation will include:

- description of the collection process
- description of the type of edits we are using for 3 applications, including the Consumer Price Index
- the impact that moving to CAI has on the operations
  - resources
  - interviewer training/manuals
- application development
  - problems
  - sharing code. ■

A Feasibility Test of On-Line Editing for Touchtone Data Entry Collection

David O'Connell, Bureau of Labor Statistics

Chapter 11

Abstract

The Current Employment Statistics (CES) Survey is a monthly panel survey of over 390,000 business establishments. Nearly 40 percent of the units in the CES sample report their payroll data by telephone to the survey's Touchtone Data Entry (TDE) system. This automated data collection system has made data available faster and has reduced program costs. Because TDE captures the data in machine-readable form that can be easily fed into an editing system, the CES is working to integrate an editing system into the data collection system. With such a system, a series of edit checks could be run on data as they are reported. In the event that any items fail edit, appropriate feedback could be provided to the respondent. Such a system could speed up the editing process by resolving many edit failures automatically and thus reduce the editing workload of the human analysts.

Two approaches are under consideration:

- The first approach is an on-line edit message. If a report fails one of the basic edit checks, the system would speak a short message to the respondent. This message would explain the edit failure and ask the respondent to check that they entered their data correctly. The system would then allow the respondent to re-enter their data items or to add enter an explanation code.

- The second approach is a "FAX-back" system. For any reports failing one or more checks, the system sends a FAX message to the respondent. This message explains the nature of the edit failure and asks the

Abstract (Cont'd)

respondent to check their data and either call in a corrected report or FAX back an explanation. Because it is printed, the FAX would be able to explain more sophisticated edit failures than would a short spoken message.

For either approach, there are, however, a number of conceptual and methodological questions which must be addressed. These include:

- What type of question scripting/branching should be used?
- How sophisticated can the edits be?
- How will respondents react to these types of questions?
- Are respondents sufficiently knowledgeable about the data to provide accurate responses?

The paper provides a full review of these issues as well as the results of the first feasibility test of the FAX-back system.

A Feasibility Test of On-Line Editing for Touchtone Data Entry Collection

David O'Connell, Bureau of Labor Statistics

Background

The Current Employment Statistics (CES) survey is a monthly panel survey of over 390,000 business establishments. The survey publishes key economic statistics including employment, average hourly earnings, and average weekly earnings for the nation, as well as by industry, state and area. The employment estimates are closely watched by businesses, financial markets and policy-makers as a leading economic indicator.

The CES is a time-critical survey. Each month, there are only ten to fifteen days to collect and process the data before the preliminary estimates are published. Historically, most CES establishments have reported data by mail. Average response rates for mail are only 55 percent by the cut-off date for preliminary estimates. Because of the low response rate for mail, the preliminary estimates have been subject to considerable revision. Furthermore, collecting data by mail is labor-intensive and costly.

Conversion to Automated Collection

In an initiative to improve estimates, raise response rates and reduce program costs, the CES has developed and implemented a number of automated collection methods.

In 1984, CES began to develop *Computer Assisted Telephone Interviewing* (CATI), in which live interviewers call respondents and collect their data over the phone. CATI regularly achieves response rates in excess of 90 percent, but because it is relatively expensive, it is used by CES primarily as a transition mode. Larger reporters -- 50 or more employees -- spend six months on CATI and then are converted to automated self-response. Smaller reporters are converted with a single telephone call without collecting data.

Most reporters are converted to *Touchtone Data Entry* (TDE), a system CES started developing in 1986. Under TDE, the respondent initiates a phone call to a computer system and enters data directly using a touchtone telephone. Data are available to the survey in minutes instead of days, and in machine-readable format. The costs of postage, handling and key-entry are greatly reduced compared to mail, and self-response makes it much cheaper than CATI. Respondents have been very receptive to this collection mode, producing average response rates of 80 percent. Over the past decade, CES has converted 47 percent of the sample to TDE.

CES has also developed a number of other automated self-response data collection systems for reporters meeting special criteria. For respondents without touchtone service, CES developed a *Voice Recognition* (VR) system that recognizes spoken responses to the automated interview. For respondents who report for dozens of establishments each month, *Electronic Data Interchange* (EDI) enables them to transmit all of their reports at once quickly and easily, computer-to-computer. For respondents with Internet access, CES developed a secure site on the World Wide Web in 1996 that collects data and provides interesting and useful feedback to our respondents.

Features of Touchtone Data Entry

Over 170,000 units currently report by TDE, with over five thousand new units converted from Mail each month. Each TDE reporter receives a report form and instructions.

In the middle of each month, each reporter gets an "advance notice" message either by FAX or postcard, which shows suggested reporting dates and the toll-free number. When the establishment's payroll data are available, the respondent initiates a call to the computer and is guided through a brief interview. The respondent enters data by using the numeric keypad. Each response is repeated for verification. The average interview lasts two minutes, and the respondent may call at any time of the day. As the reporting deadline approaches, missing reporters are prompted with either a phone call or a FAX.

Editing Microdata

CES microdata are edited for validity, for adherence to tolerance ranges, and for month-to-month consistency. These edits are currently performed on the mainframe in batch mode. Any report that fails edit must be coded or corrected by the data analysts. Performing edits in this manner makes sense when reports are collected by mail and must be key entered by the survey's data entry staff. It makes less sense in the TDE environment, when the data enter the survey office in machine-readable form. Editing these reports on-line may achieve significant savings.

Edit reconciliation is a labor-intensive, time-consuming activity. Analysts review the editing output manually, contact the respondents when necessary, correct errors and code valid reports. There is often a considerable time lag between the time the respondent reported the data and the eventual editing follow-up contact. In addition, most "failed" data are found to be correct. Time constraints limit the amount of review time available before the initial estimates are published, and the latest data received are least likely to be edited.

TDE collection may enable the CES to address some of these editing challenges through automation. Automatic on-line editing would treat each data report as it is entered and notify the respondent immediately in the event that a data item fails edit. This notification could either be part of the interview that the respondent hears, or it could be a FAX message, generated during the interview and sent to the respondent within minutes.

One benefit of hearing an editing message as part of the TDE interview is that the respondent can make corrections or enter explanation codes at once during the same call. Possible drawbacks include the increased length of the interview and the possibility of confusing the respondent. In the event that the respondent needs to check other records or recalculate the data, it is unlikely that the editing question would be resolved in the same call.

The FAX message option would provide the respondent with a clear, printed version of the editing question. The data for the current month and previous month could be shown with the questionable items highlighted. Because some edit failures are complex, it may help to see such a message as opposed to hearing it. The message would ask the respondents to check their data sources and then either phone in a corrected report or send an explanation by FAX. Some potential drawbacks of the FAX message are that the FAX may not get to respondent in a timely manner, the confidentiality of the data may be impaired (within the respondent's company), and those explanations returned by FAX would require further attention.

The CES program has found that FAX is widely available. Eighty-five percent (85 percent) of CES respondents have FAX machines, and this medium is accepted for receiving important messages. Currently, CES uses FAX to send monthly messages to its Touchtone respondents for advance notice and non-response prompting. FAX is also used to send CES informational materials during solicitation and Mail-to-TDE conversion.

Feasibility Test

Before proceeding with developing an on-line editing system, certain methodological concerns must be addressed. The respondents' reaction to receiving on-line edit messages is the most important factor in determining whether this is a useful area for development. Therefore, an initial feasibility test was conducted to determine the following: respondents' willingness to receive this type of message; their ability to understand the message; their responsiveness to the message; their ability to provide corrections or explanations; and the timeliness of their responses.

The objective of the feasibility test was to determine TDE respondents' reaction to on-line edit messages. The approach selected was to use:

- edits in the CES CATI system to edit recently entered TDE data,
- a word processor to customize editing messages, and
- a FAX machine to send the messages to the respondents.

Therefore, the FAX editing option was the one that was tested.

The test sample was selected from the TDE units in our Atlanta Data Collection Center. During the test period in May 1995, 400 units reported data. These reports were edited with the CATI system, and approximately 80 failed at least one edit. Of these, 20 met the research criteria of having a valid FAX number, an edit failure that was not resolvable by human inspection, and a generally "good attitude" toward the survey (respondents who seldom needed prompting and had been receptive to past editing phone calls).

We experimented with three different treatments. In the first treatment, the analyst called five respondents to resolve the edit failure. They were informed them that we were sending a sample of a FAX message that we were considering in place of such calls, and we asked for their reactions. The feedback was positive. Respondents said that they understood the message and would view it as important. Respondents said they would respond to such a FAX within a day or two, and were very willing to FAX-back responses. One respondent expressed some concern about the fact that the BLS FAX message showed the company's payroll data, which could be viewed by other employees in her immediate office.

In the second treatment, six respondents were called only to notify them that we were sending them a FAX message about their data. Then the message was FAXed and we awaited responses. Of the six reporters, three FAXed explanations; one called us to say that she refused to check her data; and two did not respond.

In the third treatment, 9 reporters were sent the FAX message without pre-notification. Of these, seven FAXed back explanations; one phoned in a corrected report; and one did not respond.

Summary

Combining the results of the 2nd and 3rd treatments, 80 percent (12 of 15) responded to the message in some way: ten by FAX; one by phoning in corrected data; and one by telephone (a refusal). All responses came within three working days. Most FAXed explanations were sufficient for the analyst to code the data. Two required further explanation by phone. While not statistically significant, the results of this test suggest that some of our TDE respondents would be willing to receive FAX editing messages, would understand them, and would provide useful, timely responses.

Future Developments

The results of the first feasibility test are encouraging enough to consider further work in this area. One of the next logical steps is expanded testing, to obtain statistically significant results. Another is to truly automate the process by programming various edits into the TDE system and utilizing the system's existing FAX capability to create and send FAX messages automatically. The messages themselves also must be refined, so that they are clear and appear official. The data confidentiality issue must be addressed, perhaps with the use of a "consent" flag. Finally, we must consider ways to handle the responses the system will generate. Because of its potential to save considerable resources and improve the quality of data collected, on-line data editing may one day become an integral part of the TDE system. All of the lessons learned through these tests are also contributing to on-line editing screens and prompts in the Internet collection system. ■

Note: Any opinions expressed in this paper are those of the author and do not constitute policy of the Bureau of Labor Statistics.

Chapter 11

CAI and Interactive Editing in One System for a Survey in a Multi-mode Environment

Mark Pierzchala, Westat, Inc.

Abstract

The coordination of Computer-Assisted Interviewing (CAI) and interactive editing for one survey is discussed. Computer-Assisted Interviewing seeks to obtain clean data at the time of data collection. CAI controls the routing depending on the responses recorded during the interview, ensuring that interviewers ask appropriate questions. Other kinds of edits, for example consistency edits, can be placed into the interview and reconciled with the respondent. Interactive editing is a form-by-form review of all edit failures after data collection. It speeds the review by giving immediate feedback to data editors on their actions, allowing them to clean a form in one session. It eliminates the hand inspection of paper forms before data entry and the need for keypunch of corrections after the edit. Interactive editing may take place after CAI or after data are gathered by paper questionnaires. A third form of data editing, top-down editing (also known as macro-editing or statistical editing), is briefly discussed, including its effect on how interactive editing is done. This paper is based on work done in the Blaise system from Statistics Netherlands on the large and complex June Area Frame survey conducted by the National Agricultural Statistics Service (NASS).

In developing a dual-mode instrument for both interviewing and data editing, developers must keep in mind the differing needs and preferences of the interviewers and the data editors. In NASS, it is necessary to additionally adhere to processing conventions that were developed over the years. These conventions were developed separately for batch processing of paper-collected data and for Computer-Assisted Telephone Interviews (CATI) using different systems. Blaise effectively and efficiently integrates these two modes of survey data processing in most respects. In NASS' implementation, there is some additional design and programming work needed to mediate between the two modes of use, due to the previously adopted conventions. On the other hand, the use of one system for both modes eliminates much double programming that is done when two different systems are used and presents a large net savings in preparation effort. There are also some methodological gains.

CAI and Interactive Editing in One System for a Survey in a Multi-mode Environment

Mark Pierzchala, Westat, Inc.

The June Area Frame Survey

The June Area Frame survey is conducted annually in all of the 48 continental states in the United States. In a two week period in early June, 120,000 people are visited by 1,500 interviewers in almost every county in the United States. About 50,000 of these contacts result in full interviews. By the end of June, livestock inventory and grain stocks are released, with other commodity estimates following in July and August. Additionally, subsampling for down-stream surveys starts around June 20. To realize this statistical production schedule, a clean data set must be produced within a few days of the end of data collection.

Interviewers visit land areas called segments, with an aerial photograph in hand. Together with the respondent, they draw off land in the segment operated by that person on the aerial photo. This is called a tract. If the tract is agricultural, they conduct a full interview. They mark off fields in the tract on the photo, and determine what is in each field. As interviewing is done at the height of the planting season, the farm respondents are usually very busy. Interviews are conducted in crop fields (much dirt and dust), brilliant sunshine, blistering heat, perhaps in a downpour, or in the barn (more dust).

The June Area Frame dual mode interviewing and editing instrument is complex, encompassing five levels of organization (segment, tract, field, crop, crop utilization), with over 7,000 possible questions (usually about 100 are asked), 8,900 edits, and thousands of computations and routing instructions. There can be a maximum of 78 tracts per segment and up to 99 fields per tract though most tracts have fewer than 20. Within a field, there can be an indefinite number of crops, though it is extremely rare to have more than two. There are 44 versions of the June Area Frame survey -- one for each of NASS' state offices. Each state has its own agricultural commodities that are surveyed, edit limits, question wording, and units of measurement. One dual mode instrument is driven the 44 different ways, at runtime, by an external Blaise database of state-specific meta data. It takes minimally 800 pieces of information to specify a state's version, and up to 1,200 specification items for more complex states. Fortunately, NASS has organized this so that every state's questionnaire can be thought of as a variation on one master theme.

NASS conducts this survey on paper but started a small Computer-Assisted Personal Interviewing (CAPI) -- three interviewers -- and interactive editing project (1/2 the sample) in Indiana in June 1993, from the Research Division. In 1994, both Indiana and Pennsylvania conducted interactive editing on all samples, with Indiana implementing CAPI with seven interviewers and Pennsylvania using two CAPI interviewers. In 1995, Indiana used all interviewers for CAPI and interactive editing on all samples, while Pennsylvania and Wyoming conducted interactive editing on all samples following paper data collection. In 1996, both Indiana and Wyoming conducted interactive editing on all samples

following paper data collection. CAPI research for the June Area Frame survey was dropped after 1995. From 1994 on, the instrument was constructed to handle all versions, even though it was used only in a few states. In 1996, the instrument was still capable of CAPI but it was not used that way. The current plans for 1997 are to implement interactive editing for this survey in 10 - 26 state offices, with implementation in the remaining states in subsequent years. Blocks needed only for CAPI will be dropped.

There will be no further attempt at CAPI for the June Area Frame survey in NASS for the foreseeable future. This survey's additional hardships of using CAPI outside in bright light or bad weather along with the tight timeframe of the June Survey is too stressful for a large implementation at this time. A small research project using laptop computers for less stressful surveys is continuing.

Blaise III from Statistics Netherlands

The Blaise III system from Statistics Netherlands is used by many organizations worldwide. It is explicitly designed to integrate many of the front end survey processing needs, including data collection (CATI, CAPI, and Computer-Assisted Survey Interviewing (CASI)), interactive editing, forms-based data entry, data and survey management, tabulation, and metadata management. The recent introduction of Maniplus, a shell utility, extends the system's functionality to tabular top-down data editing and laptop CAPI management among other things.

Coordinating CAI and Interactive Editing

Even though Blaise integrates many different survey tasks into one system, the different users have their own needs and preferences. The following describes how these needs and preferences are accommodated, either naturally in Blaise, or by additional planning and work in NASS. In coordinating CAI and interactive editing, the application developer must keep in mind the needs of the users. They should also think of the optimal implementation of development standards and be prepared to change them as technology progresses.

Screen Presentation

The default screen presentations of Blaise for interviewing and editing have served well for their respective roles.

Blaise is a forms based interviewing system. Its default presentation features a split screen with question text appearing at the top and a *page* or *form* of data cells appearing at the bottom. The question displayed at the top of the screen corresponds to the position of the cursor in the *page*. Interviewers are instructed to read the question text verbatim from the top of the screen and to verify that they have typed in the correct response in the page. They can also see answers to previous questions. The forms based approach has proved to be very popular with interviewers in NASS, in part because it combines the best aspects of CAI and a paper form style of presentation.

For data editors, the default in Blaise is to use the whole screen presentation for data display. The assumption is that the editors are already familiar with question meaning and would like to see more data at one time. For most sections of the June Area Frame survey, this is a good assumption. For the more complex parts, such as the fields table, it is not. The data editor in this case can toggle the screen display

to use part of the screen for question display. Alternately, the data editor can display the question text and data definition in a pop-up window.

Question Identifiers

Different kinds of users may prefer different question identifiers. *Question names* are used in Blaise program logic to direct flow and to set edits. They also have their important uses on the Blaise screen. The interviewer needs the question name to be a *concept name*, for example, *TotalHogs* or *Wheat.Planted*. In an edit failure, it is the question name that the interviewer must choose in order to fix a problem. Thus the question name must be fully understandable in the heat of the interview. On the other hand, the data editor, usually an agricultural statistician, may prefer to see an item code as a question identifier. In NASS, the item code is a question number used for data entry. For *TotalHogs* the item code is 300. Both preferences are met by employing the concept name as the Blaise question name and the item code as the question tag. Both the tag and the question name appear on the screen next to the question's value, in the form (or page) part of the screen. A side benefit to using a concept name as the question name is that it makes for more readable code for the application developer.

Freedom of Movement

Blaise gives the data editor complete freedom to move anywhere in the instrument, regardless of routing constraints. This is especially necessary if data are collected on paper and routing rules were not applied correctly. The editor can go where necessary in order to correct or remove the data. For the interviewer, forward movement is completely controlled by the system. Backward movement is free, but the interviewer can never go off route.

Edit Message Presentation

Edit message presentation, for one edit, is the same for data editors and for interviewers. Blaise displays the edit message in an edit window, as it is written by the applications developer, and an appropriate list of possible question names and values to correct. If the edit is soft, then the user can suppress it with the <*s*> key, or fix it, whichever is needed. If the edit is hard, it must be corrected. The user can scroll in the question list. By placing the cursor on the required data item, and pressing the <*Enter*> key, the user is moved to the data item, no matter where in the instrument it is. Once the item is fixed, by pressing the <*End*> key, the user is moved to the next appropriate question.

Multiple Edits in a Form

An interviewer encounters one edit at a time, and must deal with it on the spot. The interviewer never sees more than one edit at a time. The data editor, on the other hand, has the choice of reviewing all edit messages at once, or displaying them one at a time while moving through the form. The editor can determine in which order to dispose of them and can leave them temporarily unresolved if necessary. The editor can look at all errors, just the soft ones, or just the hard ones.

Severity of Edits for CAI and Interactive Editing

There are three kinds of edit severity that Blaise keeps track of on a form-by-form basis: routing, hard edits, and soft edits. In interviewing mode, the routing is enforced dynamically. The interview is guided along the correct path as determined by the responses. In edit mode, routing is enforced pas-

sively. A routing edit marker is displayed if a cell has an answer when it should be off the route, or if it is empty when it should be answered. In edit mode, a routing edit must be cleaned up or the form is considered dirty, while in interviewing mode, the route is never violated.

Any other within-form edit can be implemented. These include additivity, relationship, consistency, linear, ratio, range, and other kinds of edits. Whether an edit is hard or soft in a Blaise instrument depends on its toggle. The key word CHECK denotes a hard edit while the key word SIGNAL denotes a soft edit. It is possible to make an edit soft in the interview and hard in data editing. For example:

```
IF CADI THEN        {data editing mode}
   CHECK            {hard edit toggle}
ELSE                {any data collection mode}
   SIGNAL           {soft edit toggle}
ENDIF
Edit statement
```

will make the edit hard for data editing and soft for interviewing. The question is when to have hard edits in the interview. Let us examine, first, an easy-to-handle edit:

```
CHECK
Planted >= Harvested
"Planted acres must be greater than or equal to Harvested acres."
```

In Blaise, the edit is stated in terms of what is correct. The edit will be invoked only if harvested acres are greater than planted. This edit is well within the interviewer's and respondent's understanding, it points definitely to an error, and it is easy to fix. Thus it is kept as a hard edit in the interview as well as in data editing mode. A more difficult edit follows:

```
Tract.Acres = Field[1].Acres + . . . + Field[99]. Acres
"Tract acres must equal the sum of field acres."
```

This edit states that the tract acres must equal the sum of all the fields within the tract, for up to 99 fields. This edit is understandable to the interviewer and the respondent, but may be difficult to correct during the interview. The required accuracy is to the tenth of an acre, and the respondent may not even know the size of the fields to the tenth. Indeed, the respondent may not know the acreage at all and thus it would be foolish to stop the interview when the respondent is incapable of giving more accurate information. This edit is soft in the interview but hard for data editing. The interviewer or data editor can use measuring devices on the aerial photo after the interview in order to determine the correct acreage of the fields. The edit should be invoked during the interview because the actual fixing of it may not be too difficult in most situations. However, for some interviews, fixing it may be very difficult, thus this edit is kept soft.

There are other edits which are hard in the traditional batch edit system. Whether they are hard or soft in the interview depends on the judgement of the developer, perhaps in consultation with other statisticians. If there is any chance that a hard edit can in truth be violated, then it is softened during the interview. In the June Area Frame survey, most edits which are hard in the traditional batch edit are hard in the interview.

Don't Know and Refusal

There are some questions that farmers are not able to answer or refuse to answer. The CAPI instrument allows for *Don't Know* and *Refusal* for selected questions. For example, the total land question can take the *Don't Know* response in CAPI. The summary system, on the other hand, expects important items to be filled in. The method of imputation is hand imputation by the editor during the edit process. Sources of information available to the data editor include the respondent's previous survey forms if any, and since this is an area frame survey, interviewer observations from the road. It is felt that manual imputation of observed values, or values from previous years, is more accurate than automated imputation. The result of this is that questions are not allowed to have a *Don't Know* or *Refusal* response in edit mode and thus there is a subtle difference in data definition between the two modes. The developer has to put in edits to flag these nonresponse codes in the edit.

Zero and Empty

In data collection, according to NASS standards, an *EMPTY* cell on the route represents a question that has not been asked yet. The interviewer must enter a number, even if it is zero. In paper data collection, and in data editing, an *EMPTY* cell represents a zero. This represents another subtle difference in data definition between modes of use of a Blaise instrument in NASS. The instrument has edits that allow the *EMPTY* in data editing, but not in data collection. An enhancement NASS would like to see for its dual use instruments is a key word such as *CADIEMPTY* that would be written as part of data definition. This would tell the system to allow *EMPTY* for data editing, but require a response for interviewing.

Screening Questions

On paper forms, and in the CAPI instrument, there are screening questions at the start of each section. For example, in the hogs section, questions are asked to determine if there are hogs on the farm. If not, then the rest of the section is skipped. However, when data are collected on paper forms, the screening questions are not key punched in NASS. This is done to save the data entry people a few key strokes. Routing in CAPI is dependent on the values of the screening questions as entered by the interviewer. Routing in data editing mode is dependent upon the presence and values of the commodity items themselves. In this latter situation, it is up to the instrument to determine, from the values in the section, what the values of the screening questions should be, and impute them. NASS should consider key punching the screening questions.

Switching Between Modes

Blaise can allow the user to switch between its four modes of operation. It also allows the application developer to disable this feature. The mode switching is disabled for interviewers but is enabled for data editors. The editors may find it advantageous to switch to the interviewing mode to follow the course of an interview.

CAI Administration Blocks

CAI administration blocks include blocks of questions that are necessary to administer a CAI questionnaire. As implemented in the June Area Frame Survey, these blocks include a *Name and Address* block for the respondent and partners, an *Appointment* block, a *Nonresponse* block, and a *Time and*

Date Stamp block. These blocks have marginal use in data editing, but in the multi-mode nature of Blaise, they are present in edit mode as well as data collection mode. This takes up space in the data set and increases the number of screens it takes to present the data. This has not been much of a problem in practice. The developers give the data editors jumping points that get them to the commodity data quickly, and save them the trouble of paging through many administration screens. Navigation through edit messages in Blaise is superb and gives another possibility to get to the problems without paging. It is also possible to hide the administration screens in edit mode, but this has not been done.

Mouse and Operating System

Blaise III allows functions to be executed either by mouse or by function keys. The mouse is turned off for interviewers but left on for data editors. The editor has a choice of which to use.

Blaise III is DOS-based, but can be operated in a DOS window of Windows 3.1, Windows 95, or Windows NT. The interviewers are generally set up in a straight DOS environment. Their job is to conduct interviews, they do not need to switch between tasks. The data editors usually operate out of a DOS window in Windows 3.1. The data editors in NASS are agricultural statisticians with a wide variety of duties to carry out. When the Blaise edit is in a DOS windows, the data editor can switch to other tasks without having to close the Blaise application.

Data Read-In from Paper Forms

When data are collected on paper, it is necessary to read them into the editing instrument. If data are collected only in CAPI, then edited in the same software system, then no data transfer is necessary. Data entry in NASS is done in another system. The kind of data entry is known as item-code data entry, where the 3 digit question number is entered followed by its value. The data are put out into 80 column records where many physical records represent one logical record. The Manipula utility is used to read data into Blaise format. For each data item, a mapping from the Blaise record to an item code must be made. This mapping is easily accomplished by using the item code as the question tag for most of the June Area Frame instrument. Cameleon, a meta-data utility, generates the necessary Manipula program to do the data read-in. However, for the grain stocks table and the crops table, much of the metadata are held in an external Blaise file. A separate Manipula program is required to read these data in. Since the meta data in this case are not part of an instrument's code, Cameleon cannot generate this Manipula program. This part of the data read-in is accomplished by using the external files to mediate the meaning of each data item for each state.

Farmer in More than One Segment

A data collection challenge is that a farm respondent may operate land in two or more segments. However, the traditional unit of processing is based on, and the paper questionnaire represents, a tract. The questionnaire has farm-level questions and tract-level questions. When data are collected on paper, and the interviewer finds out that the respondent operates land in two or more segments, she collects farm-level questions on one just form. For the tract-level questions, she shifts back and forth between paper forms.

Instantly shifting between forms was not so easy to do in CAPI in Blaise in 1995. One method of handling a respondent in 2 segments that was tried in 1995 was to have two instances of all tract-level blocks in the instrument, including the massive field table (each instance of which had 5,500 possible

questions). Mechanically this worked in the interview, and was easily implemented by the instrument authors, but was not optimal for the edit. This scheme greatly enlarged the instrument in terms of numbers of questions, edits, and screens for a situation that rarely occurred. It slowed down batch processes such as combining forms from different CAPI machines into one data set. Another problem with this approach is that there is no upper limit on the number of segments that a farmer might appear in. As soon as two instances of tract-level questions were allowed, someone might find a farmer who operated land in 3 or more segments. If data for two tracts were collected in one CAPI form, there was the problem of transferring data from one form to a second. In future, this difficult data collection problem would be handled through the use of the new shell utility Maniplus. This utility makes it far easier to shift between forms for tract-level questions for one farmer in multiple segments, and then to automatically transfer farm-level data from one form to the rest.

Benefits of One System for CAI and Interactive Editing

While the above paragraphs enumerate some differences between implementation of editing and interviewing in one system, there are powerful reasons for the adoption of one system for both tasks. Despite the differences that must be accounted for, 80 percent or more of the program code is applicable between the two modes of processing, including data definition, routing, edits, and data base structure. The percent of shareable code would be higher if NASS were to adopt different conventions to eliminate the difference in data definition between the two modes. Though there is more programming code than for CAPI alone or for interactive editing alone, it is still far less code than if two programs were written in two systems. Where there are two systems, additional conversion code must be written or generated to transfer data from CAPI to edit. Data conversion itself can be problematic between two different systems. Conventions for handling interviewer remarks or nonresponse codes such as *Don't Know* or *Refusal* may be incompatible, or may require a lot of work to bridge. Database structure for a complex multi-level instrument such as the June Area Frame survey may only be handled adequately in a data collection system.

With everything in one program code, the developers are forced to make explicit methodological choices between what happens in data collection and what happens in edit. If there are two systems, sometimes these choices are made implicitly without anyone realizing it.

The Implementation of Top-Down Edit Methods

NASS for some time has had paper-based top-down editing tools. These are SAS programs that generate various kinds of data listings, tables, and other macro views of the data that allow the agricultural statistician in the state to quickly spot problems. Recently, a SAS-based top-down system has implemented most of these ideas in an interactive mode for the Hog Report. This system, called IDAS (Interactive Data Analysis System) was demonstrated at this conference (Hood, 1996). For the Hog Report, it is the intention of NASS to implement this system in 30 quarterly states and to retain most of the SAS paper-based reviews for other states which do the survey only annually. NASS (Knopf et al., 1996) has just embarked on a systematic review of the overall Hog Report editing process, from data collection, through a micro-level interactive edit, and the macro-edit tools. Knopf and colleagues have made explicit decisions about which edits to invoke at each stage of the process. Some edits have been eliminated or consolidated. A few of the range edits are no longer implemented in the micro-level, post-collection, interactive review, but are invoked in CATI and at the macro-level.

A new utility in Blaise, Maniplus, is recently available. Things that can be done with it include the production of a laptop computer CAPI management system, a data and survey management system for the in-office processing of forms, and tabular top-down editing methods. The latter function has been demonstrated in the form of interactive data listings though there is no limitation on the form of tabular review. It is possible to scroll through a list of records, sort them in any order, invoke a form, change the data, get back into the interactive table, and immediately see the effect of the changes on the table and on estimates. Maniplus does not have a natural graphical review capability as the IDAS system does. On the other hand, Maniplus knows the metadata of Blaise instruments, and can read survey data directly from the Blaise database. This allows for top-down review of data without having to convert data to a different system. NASS should consider its use for interactive data listings, getting away from the paper-based data listings it plans to continue.

References

Knopf, D.; Anderson, C.; Apodaca, M.; Pallesen, M.; Prusacki, J.; and Tesky, M. (1996). Report of the Hog Editing and Analysis Team, internal NASS report.

Hood, R. (1996). Improving the Quality of Survey Data Through an Interactive Data Analysis System, demonstration at the *Data Editing Workshop and Exposition*, Washington, DC. ■

Note: The work represented in this paper was done while the author worked for the National Agricultural Statistics Service (NASS), U. S. Department of Agriculture. The views and conclusions are the author's, not those of NASS or Westat, Inc. The author thanks several people in NASS and Westat, Inc. for reviewing and commenting on a draft of this paper.

12
Chapter

Statistical Techniques -- II

Chair: Linda Stinson, Bureau of Labor Statistics

David A. Pierce ◆ Laura Bauer Gillis

Peter Ochshorn

James Kennedy

Time Series and Cross Section Edits

David A. Pierce and Laura Bauer Gillis
Federal Reserve Board

12
Chapter

Abstract

Much editing of data from repeated surveys and reports is based on comparing the current or incoming value for a variable or item to that variable's value for the previous week, using a set of published **tolerances**. The previous value represents an estimate or forecast of what the current value would be in the absence of error or unusual circumstance. This paper investigates two generalizations of this editing method, which both involve incorporating information beyond that contained in the previous week's value. One of these is to base this estimate on the item values from a **cross section** of similar institutions in the current time period which have already reported, and the other is to calculate a forecast based on the **time series** of past values of the item. A composite estimate combining these two methods is also presented.

These methods are applied to data from the major deposit reports submitted by commercial banks to the Federal Reserve System. Edit simulations are performed to measure the improvement from this approach (in terms of fewer edit exceptions which are correct and/or increased detection of errors), which is found to be substantial for some items and size groups. Efforts thus far to implement these enhancements are described, and possible further generalizations are mentioned.

Time Series and Cross Section Edits

David A. Pierce and Laura Bauer Gillis
Federal Reserve Board

Background and Introduction

Data for the U.S. Money Supply are regularly transmitted to the Federal Reserve System by commercial banks and other financial institutions at weekly and other intervals. A major vehicle for this transmission is the "Report of Transactions Accounts, Other Deposits and Vault Cash," or simply the "Report of Deposits," on which banks and other financial institutions report weekly data for 25 deposit categories and related items. Based on these data and on similar information contained in other reports, the money supply measures are constructed and reserve requirements are maintained.

The money and reserves figures are important both as barometers of economic activity and in enabling the Federal Reserve to perform its economic stabilization and bank regulatory functions, and it is essential that the data submitted on the Report of Deposits and other reports be reliable and of high quality. To ensure their accuracy, all such data are subjected to numerical edits to detect unusual or deviant values. These edits are to two general types, **validity** edits to ensure that adding-up and other logical constraints are satisfied, and **quality** edits based on statistical or distributional aspects of the data.

The most commonly used quality edit involves the comparison of an incoming weekly figure to the previous value of that variable (in both dollar and percentage terms), using a tolerance band constructed about that value. The **tolerances**, or half-widths of the tolerance bands, are determined from previous estimates of the variable's distribution, in particular measures of spread, and are published in a Technical Memorandum or "Tech Memo" (Federal Reserve Board, 1993). An edit "exception" occurs if the incoming value falls outside this tolerance band; when this happens, the reporting bank or other institution may be contacted for verification or correction. All tolerance-table comparisons are made (and edit exceptions generated) by machine, whereas the decision to contact the respondent is made by data analysts. The editing is done at both the Federal Reserve Board and the 12 Federal Reserve Banks.

Edits are in essence hypothesis tests, and both Type I and Type II errors can occur. A major task in setting edit tolerances is to ensure adequate sensitivity without generating unnecessarily large quantities of "false positive" edit exceptions. It is because of the large number of these exceptions that editing at both the Reserve Banks and the Board is currently quite labor intensive. All exceptions are reviewed by data analysts who must decide which are to be referred to the respondent institution for verification or revision. At the same time, a large majority of the data errors are not caught by these edits, based on the historical record of revisions submitted by respondents (they may be detected by other edits at a later date). There is consequently a need both to increase the sensitivity of the edits and to streamline the data editing process.

The value to which the tolerances are applied is in effect an estimate or forecast of the incoming figure that is being edited in the absence of error or unusual circumstance. By basing this forecast or estimate on information beyond that contained in the previous week's value, we obtain the generalizations of the current editing method that are investigated in this paper. One generalization is to base this estimate on the item values from a **cross section** of similar institutions in the current time period which have already reported, intending to capture economic, institutional or calendar movements which tend to affect similar respondents in a similar manner. The other is to calculate a forecast based on the **time series** of past values of the item for that respondent, including possibly last month's or last year's figures in addition to the one for last week as in the current procedure. A composite estimate combining these two methods is also investigated, the idea being that each method may incorporate information not captured by the other. (We also generated a composite of the cross section and current edits).

The paper's focus is on the data submitted on the Report of Deposits, also known as the Edited Data Deposits System (EDDS) Report. We investigated four of the more important items on this report, total transactions deposits, total savings deposits, and large and small time deposits. The study was motivated by the desire for greater automation in the Federal Reserve Board's Division of Information Resources Management, which carries out the edits. The improvements resulting from the study are being incorporated into a new software package called DEEP (Distributed EDDS Editing Project), for interactive editing on the PC. (For more detail see, Pierce and Gillis, 1995.)

Our results vary greatly according to item, entity type (e.g., commercial bank, credit union, etc.), and the amount of data in an institution group -- the latter being important for reliable cross-section estimates. In most cases we find that, with sufficient data, the cross section approach is as reliable as the current editing procedure. For total transactions deposits almost uniformly, and for total savings deposits for most commercial bank categories, time series modelling plays a significant role in the edits.

The following section of the paper discusses in greater detail the methodology underlying the different data editing approaches investigated. The third section then describes a set of edit simulations we performed with each of the five types of edits studied, and presents the results of these. Based on the simulation results, we provided a set of recommendations for experimental edits for DEEP, for each entity type and item, which have recently become operational.

Methodology

Given a variable or item of interest, many data editing procedures can be characterized as first generating a forecast (a point estimate) of the incoming value for that item, next applying a tolerance to the forecast to form a tolerance interval (an interval estimate) for the incoming value, and then flagging that value if it is outside the tolerance interval. In the current editing framework, that forecast is taken to be the previous week's item value, and the tolerance is as given by the Tech Memo (Federal Reserve Board, 1993). In this section the two generalizations to the forecast noted above are presented, along with composite procedures, after first describing the data and framework used.

Choice of Items and Statistical Form

The current approach to editing data from financial institutions is to subdivide them into homogeneous "cells," which are combinations of an institution's size group, entity type, and geographic location. There are six size groups for commercial banks and a smaller number of size groups for each of

the other entity types, which are credit unions, savings and loans, savings banks, agencies and branches of foreign banks, and Edge and Agreement Corporations. The geographic locations are defined in terms of 12 Federal Reserve districts.

There are thus a great many edit cells, and to make our task manageable, and to achieve comparability with the current edits, we have simplified this study in the following ways:

- Staying with the **same cells** of the current EDDS edits. This will facilitate assessing the effects of the cross section estimates, model forecasts, and composite procedures. We recognize that more sophisticated groupings into cells may enhance the performance of the edits and plan to work with these in the future. Also we have eliminated all acquisitions and mergers from the institutions studied and have placed "credit-card banks" in a separate group.

- Maintaining the **same tolerance** widths as currently (applied, however, to the time series / cross section estimates that we generate, as well as to the most recent value as currently done). This may at first seem unnecessary, since standard deviations, percentiles, and other aspects of the distribution can be determined from either the cross section data or the historical model. However, such calculations can sometimes be unreliable, especially with cross sections without at least several hundred institutions in a group, as we are working with the extremes of distributions. And as with the cells themselves, keeping the current cell tolerance-interval widths facilitates comparisons among procedures.

We have also confined our attention in this study to the smaller institutions ("Priority-3" or P-3 institutions), where there may be the greatest potential for human resource savings from this approach. (Essentially this excludes the largest three size groups for commercial banks and a portion of the largest size group for other entity types.) For these institutions, we have examined the following items:

Total transactions deposits	Large time deposits
Savings deposits	Small time deposits.

Current EDDS editing is performed with both dollar and percentage changes of the item being edited, with both required to exceed tolerances ("and" condition) for an exception to occur. The modifications outlined in this report are only for percentage changes; the Tech Memo tolerances continue to be applied to the dollar changes. There are several reasons for choosing percentage changes as the focus. Since they are used in current edits, the present edit cells and tolerances can be employed, and comparisons with current procedures can be made. They (or their annualized versions, growth rates) are also used in other analyses, such as with the Small Bank Sample of early reporting institutions. They are more homogeneous than dollar changes among different sized institutions, so that fewer edit groupings should eventually be needed. Percentage changes were found to be more sensitive to reporting and other errors than ratios to other items such as total deposits, which change with the denominator as well as the numerator and moreover present difficulty when the denominator was zero.

Cross Section Edits

Period-to-period edits compare an institution's current value for an item to the previous period's value. However, useful additional information may be contained in the current values of that item for other institutions that are similar to the one being edited. For example, if most of the institutions in a

group experience a surge in large time deposits in a given week, then it would probably be inaccurate to list them as exceptions simply because they were outside the EDDS tolerances. Conversely, a very small change that week in large time deposits for a particular institution in that group may be suspicious even though current period-to-period tolerances would not be exceeded.

Cross section edits are carried out by examining the distribution of values (here, of percentage changes) for institutions within a homogeneous group, and listing as exceptions any values that were unusual compared to that distribution. Ordinarily one would calculate the mean and standard deviation of the percentage changes and flag those that were farther away from the mean than (say) two or three standard deviations; but in the present study we modified this set-up in two ways. First, because extreme values (the ones we hope to detect) would themselves influence the mean to which they would be compared, we "trimmed" the mean by eliminating the largest and smallest 5 percent of the values before calculating the estimated mean. Second, more observations are required to form a reliable estimate of the standard deviation than of the mean, and since most of the cells or groupings of institutions were too small for this, we chose to use multiples of the current EDDS tolerances as proxies for the standard deviations. As noted earlier, an additional advantage of this practice is to facilitate comparisons with the current edits.

One difficulty in using a cross section edit is that the data for an editing group need to be available in order to calculate such quantities as the average percentage change for that group. But the data for Priority-3 institutions are not due at the Board until nine days after the as-of date; and since timely estimates of the monetary aggregates and required reserves are needed, the editing process cannot be postponed this long. Our solution to this is to wait until a large enough fraction of the institutions have reported, and to form the distributional estimates (the trimmed means in this case) from the data available at that time.

For the EDDS data, more than half of the P-3 institutions' records are received by the Federal Reserve Board on the Thursday night following the as-of date (the previous Monday, on which the statement week ends), with the majority of those outstanding arriving by Friday night and the few remaining ones by the following Wednesday. For this study it was, therefore, decided to start the cross section editing on Friday morning. In either case, the trimmed mean estimates initially formed are not modified when more institutions have reported, in order not to confuse the editing process.

Some of the editing cells contain only a small number of respondents (and an even smaller number reporting by Friday), so that the estimated mean for those cells may not be very reliable. We required a minimum of 50 available observations in order to use the cross section estimate by itself. If the number of available observations is less than 50 but at least 20, a composite of that estimate and the previous week's value for the institution is employed, and with less than 20 the previous week's value alone is used.

The cross section edit is performed by comparing the deviation between the observed and the estimated percentage changes to the current EDDS edit tolerance for the item. As noted earlier, if the percentage-change condition is violated, then a second comparison of the magnitude of the dollar change versus its tolerance is performed, and the item is flagged only if both sets of tolerances are exceeded. An exception to this is that, as is done with the current edits, when the item changes from zero to a nonzero value or vice versa, the current dollar-change edit tolerances are applied without any adjustment.

Time Series Edits

These edits are based on time series **models**, which predict or explain an item's present value in terms of its past history. This usually involves the immediately previous value, on which the current edits are based, and often additional values as well, such as last year's. To the extent that these more distant values are important in predicting the incoming value, more sensitive edits should result from taking them into account.

Editing using a time series model for generating forecasts of percentage changes implies that a historical relationship exists between the item and its previous values. The "random walk" model is a time series model in which the best forecast of the current value is simply last week's value. Thus, the random walk model is implied by the current period-to-period change edits, which take last week's value as the current-period forecast around which the tolerances are applied. More complicated time series models yield forecasts which are weighted averages of several past values of the percentage change.

We first investigated the fitting of time series models for each institution separately. Some institutions' data fit the models quite well, with reductions in the standard deviation of the forecast errors (a key to the effectiveness of tolerances of a given width) of 50 percent or more, while other institutions exhibited only weak fits, or only the random walk behavior that the current editing framework already captures. Although fitting individual models is the preferable method for forecasting, it was not feasible to maintain over 7,000 models for each item edited within the DEEP framework -- at least not at this time. Thus, at this stage and for the P-3 institutions, a single time series model was fit to each editing cell's aggregate, and the coefficients from that estimated model were used to obtain an individual bank's forecast using its own previous values. While the benefits of time series modelling are reduced by doing this, the method can be easily implemented, and updated when necessary. Another constraint at present is that, because of data storage limitations, we only utilized terms in the model at lags of 1, 2, 3, 52 and 53 weeks, thus capturing nearby effects and annual seasonal influences but not, say, monthly or quarterly effects.

As an example of the model-fitting results, Table 1 provides information on time series models fit to cell aggregates of Total Transactions Deposits for three of the editing cells. Notice the highly statistically significant seasonal effect (lag 52, and in some cases lag 53). The strength of the fit declines going down the page, with the third one (Edges & Agreements, a root MSE reduction of only 9.2 percent) being not much different from the random walk model underlying current edits. On the other hand the results suggest that model-based editing may be valuable for certain commercial bank cells, for total transactions.

As with cross section edits, the deviation between the actual percentage change and the forecasted change from the time series model is compared to the edit tolerances. A tolerance exceedance both here and on the dollar change (also using current EDDS tolerances) triggers an edit exception for the record.

Composite Edits

The cross section and time series edits are based on different sets of information, past values of the institution being edited and present values of similar institutions. Thus a forecast which combined these two estimates, thereby utilizing both sources of information, may be more accurate than either one separately, and edits derived from such forecasts correspondingly more sensitive.

Table 1. --Percentage Change Models for Total Transactions Aggregates, Selected Editing Cells

----------Cell = CB, Size Group 4, Region I----------

Root MSE(orig.) = 0.0383 Root MSE(model) = 0.0211
Reduction in Root MSE = 44.9 percent

Variable	Parameter Estimate	Standard Error	T-stat	p-value
TRN_{t-1}	-0.4349	0.0483	-9.005	0.0001
TRN_{t-2}	-0.0341	0.0329	-1.039	0.2996
TRN_{t-3}	-0.1510	0.0338	-4.467	0.0001
TRN_{t-52}	0.6494	0.0318	20.391	0.0001
TRN_{t-53}	0.4668	0.0440	10.606	0.0001

----------Cell = CU, Size Group 2, Regions II&III----------

Root MSE(orig.) = 0.1067 Root MSE(model)=0.0809
Reduction in Root MSE = 24.2 percent

Variable	Parameter Estimate	Standard Error	T-stat	p-value
TRN_{t-1}	-0.2450	0.0546	-4.486	0.0001
TRN_{t-2}	-0.1160	0.0474	-2.444	0.0151
TRN_{t-3}	-0.2200	0.0486	-4.525	0.0001
$TRNt_{t-52}$	0.4922	0.0477	10.312	0.0001
TRN_{t-53}	0.1866	0.0533	3.498	0.0005

----------Cell = EA, All----------

Root MSE(orig.) = 0.0564 Root MSE(model)=0.0512
Reduction in Root MSE = 9.2 percent

Variable	Parameter Estimate	Standard Error	T-stat	p-value
TRN_{t-1}	-0.3776	0.0569	-6.632	0.0001
TRN_{t-2}	-0.1547	0.0586	-2.642	0.0087
TRN_{t-3}	-0.0449	0.0553	-0.815	0.4181
TRN_{t-52}	0.2432	0.0524	4.638	0.0001
TRN_{t-53}	0.1057	0.0540	1.955	0.0514

For a given institution (e.g., bank) and a given item, if T denotes a time series estimate (forecast) for a given week, C represents a cross section estimate, and A the actual value that is reported, then the **composite estimate** is a weighted average of T and C which is of the form

$$wT + (1-w)C.$$

The weights w and 1-w depend on the relative sizes and the correlation between the estimation / forecast errors of T and C. If these errors are given by

$$ET = A - T \text{ and } EC = A - C,$$

then

$$w = [Var(EC) - Cov(ET,EC)] / Var(ET-EC).$$

A composite forecast is thus a weighted average of individual component forecasts where the relative weights are chosen to minimize the sum of the squared forecast or estimation errors, and where the sum of the weights is one.

Using past data, we investigated a composite estimate of the cross section (CS) and the time series (TS) forecasts, denoted "CSTS", for each editing cell and each variable or item. The composite forecast defaults to the time series forecast with fewer than 20 available observations in the cell average. (With exactly 20 and using the 90 percent trim, 18 observed changes would be used in the cell estimate.)

The other type of composite edit we considered combines the cross section and the random walk forecasts (CSRW). We employed this edit when a CS edit was indicated but the sample size -- the number of observations available on Friday morning when the cell means are formed -- was insufficient (less than 50) to obtain an adequately reliable cross section estimate. For very small sample sizes (less than 20), our procedure is to revert to the use of only the RW edit.

Modelling and Simulation

To examine the relative performance of different types of edits, we conducted simulations of these edits over the 1991-1992 time period. For each cell (choice of item, entity type, size group and geographic region), we performed five sets of simulations, corresponding to the different types of edits under consideration: current (random walk), cross section, time series, cross section/time series composite, and cross section/random walk composite.

Simulation Procedure

Data preparation was a time consuming task. First, all Priority-3 reporters' weekly average data were compiled for the period from January 1986 through December 1992. While the edits were simulated only for the most recent two years, the additional data were used for fitting time series models with potential annual patterns. To avoid distortions, we eliminated all banks involved in mergers during this period. We next partitioned the data set into the editing groups or cells. We found that not all cells had a sufficient number of reporters to fit a model or to obtain reliable cross section estimates, and so some of them were combined. For commercial banks of size group 3 (total deposits between $1B+ and $3B), there were too few P-3 reporters to employ any of the new approaches. In addition, we added an editing category for known credit card banks. In total there were 40 edit cells, 37 of which were involved in the simulations.

Once the data were prepared, time series models were fit to the percentage changes in each cell's aggregate, as described above. Using the fitted model for a cell, predicted values for the last two years were generated for each institution in the cell. (Although forecasted values of the percentage change were generated for all periods, those in which a change of zero to a value or a value to zero were edited using special tolerances). Both the model-based and the zero-valued random walk forecasts were assigned to each observation in the cell. The 10 percent trimmed mean of the percentage changes was also calculated for each cell and each week of the two year simulation period, for use in the cross section edits. Since the cross section simulation employed all the data within a cell to calculate the current-period forecast, rather than the available data as of Friday morning when editing begins, the simulated results will differ from those in practice. In order to generate the two composite forecasts, the prediction errors from the original three forecasts were computed and the formulas in the section on methodology applied by institution. A cell root mean square prediction error (RMSE) was also computed.

Since the composite forecast combines the component forecasts in such a way as to minimize the sum of the squared prediction errors, we chose to estimate the appropriate weights for each bank in a cell and then to average those weights over the cell in order to obtain the composite for editing. Since the composite is a weighted average of the individual forecasts, the sum of the weights must equal one. For some institutions, where the prediction errors were very highly correlated between methods, we obtained pairs of weights with one value less than zero and the other greater than one. Evidently it only requires a small number of observations away from that correlation structure to cause such disproportionate weights. In calculating the average pair of composite weights for each cell, therefore, we first screened out those sets of weights not within the (0,1) range. After the two composite weighting schemes were determined for each cell, the mean square prediction errors were computed for these two forecasts as well.

For each of the five edit methods, Table 2 presents the root mean square prediction errors and composite weights for the commercial bank cells, for total transactions and total savings deposits. Similar calculations for other entities (savings and loans, etc.) and other items were also made. We anticipate the method with the smallest forecasting error to have the best potential as an edit, but until our tolerances are better tailored to the actual editing method, this potential may not be realized.

To apply the edits, we first looked for percentage changes that differed from the forecasted percentage changes by more than the appropriate tolerance (whether taken from the Tech Memo or generated as described in this paragraph), and for those ascertaining whether the dollar change tolerance was also exceeded. Since total savings and large time deposits are currently edited items, their current tolerances can be used. However, for total transactions and small time deposits, current tolerances do not exist. We therefore generated tolerances in a manner similar to that used for the creation of the current ones. This involved iterative steps with the intent of flagging approximately 0.3 percent of the observations per cell on average (the maximum percentage of observations flagged using current editing methods for other items for the year 1991). Using the components of total transactions and items that were related to small time, such as total and large time, we first compiled a range of feasible values for the tolerances. We then examined where these values occurred on the distribution of percentage changes over each cell for the two-year period. Given a reasonable proportion of the changes exceeding the initial values, we then examined the dollar change distribution for the subset of percentage change exceptions. Percentiles of this distribution were then determined in order to obtain the expected 0.3 percent edit failures under the current random walk model. These percentiles became the dollar change tolerances.

Table 2. --Root Mean Square Errors for Forecasts, Commercial Bank Cells

A. *Total Transactions*

	Root Mean Square Error					Weight of CS in Composite	
Cell	RW	CS	TS	CSRW	CSTS	CSRW	CSTS
Region 1							
-Size 4	0.077	0.073	0.077	0.074	0.071	0.72	0.51
-Size 5	0.096	0.094	0.097	0.094	0.089	0.73	0.55
-Size 6	1.190	1.190	1.276	1.190	1.204	0.70	0.58
Region 2							
-Size 4	0.064	0.059	0.236	0.060	0.121	0.77	0.58
-Size 5	0.210	0.209	0.223	0.209	0.212	0.62	0.55
-Size 6	0.331	0.330	0.344	0.330	0.333	0.68	0.57
Region 3							
-Size 4	0.102	0.099	0.108	0.100	0.100	0.75	0.51
-Size 5	0.054	0.048	0.051	0.050	0.046	0.74	0.58
-Size 6	0.067	0.063	0.071	0.064	0.062	0.70	0.60

B. *Total Savings*

	Root Mean Square Error					Weight of CS in Composite	
Cell	RW	CS	TS	CSRW	CSTS	CSRW	CSTS
Region 1							
-Size 4	0.042	0.042	0.045	0.042	0.042	0.64	0.73
-Size 5	0.054	0.054	0.056	0.054	0.054	0.64	0.67
-Size 6	0.048	0.048	0.055	0.048	0.048	0.60	0.72
Region 2							
-Size 4	0.038	0.038	0.099	0.038	0.043	0.65	0.76
-Size 5	0.235	0.234	0.244	0.234	0.236	0.64	0.64
-Size 6	0.055	0.055	0.067	0.055	0.055	0.64	0.66
Region 3							
-Size 4	0.051	0.051	0.998	0.051	0.274	0.68	0.74
-Size 5	0.041	0.040	0.041	0.040	0.040	0.63	0.66
-Size 6	0.055	0.055	0.065	0.055	0.055	0.61	0.75

Once all the forecasts and tolerances were in place, the editing experience for the 1991-1992 period was simulated for each of the five forecast methods. For each method we observed which observations were flagged as edit exceptions. Then, based on a history of weekly revisions to the EDDS file maintained by the Federal Reserve's Statistical Services branch, we were able to determine the rate of type I and type II errors for each method. (A type I error or "false positive" refers to an item that was flagged but not in error, or at least not revised. A type II error occurs when an item is not flagged but is erroneous -- as evidenced by a later revision.)

Simulation Results

For reference in this section, Table 3 shows our recommended edits based on these simulations. As mentioned in Section 1, these have been implemented as part of the Federal Reserve Board's DEEP editing software. In Table 3, the left column lists the entities (with the included size groups in parentheses), followed by the chosen edit for each item.

Details of the results on which this table is based are contained in earlier reports available on request. To give the flavor of the analysis, Table 4 summarizes the editing simulations for commercial banks' total transactions deposits; those for other entity types and other variables were similar. To assess the magnitudes and the implications of errors caught and errors missed by the editing schemes, Table 4 breaks down these errors in terms of their size (i.e., the size of the revision -- we assume, however accurately, that revised data are correct and the revision is the error in the unrevised data). It is clear from these simulations that there is room for improvement, especially regarding the type II error probabilities, which range from 98 percent to 99 percent. And although the type I error probabilities appear small, the number of flagged items that are not in error is quite large (between 87 percent and 94 percent).

Wherever the fitted time series model indicated a potentially substantial payoff relative to the random walk model (as in the first model in Table 1), the time series edit tended to be the most accurate, yielding the smallest number of edit exceptions and with fewer errors missed that were captured by other methods than vice versa. The reduction in the number of edit exceptions was not as great for the CS and CSTS composite methods, but often the composite method caused less of an increase in the type II error probability. The CS and the CSRW composite often mimicked the current RW results. Where there was doubt regarding the preferable edit method, we tended to favor the CS or CSRW -- even when the reduction in RMSE and the number of edit exceptions was small relative to the current (RW) method -- since cross section edits would allow possibly large shifts in behavior for a given week to be incorporated into the editing norm, and the DEEP software is well-suited to this type of edit. Also, we gave some preference to a uniformity of editing method across related cells (e.g., adjacent size groups within an FR region, or like size groups between regions).

For commercial banks, the alternative edits on the whole did quite well. The time series edits for total transactions and total savings were effective in reducing the total number of exceptions while missing only 3 small revisions and actually finding an additional error of over $25M. (This revision was generated either by an outside source or by an edit of another report that is not being considered here. This occurrence brings to light that some errors are detected by other sources - not the Reserve Banks or the Board. What we gain from this additional edit exception an earlier detection of the error; it would not necessarily go undetected permanently.) For the other entity types, total transactions was the only item that allowed for an alternative other than the CSRW method (CSRW was selected for

Table 3. --Experimental Edits for DEEP

Institution	Total Transactions	Total Savings	Large Time	Small Time
Commercial Banks (3,Ccd)	RW	RW	RW	RW
Commercial Banks (4,5,6)	CSTS	TS	CS	CS
Credit Unions (1,2,3,4)	TS	CSRW	CSRW	CSRW
S&Ls, Coops, Sbs (1,2,3,4)	$\underline{\mathfrak{R}I}$ $\underline{\mathfrak{R}II\text{-}IV}$ TS CSTS	CSRW	CSRW	CSRW
Agencies & Brs.(1,2,3)	CSTS	CSRW	CSRW	CSRW
Edges & Agr. (1,2)	CSRW	CSRW	CSRW	CSRW

The numbers in parentheses are the size groups, with "Ccd" denoting credit card banks. CB size groups 1 and 2 are omitted, as they are priority 1 and 2 institutions. $\mathfrak{R}$ denotes the FR Region, as in TM#16. The other entries in this table have the following explanations:

TS: The time-series model-based forecast, utilizing the institution's past percentage changes (of 1, 2, 3, 52, and 53 weeks ago).

CS: The cross-section forecast, or estimate of the average percentage change over all the institutions in the editing group or cell. Uses only the data received by the Friday after the as-of date and is calculated as the 90 percemt trimmed mean of the individual percentage changes in the cell.

CSTS: A weighted average of the TS and CS percentage-change forecasts, with statistically determined weights. When the number (n) of institutions in the group available on Friday for calculating the mean is less than 20, the weights are 1 and 0 (only the TS forecast is used).

RW: The forecast based on the "random walk" model, or the time series model giving a zero period-to-period change as the best forecast -- and is thus the implicit model underlying the current edits. This translates into a percentage-change forecast of zero.

CSRW: The forecast based on a composite of the CS and RW estimates of the percentage change, again depending on the number n of available observations in the cell. Thus:
if $n \geq 50$, use CS only;
if $20 \leq n < 50$, use weighted average of the CS and RW estimates;
if $n < 20$, use the RW estimate (zero percentage change forecast).

Table 4.--Selected Editing Simulation Results for Commercial Banks

A. *Total Transactions*

1. Random Walk (Standard Edit)

Frequency/ Percent	Not Revised	<$5M	$5M <$10M	$10M >$25M	< $25M	Total
Not flagged	557,166	9,732	791	508	168	568,365
	97.76	1.71	0.14	0.09	0.03	99.73
Flagged	1,444	75	17	12	6	1,554
	0.25	0.01	0.00	0.00	0.00	0.27
Total	558,610	9,807	808	520	174	569,919
	98.01	1.72	0.14	0.09	0.03	100.00

Pr(type I error) = Pr(Flag Item | Item not in error) = 0.26 percent
Pr(type II error) = Pr(Do not Flag Item | Item in error) = 99.0 percent
Pr(Item not in error | Item Flagged) = 92.9 percent

2. Cross Section -- Time Series Composite

Frequency/ Percent	Not Revised	<$5M	$5M <$10M	$10M >$25M	<$25M	Total
Not flagged	557,326	9,743	792	509	167	568,537
	97.78	1.71	0.14	0.09	0.03	99.76
Flagged	1,444	75	17	12	6	1,554
	0.23	0.01	0.00	0.00	0.00	0.24
Total	558,610	9,807	808	520	174	569,919
	98.01	1.72	0.14	0.09	0.03	100.00

Pr(type I error) = Pr(Flag Item | Item not in error) = 0.23 percent
Pr(type II error) = Pr(Do not Flag Item | Item in error) = 99.1 percent
Pr(Item not in error | Item Flagged) = 92.9 percent

Reduction in edit exceptions = 11.1 percent
Reduction in type I error probability = 11.5 percent
Increase in type II error probability = 0.1 percent

these entity types in place of CS in order to accommodate smaller sample sizes in the preliminary data). Those credit unions and savings institutions which would have more activity in transactions accounts than the other entity types, do exhibit cyclical patterns which the time series model was able to capture. Agencies and branches also exhibited improved editing results with the CSTS method. As mentioned, this combination of alternative strategies yielded an 11 percent reduction in both the type I error probability and the number of edit exceptions, with only a very slight increase in the type II error likelihood (about 0.1 percent).

All of these results are based on simulations using data reported to the Board by the Reserve Banks. Thus, any errors caught at an earlier stage are not reflected in these data, nor are errors undetected by Banks or Board that do not show up in the revision files. And as previously mentioned, the other factor to be monitored is the use of preliminary data in cross section estimates of the mean percentage change. Depending on where the preliminary data fall in the distribution of all percentage changes for an item, the operational results based on the CS, CSTS, or CSRW methods may differ significantly from what is expected based on the simulation results.

This investigation is still in progress, and further generalizations of the work are underway or planned. Among these are examining time series models with regression components to account for such phenomena as tax dates, calendar effects or related variables, alternative groupings of the data according to size or geographic region, modelling larger banks individually, and examining additional items or variables.

References

Federal Reserve Board (1993). "Processing Procedures for the Report of Transaction Accounts, Other Deposits and Vault Cash (FR2900)," Technical Memorandum No. 16, Publications Section, (December).

Pierce, David A. and Laura Bauer Gillis (1995). "Time Series and Cross Section Edits with Applications to Federal Reserve Deposit Reports," *Seminar on New Directions in Statistical Methodology*, Statistical Policy Working Paper No. 23, Part 1, pp.152-171 (June). ■

Chapter 12

Inflation Factors for Stratified Samples with Control Information

Peter Ochshorn, University at Albany, SUNY

Abstract

This paper looks at the "inflation factor" or "weight" for the members of a stratified sample. As the inverse of the sampling ratio, it is often interpreted as the number of members of the population "represented" by each sample member. In the first section, standard estimators are reexamined to construct corresponding "implicit" inflation factors. Then, inflation factors are chosen to minimize error functions. Two examples are discussed, both of which are easily computed by linear programs. An extract of the data for the problem will also be displayed.

Inflation Factors for Stratified Samples with Control Information

Peter Ochshorn, University at Albany, SUNY

Introduction

The quantity $w_h = N_h/n_h$ is sometimes referred to as an "inflation factor" or "weight" for the members of a stratified sample survey in stratum h. As the inverse of the sampling ratio, it is often interpreted as the number of members of the population "represented" by each sample member. By applying the factors to the sample, sums of variates are "inflated" from the sample to the population, providing standard estimators such as $\sum_{h,i} w_h y_{hi}$ for the sum of variate y over the population. When covariate parameters or "controls" such as μ_{xh} are known, it is usually possible to construct more efficient estimators, such as ratio and regression estimators. For the theory, compare any sample survey textbook, e.g., (Cochran, 1977). These in turn require a degree of statistical sophistication not always present in the user of the sample, especially if the sample is made available to "outside" users. Additionally, tables of control information are not always available to these "outside" data users. Thus, there is some motivation to adjust the inflation factors, taking account of control information in some way to force better estimation.

Inflation Factor Form of Control-Based Estimators

In this section standard estimators are reexamined to construct corresponding "implicit" inflation factors. This can be thought of as "a priori" construction because the estimators usually are chosen according to their average behavior, i.e., prior to sampling. The advantages of the different choices are discussed in the textbooks. Equation numbers below beginning with "c" refer to (Cochran, 1977) (e.g., "c6.44/p164" is Cochran's equation 6.44 found on page 164).

The Separate Ratio Estimator

The inflation factor corresponding to the separate ratio estimator can be derived starting with equation c6.44/p. 164:

$$\sum_h \frac{y_h}{x_h} X_h = \sum_h \frac{N_h}{n_h} \frac{\mu_{xh}}{m_{xh}} \sum_i y_{hi} = \sum_h w_h \frac{\mu_{xh}}{m_{xh}} \sum_i y_{hi}$$

so that the implicit inflation factors are just $w_h(\mu_{xh}/m_{xh})$, the adjustment being the ratio of population to sample mean of x in each strata. In other words the sample is inflated by the ratio of x totals rather than counts.

The Combined Ratio Estimator

To derive the inflation factor corresponding to the combined ratio estimator, one can start with equation c6.48/p165:

$$\frac{\hat{Y}_{st}}{\hat{X}_{st}}X = \frac{N\mu_x}{\sum_h N_h m_{xh}} \sum_h \frac{N_h}{n_h} \sum_i y_{hi} = \frac{N\mu_x}{\sum_h N_h m_{xh}} \sum_h w_h \sum_i y_{hi}$$

so that in this case the adjustment to w_h is constant across strata, being the ratio of population to count-inflated sample totals of x. Another view of the adjustment factor is given by:

$$\left[\sum_h \frac{N_h}{N}\left(\frac{\mu_x}{m_{xh}}\right)^{-1}\right]^{-1}$$

which is a harmonic mean of the strata ratios of population to sample mean of x, weighted by population count shares.

Separate Regression Estimator

Regression estimators can also be the starting point for control-based inflation. For the separate regression estimator the starting point can be equation c7.49/p200 (multiplied by N):

$$\sum_h N_h[m_{yh} + b_h(\mu_{xh} - m_{xh})] = \sum_h w_h \sum_i [1 + n_h(\mu_{xh} - m_{xh})a_{hi}]y_{hi}$$

where a_{hi} are the linear least squares transforms from $b_h = \sum_i a_{hi}y_{hi}$ and are equal to:

$$a_{hi} = (x_{hi} - m_{xh}) / \sum_i (x_{hi} - m_{xh})^2.$$

In this case the inflation factors are adjusted not only according to the difference (not ratio) of population and sample mean of x by stratum, but also to the distance of the covariate from its sample stratum mean. Thus each sample member receives an individual inflation factor, which varies within as well as among the strata.

Combined Regression Estimator

The final "a priori" inflation factor of this section is derived from the combined regression estimator. Starting with c7.50/p200 (multiplied by N):

$$(N_h/n_h)y_{hi}+b_c(N\mu_x-m_{xh}) = \sum_h w_h y_{hi}+b_c N_h(\mu_{xh}-m_{xh}) = \sum_h \sum_i [w_h+N_h(\mu_{xh}-m_{xh})a'_{hi}]y_{hi},$$

where $b_c = \sum_{h,i} a'_{hi}y_{hi}$ is given by equation c(mid-page)/p202 weighted combined regression. As in the separate regression estimator, the implicit inflation factor depends both on the difference between population and sample means in the strata, and on the distance between the covariate and its sample stratum mean.

Inflation Factors as Solutions to Ex-Post Optimal Programs

In this section inflation factors are chosen to minimize error functions. Instead of starting with an estimator to construct inflation factors, one can start with a comparison between control totals and inflated sample totals, minimizing some objective function of the discrepancies by choosing "optimal" inflation factors. An advantage of this approach is the flexibility in choosing the objective, allowing for various criteria to assume their respective importance in determining the solution. Two examples are discussed below, both of which are easily computed by linear programs.

Example 1

Suppose $\{z_j\}$ are a set of variables for which the population totals by strata Z_{jh} are known. Usually the unit variable whose totals are N_h and n_h for the population and sample respectively will be included in this set. Construct the ratios $w_{jh} = Z_{jh}/z_{jh}$ (again usually including $w_h = N_h/n_h$. Let $\{w_h^*\}$ be the set of inflation factors to be determined. Then for each stratum the discrepancy between sample and population for variable z_j is $w_h^* z_{jh} - Z_{jh}$. One possibility for an optimality criterion is the sum of the absolute percentage errors, over variables and strata, or

$$\sum_{j,h} |1 - w_h^*/w_{jh}|.$$

This objective is minimized by minimizing separately for each stratum $\sum_j |1 - w_h^*/w_{jh}|$, summing over the variables z_j. As a function of w_h^* this is a sum of piecewise-linear convex functions, and so inherits these traits. The optimum is therefore one among the $\{w_{hj}\}$. In many applications the solution will be close to the median of these values, but in any case never greater than the median.

To show this result, assume generically that $1 \leq w_{jh} < w_{j+1,h}$. Then the slope of the objective function between is just $w_{j'h}$ and $w_{\{j'+1,h\}}$ is just

$$\sum_{j \leq j'} -1/w_{jh} + \sum_{j > j'} +1/w_{jh}$$

which, since the individual terms $1 / w_{jh}$ are decreasing over j, must be positive whenever j' is such that the number of terms in the first sum is equal to greater than the number in the second sum. The objective minimum must therefore be less than or equal to all such j'.

In cases where, for each stratum h, the $\{w_{jh}\}_j$ are roughly of the same order of magnitude (a "nice" sampling outcome) then the median value is a likely solution (for this problem, for an even number of values, the lesser of the two middle values is taken as the median.) For example with three ratios, the slope between w_{h1} and w_{h2} is given by

$$+1/w_{h1} - 1/w_{h2} - 1/w_{h3},$$

and sufficient conditions for the negativity of this slope are given by $w_{h2} < 2w_{h1}$ and $w_{h3} < 2w_{h1}$.

Example 2

The final example is an actual application to an annual microdata file maintained by the New York State Department of Taxation and Finance, Office of Tax Policy Analysis. The data consist of approximately 90,000 records randomly sampled from a stratified population of about 8,000,000 tax filers. Stratification is by type of tax return (long form, short form, etc.) and by income class. Control information consists of strata totals of return counts, income, and tax liability.

Part of the art of constructing mathematical programs is in eliciting a hierarchy of preferences from users of the final "product." In the present instance it has been determined that it is important to reduce discrepancies in income totaled by income class, in counts totaled by return type, and in overall total tax liability. An implementation has been made using the AMPL (Fourer et al., 1983) algebraic modeling language with the MINOS (Murtaugh and Saunders, 1993) program solver. The objective function has been implemented as a weighted sum of absolute values of the discrepancies listed above. Such a minimization is inherently linear, and can be solved by the usual linear programming techniques. The appendix displays an AMPL model for this problem.

Acknowledgment

The work reported here was begun while the author was employed at the NYS Department of Taxation and Finance. The suggestion to explore the use of linear programs to inflate the personal income tax sample came from Frans Seastrand, Director of the Bureau of Revenue Analysis and Data.

References

Cochran, W. G. (1977). *Sampling Techniques* (third edition), Wiley, New York.

Fourer, R.; Gay, D. M.; and Kernighan, B. W. (1993). *AMPL: A Modeling Language for Mathematical Programming*, The Scientific Press, San Francisco.

Murtaugh, B. A. and Saunders, M. A. (1993). *MINOS 5.4 User's Guide*, Stanford University, Stanford CA.

Appendix: AMPL Model

The AMPL model below specifies the minimization of an error function as explained above. The inflation factors to be determined are coded in the statement beginning "var Infl ..."; the objective statement starts with "minimize Errors: ..."; and the constraints follow with the phrase "subject to ..." (the double construction for each constraint is a trick to convert absolute value problems into linear form)

```
set RetType ;
set IncCls ;
set Cells within {RetType,IncCls} ;

param CtrlInc {Cells} ;
param CtrlCnt {Cells} ;
param CtrlTax {Cells} ;
param SmplInc {Cells} ;
param SmplCnt {Cells} ;
param SmplTax {Cells} ;

param ClassCtrlInc {i in IncCls} := sum {(r,i) in Cells} CtrlInc[r,i] ;
param TypeCtrlCnt {r in RetType} := sum {(r,i) in Cells} CtrlCnt[r,i] ;

param TotCtrlInc := sum {(r,i) in Cells} CtrlInc[r,i] ;
param TotCtrlCnt := sum {(r,i) in Cells} CtrlCnt[r,i] ;
param TotCtrlTax := sum {(r,i) in Cells} CtrlTax[r,i] ;

param ClassIncWgt {i in IncCls} := ClassCtrlInc[i]/TotCtrlInc ;
param TypeCntWgt {r in RetType} := TypeCtrlCnt[r]/TotCtrlCnt ;

param IncWgt default 1 >0 ;
```

```
param InvCntRatio {(r,i) in Cells} := CtrlCnt[r,i]/SmplCnt[r,i] ;
param InvIncRatio {(r,i) in Cells} := (if not SmplInc[r,i]
    then InvCntRatio[r,i] else CtrlInc[r,i]/SmplInc[r,i]) ;
param InvTaxRatio {(r,i) in Cells} := (if not SmplTax[r,i]
    then InvCntRatio[r,i] else CtrlTax[r,i]/SmplTax[r,i]) ;
param MaxRatio {(r,i) in Cells} :=
    max ( InvIncRatio[r,i], InvCntRatio[r,i], InvTaxRatio[r,i] ) ;
param MinRatio {(r,i) in Cells} :=
    min ( InvIncRatio[r,i], InvCntRatio[r,i], InvTaxRatio[r,i] ) ;

var Infl{(r,i) in Cells} >=MinRatio[r, i], <=MaxRatio[r,i],
    := InvCntRatio[r,i] ;

var ClassInflSmplInc {i in IncCls} = sum {(r,i) in Cells}      Infl[r,i]*SmplInc[r,i] ;
var TypeInflSmplCnt {r in RetType} = sum {(r,i) in Cells}    Infl[r,i]*SmplCnt[r,i] ;
var TotInflSmplTax = sum {(r,i) in Cells} Infl[r,i]*SmplTax[r,i] ;

var ClassIncErr {i in IncCls} = (if ClassCtrlInc[i]=0 then 0 else
    ClassInflSmplInc[i]/ClassCtrlInc[i] - 1) ;
var TypeCntErr {r in RetType} = TypeInflSmplCnt[r]/TypeCtrlCnt[r] - 1;
var TotTaxErr = TotInflSmplTax/TotCtrlTax - 1;

var AbsClassIncErr {IncCls} ;
var AbsTypeCntErr {RetType} ;
var AbsTotTaxErr;

minimize Errors:
    IncWgt*(sum {i in IncCls} ClassIncWgt[i]*AbsClassIncErr[i]) +
    (sum {r in RetType} TypeCntWgt[r]*AbsTypeCntErr[r]) +
    AbsTotTaxErr ;

subject to IncPos {i in IncCls}: AbsClassIncErr[i] >=  ClassIncErr[i] ;
subject to IncNeg {i in IncCls}: AbsClassIncErr[i] >= -ClassIncErr[i] ;
subject to CntPos {r in RetType}: AbsTypeCntErr[r] >=  TypeCntErr[r] ;
subject to CntNeg {r in RetType}: AbsTypeCntErr[r] >= -TypeCntErr[r] ;
subject to TaxPos: AbsTotTaxErr >=  TotTaxErr;
subject to TaxNeg: AbsTotTaxErr >= -TotTaxErr;
```

■

Empirical Data Review: Objective Detection of Unusual Patterns of Data

James Kennedy, U. S. Bureau of Labor Statistics

12
Chapter

Abstract

Hard edits are often coded into automated editing routines. Illogical and inconsistent responses are flagged; occasionally "soft edits," such as extreme numbers, are flagged as well. These responses are not necessarily incorrect, but require documentation. Often individual data are within normal limits, but the pattern of responses is unusual. The current paper discusses an empirical method for determining when patterns of data fall outside of a normal range.

An observation can be represented as a vector of N variables defining a point in N-dimensional hyperspace. Similar patterns are conceived to be located near one another in this hyperspace; cluster analysis produces the means of multidimensional clusters. The "unusualness" of patterns of data can then be defined in terms of an observation's distance from cluster centers.

The method is demonstrated with data from the COMP2000 generic leveling field test. Data patterns from nine generic leveling factors were analyzed into ten clusters. A Windows program demonstrates the comparison of new data with clusters found in previously-collected data. ■

13
Chapter

Case Studies -- III

Chair: Charles Day, National Agricultural Statistics Service

Anusha Fernando Dharmasena

Clancy Barrett ◆ Francois Laflamme

Gregory D. Weyland

Chapter 13

Time-Series Editing of Quarterly Deposits Data

Anusha Fernando Dharmasena, Federal Reserve Board

Abstract

The Statistical Services Branch in the Division of Information Resources Management at the Federal Reserve Board is responsible for ensuring the accuracy and reliability of deposits data reported by the Federal Reserve Banks. This research attempts to provide a statistical methodology for editing these data using forecasting techniques, to identify "acceptable" and "unacceptable" data. The study will show that changes in the quarterly deposits data are a result of changes in seasonality, the number of respondents, and "micro level" data fluctuations.

These consistent fluctuations in the aggregates have been modeled using regression techniques. The data for this study consists of twelve quarterly deposits items that were summarized by five entity types.

Time-Series Editing of Quarterly Deposits Data

Anusha Fernando Dharmasena, Federal Reserve Board

Introduction

The Deposits Unit of the Statistical Services Branch at the Federal Reserve Board is responsible for editing and refining deposits data reported by the Federal Reserve Banks. One such project is the editing of quarterly deposits data to ensure accuracy and reliability.

Examining the Quarterly Edited Deposits System (QEDS) has became more important as we realize that the present system of data analysis at the micro level is less useful than an analysis of QEDS aggregated data. This research attempts to provide a statistical methodology for editing these data using forecasting techniques, to identify "acceptable" and "unacceptable" data. We assume that the changes in the QEDS data are caused by "micro level" data fluctuations, seasonality, and other macro influences. These consistent fluctuations indicate that the aggregates could be modeled using regression techniques. The close fit of the final model shows that this assumption is true.

The QEDS data consist of the quarterly reported deposit items which are listed below:

- Vault cash
- Total demand deposits
- ATS & NOW accounts
- Total savings
- Small time deposits
- All time deposits
- U.S. Government demand deposits
- Demand deposits due to
- Cash items in process of collection
- Demand balances due from
- Other demand deposits
- Total net transactions
- Total nonpersonal savings and time deposits.

The data have been aggregated by entity type as described below:

- Commercial banks (member banks)
- Commercial banks (nonmember banks)
- Mutual savings banks
- Savings and loans
- Credit unions.

Exploration

The research began by looking at various forms of the dependent variable: the QEDS quarterly value for a given item, together with independent variables that would result in good estimates of QEDS data for a given quarter. To this end many linear regression models were investigated to obtain the best model that fit the data and that helped in the formulation of the final model. This final model serves as the basis for our aggregate data editing procedure.

The data for the initial explorations were taken from a SAS dataset that was created for performing exploratory data analysis with QEDS using SAS/Insight®.

The first attempt was to look at the following multiple linear regression expecting meaningful estimates and useable models:

$$\tilde{Y}_t = \beta_0 + \beta_1 \tilde{Y}_{t-1} + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + \beta_5 X_{5t} + \beta_6 X_{6t} + \beta_7 X_{7t} + \epsilon_t \quad ,$$

where

$\tilde{Y}_t$ = The percentage change of the QEDS item from quarter to quarter after ensuring that the number of respondents stayed the same for each quarter. This was accomplished by dividing the QEDS value by the number of respondents for each quarter.
$\tilde{Y}_{t-1}$ = The lagged dependent variable (% change from previous quarter)
X_{2t} = Monetary Aggregates (MA) growth rates
X_{3t} = Number of banks per quarter
X_{4t} = Seasonal Dummy 1 representing quarter 1
X_{5t} = Seasonal Dummy 2 representing quarter 2
X_{6t} = Seasonal Dummy 3 representing quarter 3
X_{7t} = Seasonal Dummy 4 representing quarter 4
ϵ_t = A random error term uncorrelated over time, typically called ***white noise***.

❒ **Results.**--The coefficients for the seasonal factors were significant, indicating strong influences of seasonality. Unfortunately, because the data had to be manipulated to ensure that the same respondents were reporting for two consecutive quarters, some respondents were eliminated from the calculation. This elimination resulted in a sample not completely reflective of the QEDS universe, which, in turn, led to poor models that could not be used to predict aggregate QEDS data.

The next step in the analysis was to examine the explanatory power of a different set of independent variables -- combinations of economic factors. We hoped that these indicators would be linearly related to the deposits data:

$$\tilde{Y}_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + \epsilon_t \quad ,$$

where

$\tilde{Y}_t$ = The percentage change in QEDS -- the dependent variable and the following independent variables for end of quarter reporting dates
X_{1t} = GNP

Y_{2t} = Interest rates
Y_{3t} = Consumer price index
Y_{4t} = Unemployment rates.

❒ **Results**.--The outcome of the statistical model was disappointing. These leading indicators did not produce strong estimates as hoped, because a correlation analysis of the variables showed the lack of a functional relationship between the dependent and independent variables. Since the postulated model did not describe the data satisfactorily and no fundamental conclusions were recovered from the fitted equation, the model could not be used.

The research continued to use the data described in the first attempt with the dependent variable (Y). One of the new predictor variables (X) reflected the percentage change in the number of respondents from quarter to quarter. In addition, to incorporate an important guide to the properties of time series analysis, we introduced a series of lags of the dependent variable:

$$\tilde{Y}_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + \beta_5 Y_{t-1} + \beta_6 Y_{t-2} + \beta_7 Y_{t-3} + \beta_8 Y_{t-4} + \varepsilon_t ,$$

where

Y_t = The percentage change in QEDS -- the dependent variable
X_{1t} = MA growth rates
X_{2t} = Seasonal factors
X_{3t} = Percentage change in the number of respondents
X_4 = Number of respondents
$\beta_5\tilde{Y}_{t-1} - \beta_8\tilde{Y}_{t-4}$ = Lag dependent variables.

❒ **Results**.--Although some of the lags were very significant, indicating effects from previous time periods, the overall estimates produced by the model did not produce a good fit because of serial correlation among the residuals.

After much research and model testing, we decided to get data directly from the QEDS archival and to construct quarterly data by item and entity to fit the needs of the project. Thus, the final model that helped in predicting aggregate QEDS data is as follows:

$$¥_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + \beta_5 X_{5t} + \beta_6 X_{6t} + \varepsilon_t ,$$

where

$¥_t$ = Dollar difference in levels of QEDS data ($Y_t - Y_{t-1}$) - the dependent variable
β_0 = Intercept + seasonal dummy quarter 1 + seasonal dummy quarter 2
X_{1t} = MA QEDS estimate
X_{2t} = Number of respondents
X_{3t} = Seasonal dummy quarter 3
X_{4t} = Seasonal dummy quarter 4.

Here the dependent variable reflects the change in levels. The set of regressors for this model was also reduced to reflect seasonal dummies, the MA QEDS estimates, and the number of respondents for the quarter. The QEDS estimates were constructed from the growth rates and the panel shift data that were obtained from MA. These estimates were constructed in a manner similar to that used by MA to obtain the money supply estimates.

- **Results**.--The results from this model fully used the QEDS estimates from MA which included the panel shift data to help explain the variability in our dependent variable. The next section on methodology will cover the model and its results in detail.

The previous discussion was concerned primarily with finding the best model from a group of candidate models using the least squares method for estimation of model coefficients. Implicit in the least squares method are the assumptions that $E(\varepsilon_i) = 0$ and that the ε_i are uncorrelated with homogeneous variance σ^2. In addition, ***normality*** on the ε_i is required for the estimators to attain the property of minimum variance of the class of unbiased estimators. Thus, to address the issue of variances in the dependent variable from observation to observation, the proper estimator of ß should take the normality of ε_i into account by weighing the observations in some way that allows for the differences in the results. Sample autocorrelations coefficients were created to measure the correlation among observations at different distances apart. These autocorrelations were used to account for adjustments that take place over time. Therefore, all models were checked for autocorrelation and heteroscedasticity.

Prediction Model

The primary function of the preceding model-building exercise was to determine which regressor variables truly explained the response variable y; the QEDS reported value for a given quarter. The final model had the following regressor variables that were responsible for a significant amount of variation in the dependent variable. These variables are the seasonal dummies, the constructed MA QEDS estimates, and the number of respondents for the quarter.

The main focus of the analysis is to use econometric modeling techniques to make a good prediction of quarterly data. Although an attempt is being made to use a set of mathematical formulas and assumptions to describe this deviation, the uncertainty inherent in statistical prediction methodology will introduce errors. In an attempt to be parsimonious, this scientific methodology will try to capture the systematic behavior of the data and represent the factors that are nonsystematic and cannot be predicted as error terms.

Using the traditional linear regression equation -- the least squares method, the research will attempt to explain the relationship between the dependent variable and the regressors for forty-four quarters as follows:

$$\psi_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + \beta_5 X_{5t} + \varepsilon_t \ ,$$

where

ψ_t = The dependent variable -- the initial QEDS reported item measured at time t.

X_{1t} - X_{3t} = Three dummy variables that represent seasonality -- the four quarters of the year. We suspect that seasonality as a qualitative regressor variable will help improve prediction together with other quantitative variables.

X_{4t} = MA QEDS estimates constructed from MA growth rates and panel shift data.

X_{5t} = Number of respondents for each quarter by entity type.

The first step of the model estimation phase was to look at the results of a multiple regression equation to assess the functional relationship between the dependent and independent variables. The search for outliers in the output was the primary goal.

If the statistics indicated that the observation was both an outlier and an influential point, the observation was marked for re-estimation by the regression procedure instead of eliminating it altogether. This process ensured that the number of observations did not diminish while providing better input to the regression process. Estimating the effect of the outlier and removing the effect from the data point will eliminate its adverse influences on the final coefficient estimates.

As the estimation procedure is discussed, output for total demand deposits adjusted (2212) aggregated over member commercial banks (entity 1) will be presented for illustration purposes. Table 1 shows the initial estimation of outlier effects for this item and entity aggregates.

The listing is the first regression procedure that estimates the effects of the outliers Q1, Q7, Q19, and Q43. Each of these outliers is extremely significant, as can be noted in the T-statistic and the appropriate probability. The other independent variables that are significant in this model are the QEDS estimate from monetary affairs, the number of banks, and the seasonal effect of quarter three.

Table 1.--Model Selection and Estimation

Demand Deposits Adjusted (2212) - Commercial Member Banks (Entity 1)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	7	1.1765984E14	1.6808548E13	888.588	0.0001
Error	36	680976490487	18916013625		
C Total	43	1.1834081E14			

Root MSE	137535.49951	R-square	0.9942
Dep Mean	5024979.59091	Adj R-sq	0.9931
C.V.	2.73704		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	-248999	99782.214546	-2.495	0.0173
D3	1	-340885	54426.852218	-6.263	0.0001
F	1	461.528841	106.19117853	4.346	0.0001
QEST	1	0.909282	0.02327381	39.069	0.0001
Q1	1	2136875	157101.47043	13.602	0.0001
Q7	1	2179965	168519.89826	12.936	0.0001
Q19	1	2154042	161564.34373	13.332	0.0001
Q43	1	1519245	146780.11889	10.350	0.0001

As shown in Table 2, outliers (in the dependent variable) have potential to "pull" the regression equation in the wrong direction causing inadequate explanation of the "true" data. In addition, it will not predict future values well. This output shows that observations 1, 7, 19, and 43 have been perfectly estimated from the data and that the next regression is ready for processing once the outlier effects have been removed.

Table 2.--Estimation of Outlier Effects

	Dep Var Obs	Predict QED	Std Err Value	Predict	Residual	Std Err Residual	Student Residual	-2-1-0 1 2	Cook's D
*	1	2281560	2281560	137535.5	-568E-12	.	.		.
	2	2372881	2265323	49724.71	107558	128232.1	0.839	| |* |	0.013
	3	1370009	1868559	72212.91	-498550	117052.6	-4.259	|****** |	0.863
	4	1483171	1385466	58304.64	97705.1	124565.6	0.784	| |* |	0.017
	5	1425684	1338455	58630.46	87228.5	124412.6	0.701	| |* |	0.014
	6	1521271	1437119	58192.14	84152.3	124618.2	0.675	| |* |	0.012
*	7	3674690	3674690	137535.5	-276E-12	.	.		.
	8	3859955	3938499	36784.03	-78543.5	132525.3	-0.593	| *| |	0.003
	9	3636632	3732860	39588.95	-96228.0	131714.6	-0.731	| *| |	0.006
	17	4225384	4275400	32474.31	-50015.9	133646.7	-0.374	| | |	0.001
	18	4260961	4262018	30884.77	-1056.9	134022.9	-0.008	| | |	0.000
*	19	6376338	6376338	137535.5	-483E-12	.	.		.
	20	6739094	6782243	37362.20	-43149.2	132363.4	-0.326	| | |	0.001
	21	6349651	6351019	37091.75	-1368.4	132439.5	-0.010	| | |	0.000
	28	6357825	6314304	30277.27	43521.1	134161.5	0.324	| | |	0.001
	29	6002545	6042162	29553.99	-39617.5	134322.7	-0.295	| | |	0.001
	42	5832947	5720894	42352.31	112053	130852.2	0.856	| |* |	0.010
*	43	6967935	6967935	137535.5	-874E-12	.	.		.

The next step in the model estimation process was to rerun the multiple linear regression model with the estimated observations. Table 3 shows how the regression output was obtained.

Table 3.--Regression Model to Determine Final Estimates

Regression Output Total Demand Deposits Adjusted for Commercial Member Banks

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	3	1.3836181E14	4.6120603E13	2709.086	0.0000
Error	40	680976490487	17024412262		
C Total	43	1.3904278E14			

Root MSE	130477.63127	R-square	0.9951
Dep Mean	4843385.79272	Adj R-sq	0.9947
C.V.	2.69393		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	-248999	88373.959762	-2.818	0.0075
D3	1	-340885	46066.671130	-7.400	0.0001
F	1	461.528841	85.31971721	5.409	0.0001
QEST	1	0.909282	0.01699985	53.488	0.0001

The output lists the final results of the model estimates that were significant for total demand deposits adjusted - for Entity = 1 (commercial member banks). It is clear that this three variable regression model explains ninety nine percent of the variation in the QED's aggregate for commercial member banks apart from outliers.

This regression estimation process was carried out for each combination of entity and item. The model estimates shown in Table 4 on the following page, depicts the observations that needed to be re-estimated and the independent variables that were significant in the final model for each combination.

The next listing examines the error variance over the quarters to ensure consistency.

Heteroscedasticity -- Check for Constant Variance

OBS	D4	D3	D2	D1	ITEM	ENTITY	_TYPE_	_FREQ_	USS	N	Size of Variance
1	0	0	0	1	2212	1	0	11	40983878421	11	61039.39
2	0	0	1	0	2212	1	0	11	76861151631	11	83590.53
3	0	1	0	0	2212	1	0	11	495430154028	11	212224.19
4	1	0	0	0	2212	1	0	11	67701306407	11	78451.67

This listing depicts the size of the variance for the residuals in the column label "size of variance." It is apparent that the variances are synchronized within each quarter with slightly higher variances for the seasonal factor three. The results are consistent with the seasonal effect reflected in the regression model above where D3 or the third quarter dummy variable was significant.

The ARIMA procedure further examines the residuals from the regression model to confirm that time series elements in the dependent variable were considered in construction of the final equation.

The results of the Q statistic (Chi Square = 5.28) clearly indicate that the autocorrelation check for residuals are all highly insignificant. This is evidence that the residuals from the regression model are ***white noise*** and that the model does not suffer from violations of assumptions.

Chi-Squared Check of Residuals

Autocorrelations

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std
0	1.54767E10	1.00000	|**********|	0
1	2012229415	0.13002	|*** . |	0.150756
2	-3.0976E9	-0.20015	| . ****| . |	0.153283
3	-1.4777E9	-0.09548	| . **| . |	0.159112
4	-1.22818E9	-0.07936	| . **| . |	0.160408
5	-2.36701E9	-0.15294	| . ***| . |	0.161298
6	1606879130	0.10383	| . |** . |	0.164561

"." marks two standard errors

Autocorrelation Check for White Noise

To Lag	Chi Square	DF	Prob	Autocorrelations					
6	5.28	6	0.508	0.130	-0.200	-0.095	-0.079	-0.153	0.104

Table 4.--Model Estimates

Item	Entity 1 Outliers	Entity 1 Final Model	Entity 2 Outliers	Entity 2 Final Model	Entity 3 Outliers	Entity 3 Final Model	Entity 4 Outliers	Entity 4 Final Model	Entity 5 Outliers	Entity 5 Final Model
0080	1,7,19,43	D3, F, Qest	1,7,19,43	D3, Qest	31,37,43, 44	F, Qest	3,19,27, 43	F, Qest	1,7,19,43	D3, Qest
2212	1,7,19,43	D3, F, Qest	1,7,19,43	D3, F, Qest	27,33,35 43	D3, F, Qest	11,15,27, 43	D4, Qest	23,35,43	Qest
6917	1,7,19,43	Qest	1,7,19,43	Qest	31,35,42, 43	Qest	3,19,43	D3, Qest	1,7,19,43	Qest
2389	1,7,19,43	D3, F, Qest	1,7,19,43	D3, F, Qest	27,31,35, 43	F, Qest	1,7,19, 43	F, Qest	1,7,19,43	F, Qest
2697	1,7,19,43	D3, F, Qest	1,7,19,43	D3, F, Qest	19,31,35, 43	Qest	1,7,19, 43	F, Qest	1,7,19,43	Qest
2604	1,7,19,43	F, Qest	24	D3, F, Qest	27,31,35, 43	Qest	1,7,19, 43	D3, F, Qest	7,19,31, 43	Qest
2280	17,36,40	Qest	21	F, Qest	14	F	15,16,28	F, Qest	17	F, Qest
2698	35	D3, F, Qest	1,7,19,43	Qest	None	Qest	3	Qest	11,12,13, 15	Qest
0020	1,7,19,43,3 9	D3, F, Qest	1,7,19,43	Qest	43	F, Qest	43	F, Qest	19,27,43	Qest
0063	1,7,19,43	D3, F, Qest	1,7,19,43	D3, F, Qest	25,27,43	Qest	1,3,7,9	Qest	31,43	Qest
2340	1,7,19,43	D3, F, Qest	1,7,19,43	D3, F, Qest	33,35,43	Qest	11,15,27	Qest	23,43	Qest
2214	1,7,19,43	D3, F, Qest	1,7,19,43	D3, F, Qest	33,35,43	Qest	3,19,27, 43	Qest	1,7,19,42	Qest
6918	1,7,19,43	F, Qest	1,7,19,43	D3, F, Qest	31,39,43	Qest	1,3,11,19	Qest		Qest

D3 = Quarter 3 Seasonal Factor D4 = Quarter 4 Seasonal Factor F = Number of Banks Qest = QEDS MA Estimate

The Current and Proposed Editing Process

The current editing of quarterly deposits data do not use regression analysis -- a collection of statistical techniques that serve as a scientific basis for drawing inferences about relationships among quantities.

Data are currently analyzed at Statistical Services by using SAS/INSIGHT®, an interactive software system that provides extensive statistical capabilities. This system, employs SAS graphical features to display observations that need further investigation. The current editing technique focuses on using box plots and scatter plots, allows the analyst to visualize the data while making decisions about deviant observations. The data for analysis are constructed as follows:

- Aggregated data (especially residuals) for all reported items are analyzed using histograms and box plots to determine historical trends. The current quarter's data are also compared with historical data to determine trends in the current data. Items with abnormal aggregated values are investigated thoroughly at the micro data level.

- Micro data focuses on similar entity types, total deposit levels, and historical fluctuations for certain items. Depending on the variable, the micro data are analyzed using box plots and scatter plots to find trends and unusual data. Institutions with unusual values are referred to the Reserve Banks for verification, explanation, or revision.

This research recommends a different approach to analyze QEDS data. Firstly, it employs panel shift adjustments from quarter three to correctly reflect aggregate deposits data. This allows for a more accurate database from which inferences may be drawn. Secondly, the sample consists of the whole aggregate panel of respondents for a given quarter including the additions to the panel. Finally, in addition to the MA estimate, predictor variables such as the quarter three factor, the quarter four factor and the number of respondents being as significant as they are adds to the models ability to make better predictions. Therefore, this study has been able to develop a statistical methodology for analyzing QEDS data by fine tuning the MA estimates together with other pertinent varibles.

The proposed editing process will be implemented every quarter by compiling a dataset that has been adjusted for panel shift data, to which growth rates have been applied to get MAQEDS estimates. Then using SAS these data will be analyzed using the model estimates to flag deviant data points for the current quarter. Finally, the analyst will use the micro data to rank positive and negative percentage contributors to investigate the individual bank/banks causing the quarters prediction to be off.

Findings and Implementation

This QEDS research project has led to some interesting findings that may prove to be useful:

- The most interesting of all regressor variables was the computed QEDS estimate variable. This predictor was calculated from data reported by another group of institutions who report similar data on a weekly basis. Panel shift information and the growth rate will be applied to obtain observations that represent Monetary aggregates computed at the board. The regression was mostly explained by this variable which was highly significant. Ninety five percent of the models generated for this project included this QEDS estimate in the regression.

- Using quarterly indicator variables in the model allowed for the estimation and significance testing of seasonal effects in predicting QEDS aggregates. It was interesting to find that although QEDS aggregate estimates incorporate Panel shift effects that occur in quarter three (Q3), the significance of the Q3 indicator coefficient indicates that additional information has been obtained from predicting QEDS aggregates. In addition, panel cutoff changes become effective during quarter three -- every three years when cutoffs are reviewed.

- This methodology was also useful in that it allowed the model to take into account the outlier estimation process, which eliminated any distortion in the input data set that finally evaluated the usefulness of the predictors.

- The results from this study varied by "entity type" and item, with similar entity types sharing similar model forms. Entities 1 and 2, which are commercial member and nonmember banks, shared similar models that predicted their aggregates. Entities 3, 4, and 5 which are mutual savings, savings and loans associations, and credit unions, had comparable models explaining their aggregates.

To implement the findings of this methodology, analysts will use the predictive equations to evaluate incoming data. The evaluation process will create regions of acceptability, and any data falling outside these regions will be marked for further examination by analysts, at the micro level. The acceptable regions were computed by adding and subtracting from the predicted value, three times an estimate of the standard error of the regression model. The goal of this analysis is to further investigate the entity and item combinations that are flagged by a 1, since those observations reflect incoming data outside of prediction levels.

The table below is an example of items flagged for investigation for quarter four of 1995 and assists the analyst in focusing on the quarterly data that requires further probing.

DATE	ENTITY	ITEM	Qeds Quarterly Value	Model Estimate	MA Estimate	+/- Tolerance	Flag indicator
960325	3	0080	11822	9883	11200	1642	1
960325	5	0063	939696	821030	842245	116862	1
960325	5	6918	138172	115824	148492	22317	1

Looking at the table above, if the analyst investigates item 0080, 1939 is the difference that needs to be accounted for in this quarters data.

DATE	ENTITY	ITEM	Qeds Quarterly Value	Model Estimate	Difference
960325	3	0080	11822	9883	1939

When reported QEDs aggregates are different from the predicted value, here are some suggestions for probing the data:

- Retrieve data for the past and current quarter for the item in question. Then compute dollar and percentage differences which will be ranked to look at the highest and lowest twenty five rankings. These fifty banks will then be graphed by rank/dollar difference and rank/percent difference to observe irregular patterns. These graphs below are for item 0080 -- Vault Cash for entity 3.

Figure 1 displays the dollar difference for Vault Cash (item 0080) by rank for mutual savings banks. The graph clearly indicates that bank 11 deviates from the general pattern of behaviour of the other banks for this quarter. Therefore, this outlier causes the aggregate data to be inflated and the predicted value to be flagged. The tables below show the data.

Figure 1.--Dollar Differences Ranked for 0080

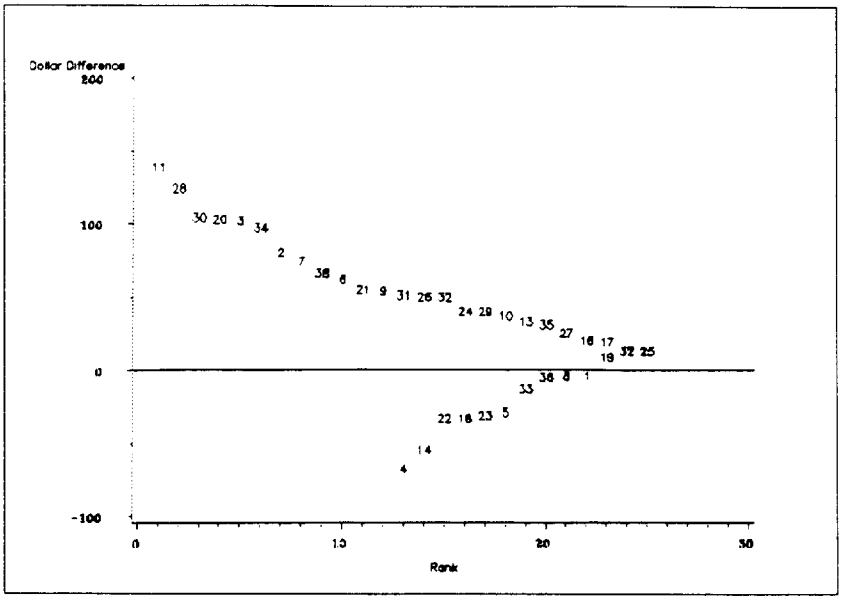
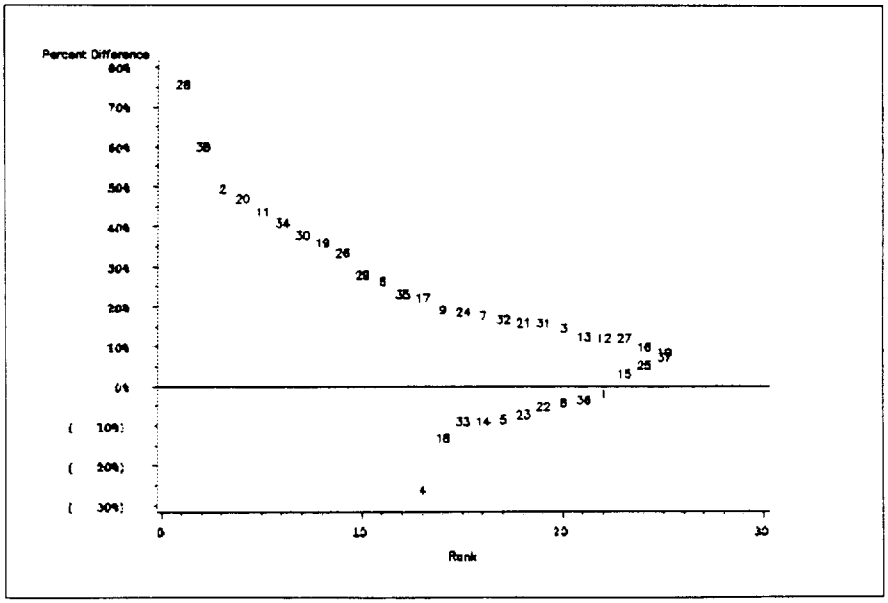


Figure 2 indicates that bank 28 and 38 are high up on the scale of positive percent changes. This indicates that the volume of activity for this bank has increased by a large amount due to structural or other changes. Bank 4 at the lower end of the negative scale should also be investigated. This is another reason for flagging the predicted value for this quarter.

Figure 2.--Percent Differences Ranked for 0080



Dollar Difference Report

Rank	QED00080	Last Quarters Value	Dollar Difference
1	454	316	138
2	288	164	124
3	378	274	104
4	322	219	103
5	803	701	102
:			
:			
:			

Percent Difference Report

17	395	426	-31
16	219	252	-33
15	601	634	-33
14	555	610	-55
13	191	259	-68

Rank	QED00080	Last Quarters Value	Percent Difference
1	288	164	76%
2	176	110	60%
3	242	162	49%

- After using the step above to locate extreme data values, the following statistics will help to identify any pattern in the data for a particular entity and item that may be reflected in the incoming data. In order to obtain an explanation of the general and uniform trend in the data that is not accounted for by an individual or a group of individuals, the following statistics will be helpful:

  - In this example, quantile statistic indicates that most of the differences calculated are positive, which means that there were more increases than decreases from the past quarter.

  - Information on the mean and standard deviation.

Statistics on Dollar Differences for 0080

N	38	100% Max	138	99%	138
Mean	32.65789	75% Q3	62	95%	124
Std Dev	49.14528	50% Med	32	90%	103
Variance	2415.258	25% Q1	-4	10%	-33
		0% Min	-68	5%	-55
		1%	-68		
Range	206				
Q3-Q1	66				
Mode	13				

Conclusion

This research has provided a statistical methodology for editing quarterly deposits data using forecasting techniques, to identify "acceptable" and "unacceptable" data. The assumptions made at the outset have proven to be very useful in building this model. There are many ways to edit "micro level" data and this research has alluded to some of these possibilities. As the model is explored there is no doubt that the prototype can be refined even further.

Bibliography

Pindyck, R. S. and Rubinfeld, D. L. (1991). *Econometric Models and Economic Forecasts*, New York: McGraw-Hill, Inc.

Myers, R. H. (1990). *Classical and Modern Regression with Applications*, California: Duxbury Press.

SAS Institute, Inc. (1991). Time Series Modeling and Forecasting, *Financial Reporting and Loan Analysis*, North Carolina: SAS Institute, Inc. ■

Experiences in Re-Engineering the Approach to Editing and Imputing Canadian Imports Data

Clancy Barrett and Francois Laflamme, Statistics Canada

13
Chapter

Abstract

The large volume of administrative data to be processed on a monthly basis and the need to use limited operational resources more efficiently led International Trade Division of Statistics Canada to the decision to base the new edit system on a combination of micro and macro approaches. The new edit system uses a series of modules that successively handled: high impact records, records in error that belong to aggregates with a high potential error, comparison of current aggregates to historical aggregates, and some special requirements. The presentation will mainly describe in a practical way the detailed mechanism for selecting and manually reviewing records. ■

Data Editing in an Automated Environment: A Practical Retrospective -- The CPS Experience

Gregory D. Weyland, U. S. Bureau of the Census

13
Chapter

Abstract

The advent of computerized data collection has opened the door to a myriad of possible scenarios for editing survey data. This includes the use of edits at the time of data capture, dependent interviewing using longitudinal data, among others, independently or in combination with detailed consistency and allocation edits after data collection is complete.

In January 1994, the Current Population Survey (CPS) became the first Bureau of the Census demographic survey to switch to a completely computerized data collection environment. This was preceded by an almost two year, large-scale test of the new data capture system and the revised processing system it required.

This discussion will review our initial plans for how the data would be edited in the new environment. Then it will review the adjustments and revisions made over the past few years to allow the editing procedures to meet the practical requirements of CPS while still improving the quality of CPS data. ■

14
Chapter

Statistical Techniques -- III

Chair: Sylvia Kay Fisher, Bureau of Labor Statistics

Thierry Delbecque ◆ Sid Laxson ◆ Nathalie Millot

Jai Choi

Richard Wendt ◆ Irene Hall ◆ Patricia Price-Green

V. Ramana Dhara ◆ Wendy E. Kaye

Statistical Analysis of Textual Information

Thierry Delbecque, Sid Laxson, and Nathalie Millot, Infoware, Inc.

Chapter 14

Abstract

Open-ended questions are a very important information source in marketing research, quality management, psychology, and survey analysis. It is critical that analysts have data mining tools that allow them to extract valuable information.

The techniques utilized to analyze complex textual information must incorporate advanced information technology in data management, linguistics, and statistics.

STATlab exploratory data analysis integrates a Natural Language Processor (NLP) module that allows users to easily analyze textual information. This module has been especially designed to meet the unique requirements for survey analysis. STATlab NLP capabilities include:

- ❒ text reduction, using lemmatization;
- ❒ filtering of the key-words;
- ❒ systematic counting of the significant terms; and
- ❒ systematic recoding and customizable recoding of the presence of terms, by creating numerical variables.

Illustrations of the use of these techniques, with classical data analysis methods such as correspondence analysis and clustering, will be demonstrated. ■

The Impact of Ratio Weighting

Jai Choi, National Center for Health Statistics

14 Chapter

Abstract

A sample is weighted to be consistent with the population it represents. The weighting procedure is an attempt to make the sample as close to certain population characteristics as possible.

A population is made up to demonstrate the impact of ratio weighting when an attempt is made to align the sample to the selected characteristics of the population. For example, as a result of multi-stage weighting, in some surveys one sample count could represent from 100 to 120,000 -- depending on the ratios being used. The ratios vary greatly due to the nature of the characteristics. The degree to which the final weighted data approximate the true population affects the size of the variance.

The Impact of Ratio Weighting

Jai Choi, National Center for Health Statistics

Introduction

Sample units are often weighted a number of times in an attempt to make the sample compatible with the characteristics of population. For instance, the National Health Interview Survey data obtained by the National Center for Health Statistics were weighted at least five times: basic weight, nonresponses, cell counts of residential areas in Nonself-representing (NSR) PSU's, alignment of the age-sex-race cells to those of population, and expansion of the two-week reference period to 13 weeks.

Such weighting may be extended to more steps to reflect other features of population and/or correct sample biases.

The following diagram shows that a random sample S is taken from the population U:

U _____> S_____>E

A sample taken by	Weighted by
Simple random	
Stratified	
Cluster	
Probability proportional	
To population size	
Equal	
Unequal	
Single stage	Basic weight: W1
Multistage stages	
With replacement	
Without replacement	
	Non-response adjust: W2
	1st ratio adjust: W3
	2nd ratio adjust: W4
	Recall adjust : W5

The sample S is then weighted five times to estimate the population. We want to have the final estimate E close to population in every aspect. But final number of estimates and cell counts are not close to those of the population after these weightings. For instance, the cells of age-sex-race table have changed after each weighting. Although we want cell estimates close to those of the population, the final results are quite different, as seen in the next section.

The weighting may reduce the difference and result in reducing variance and/or bias. It depends on each particular situation in weighting. The weighting may be repeated until the difference is minimized between the population U and final estimate E not only in number, but also in cell distributions.

In the next section, the weights are estimated five times for a sample of 12 persons. The changes of cell distribution are shown after each weighting. The impacts of weighting are discussed in the following section.

Example

A population is created for illustration, from which a sample of 12 persons is taken to show the five steps of weighting in Table 1. The population of 1,600 is divided into four strata, the first stratum of 300, the second of 300, the third stratum of 600, and the fourth stratum of 400.

Table 1.--Weighting of Doctor Visits

Str	PSU	n	Basic weight	Nonresponse		1-Ratio		2-Ratio		V	2WV	52WV
				R	W2	R1	W3	R2	W4			
(1)	(2)	(3)	(4)	(5)		(6)		(7)		(8)		
I	A300	1	100	1	100	1	100	**.8**	80	2	160	4,160
300		2	100	1	100	1	100	1	100	0	0	0
		3	100	1	100	1	100	1	100	0	0	0
II	B300	4	100	3/2	150	1	150	1	150	1	150	3,900
		5	100	n-resp								
300		6	100	3/2	150	1	150	1	150	0	0	0
III	A200	7	150	1	150	1	150	1	150	0	0	0
		8	150	1	150	1	150	1	150	3	450	11,700
600*	B200	9	150	2	300	**.9**	270	**.8**	216	1	216	5,616
		10	150	n-resp								
VI	A100	11	200	1	200	1	200	1	200	2	400	10,400
400*	B100	12	200	1	200	1	200	1	200	0	0	0
1600	**1,200**	**12**	**1,600**		**1,600**		**1,570**		**1,496**	**9**	**1,376**	**35,776**

*NSR-stratum R=ratio str=strata sp=sample
V=2 weeks Doctor visits
2WV=weighted visits for 2 weeks recall
52WV=weighted visits for 52 weeks.

The first two are self-representing PSU's, while the last four are the nonself-representing (NSR) ones. No sampling is involved in the first two PSU's got the first stage. Two PSU's are selected out of the three PSU's, each with 300 persons, in the stratum III by equal probability, and two PSU's taken from the four PSU's of equal size (100) in the stratum VI.

The weights given to these six PSU's are 1, 1, 3/2, 3/2, 4/2 and 4/2 in the respective stratum.

A sample of 12 persons is selected from the six sample PSU's by simple random sample, 3 from each of the first two PSU's, 2 from each of the third and fourth PSU's, and 1 from each of the last two PSU's.

The weights are 300/3, 300/3, 200/2, 200/2, and 100/1, 100/1 for the selection of persons from the respective sample PSU.

The basic weights are shown in the 4th column, and they are the multiplications of the two weights arising from the selection of PSU's and persons. The basic weights are 100, 100, 150, 150, 200, and 200 for the respective PSU.

The nonresponses are adjusted within the PSU, and a nonresponse ratio, used to adjust the missing numbers, is the sample persons divided by the number of respondents within the PSU. The fifth and sixth column shows the 5th and 10th samples did not responded, and adjusted accordingly.

The living areas in NSR-PSU are divided into three cells of city, urban and rural places. Six sample persons, 7, 8, 9, 10, 11 and 12, are from the NSR-PSU's in the strata III and VI, shown in Table 2; hence, they are the subject of the first stage ratio estimation. The cell ratios of population to the sample estimates are 1.0, 0.9, and 1.1, and used for the first stage ratio estimation. Since the ratio is 0.9 for the second cell, the 9th sample requires the adjustment, while no adjustment is needed for the remaining 5 sample persons in the first cell as seen in column 6, Table 1, for the lst Ratio estimation.

Table 2.--1st Ratio in NSR-PSU's

Cell	1 City	2 Urban	3 Rural
Population	510	90	100
Estimation	510	100	90
1st ratio (sample no.)	1(7,8,10,11,12)	0.9(9)	1.1(0)

Each of the 12 persons belongs to one of the 8 age-sex-race cells. Table 3 consists of the three tables of eight cells for population, estimation, and the ratios. The ratios in the last table are population divided by estimates, and shown in column 7 in Table 1. They are used for the second ratio estimation.

Table 3.--2nd Ratio of Population, Estimation					
Cell	Age	Male-White	Male-Black	Female-White	Female-Black
Population	1-49 yr	350	40	350	60
	50+ yr	350	60	350	40
Estimation	1-49 yr	350	50	350	50
	50+ yr	350	50	350	50
Ratio	1-49 yr	1(5,8,10)	0.8(1)	1(2,3,7)	1.2(-)
	50+yr	1(4,11)	1.2(-)	1(6,12)	0.8(9)

Table 4 shows the weighting process of one doctor visit of the 9th sample person. The weight of this person is changed five times, W1 through W5.

The W1 is 150, the sample persons selected from two stages, the first stage of selection of two PSU's out of three PSU's of equal size. Two persons were sampled from each sampled PSU. The basic weight is the product of these two weights, i.e., 150 = (3/2) x (200/2).

The W2 is the number adjusted for the nonresponse. Since one of the two sampled persons in the same PSU did not respond, the weight of the respondent is doubled (300 = 2 x 150) to cover the nonrespondent.

Table 4.--The Changes of Weight for One Visit of the 9th Sample Person				
W1(basic)	W2(n-resp)	W3(1st ratio)	W4(2nd ratio)	W5(52 wks)
150	300	270	216	5,616
1/(p1 P2)	W1 x 2/1	W2 x 0.9*	W3 x 0.8	W4 x 26

The W3 is 270 from the first ratio weighting (270 = 0.9 x 300). As this sample person lives in urban area, her first stage ratio is 0.9 as shown in Table 2.

The W4 is 216 by the second ratio weighting (216 = 270 x 0.8). Since this sample belongs to the cell (2,4), the black female of 50+ years, her second ratio is 0.8 for her age-sex-race. She represents 216 people for her stratum, PSU, residential area, and age-sex-race class.

The W5 are the estimated number of visits for 52 weeks or one year. She visited the doctor's office once during the past two weeks, and her one visit became 5,616 visits (= 26 x 216) for 52 weeks as shown in the column 8.

Each sample in Table 1 is weighted the same way. The 9 visits from 12 sample persons became 35,776 visits after the sample visits were weighted five times.

The eight cells in age-sex-race table are the basic weights, W1 and they have been changed three times through the three weighting processes as seen in Table 5. After these weightings, the last row of W4 is quite different from that of of the population. This difference is mainly due to sampling, non-responses of the samples 5 and 10, and the first ratio adjust of the sample 9, and the second ratio adjust of the samples 1 and 9.

Similarly, the 3 residential cells of population in NSR-PSU's differ from those of the last estimates in Table 6. This difference is also due to the sampling, empty cell, and first and second ratio adjustments of the ninth sample.

Table 5.--Eight Cells of the Age-Sex-Race Table

Cell	1	2	3	4	5	6	7	8
Pop	350	40	350	60	350	60	350	40
W1	400	100	350	0	250	0	300	150
W2	300	100	350	0	300	0	350	150
W3	270	100	350	0	300	0	350	150
W4	216	80	350	0	150	0	350	150

Table 6.--Three Cells of Residential Areas

Cell	1 City	2 Urban	3 Rural
Population	300	90	100
Estimate given	300	100	90
W1	300(8,10)	150(9)	0(-)
W2	300(8,10)	150(9)	0(-)
W3	270(8,10)	150(9)	0(-)
W4	216(8,10)	150(9)	0(-)

Remarks in Weighting

In the process of ratio weighting, we observed that each step of weighting may reduce or increase the differences between the estimates and population. This may also increase relative bias and/or variance, depending on the specific situations in sampling and weighting. Each step may have contributed to the estimation, as discussed on the following page.

❒ **The Basic Weight (W1)**

There are many ways to select a sample. For instance, if population were structured in three stages, and the sample taken by pps design, the variance would be minimized. Thus, the basic weight is decided by sampling design.

If a sample is randomly selected, the basic weighting may reduce the relative variance, while a non-random design might increase the variance and bias dramatically.

There may be empty cells when the sample persons are distributed over the cells in a large table. This may happen more likely for a small sample.

❒ **Nonresponse Adjusted Weight (W2)**

Nonresponse may be adjusted at a proper stage or stages. If nonresponses arise randomly and the nonresponse rate is low, the ratio adjustment may be valid especially for a large sample, and bias and/or variance reduced.

On the other hand, if the nonresponse rate is more than 30 percent, the ratio estimate may cause severe biases even for a large sample.

Alternative methods may be used in order to reduce bias in the presence of high rates of nonresponse. Other methods such as regression and Bayesian methods are often useful for nonresponse estimation. But such methods usually bring problems later at the stage of data analysis.

❒ **The First Stage Ratio Adjusted Weight (W3)**

We often do not have enough sample persons in sparsely populated area or among specific sub-populations, such as African Americans or older people. Consequently, the small number of a sample may not reflect the characters of population. Thus, we may use the ratio between a population and its estimation.

If the previous weights were already biased, this process may further increase biases.

❒ **The Second Stage Ratio Adjusted Weight (W4)**

The weights from the previous adjustment may not reflect the age-sex-race cells of the population. We may multiply the ratio, population to its estimate, to the previous weights. This is done for each person in the age-sex-race cell. But the resulting cells may differ from those of the population due to the empty cells, small sample, and the previous weighting. Although this process reduces the difference between the population and estimate in the age-sex-race cells, it may make the difference greater for the cells of living areas, which one may like to avoid.

❒ **The Recall Adjusted Weight (W5)**

The number of doctors' visits in the past two weeks is only 1/26 of one year; hence, we multiply 26 to make the previous weights to be the visits for one year.

The resulting number of visits per year may mislead readers for a calendar year, as two weeks could be extended to the future or past 52 weeks from the point of an interview. In this case, the visits may be counted to a different year, depending on the date of an interview.

If the nonresponse was already biased, the recalls adjust may further increase the bias.

Comments

The ratio method does not create new estimates for empty cells in a age-sex-race table. Unless we use the estimates for empty cells, no improvement can be made. However, we may put one in an empty cell for estimation or we may incease sample size if it is possible.

The high rates of nonresponses may cause bigger biases, especially when the units in the PSU are different.

The ratios may be unstable for a small sample. Since a small sample may leave more empty cells, large biases may be introduced, and nonresponse may cause more problems.

The order of weighting also has influence on the final outcome of a table. If the order of weighting were changed in the previous example, or the age-sex-race adjusted first and then residential area in NSR-PSU's, the result would be quite different. One may do the most important adjustment at the last stage.

The above example illustrates the difficulty to estimate population by ratio weighting to satisfy all of its aspects. In order to reduce the difference between the population and estimate in the age-sex-race as well as in residential areas, we may repeat steps from the first ratio W3 to the second ratio W4, leaving W1, W2, and W5 out, and stop when the difference between population and final estimates is minimum for both tables. Each time a new ratio table is created from the ratios between the population and new estimate of W3 or W4.

The ratio estimation may work better if no cells were empty, response rates high, sample size reasonably large, and cell members homogeneous. ■

Fitting Square Text Into Round Computer Holes -- An Approach to Standardizing Textual Responses Using Computer-Assisted Data Entry

Richard D. Wendt, Irene Hall, Patricia Price-Green, V. Ramana Dhara, and Wendy E. Kaye, Agency for Toxic Substances and Disease Registry

14
Chapter

Abstract

Textual responses in data collection efforts present major programming and analysis challenges. One type of text response that poses a problem in the data collection efforts of the Agency for Toxic Substances and Disease Registry's (ATSDR) Hazardous Substances Emergency Events Surveillance is chemical names. Many different chemicals are used and released into the environment in the United States each year. Additionally, varied chemical names, trade names, and mixtures add to the difficulty of establishing a uniform naming convention. Existing naming conventions or coding schemes (for example, Chemical Abstract Service Registry Numbers) are often too complicated for use by data entry personnel. Additionally, for many chemicals no codes are available. Analysis efforts involve not only the identification but also the classification of chemicals released during emergency events. Standardizing the names of chemicals is necessary to automate the analysis process.

To solve these problems, ATSDR has created a semiautomated chemical selection system that combines chemical names previously supplied by the users and chemical names supplied by ATSDR into a single database. This data entry system incorporates chemical names and chemical category assignments from previous data collection years. Users scroll through a window containing the list of chemical names and select the substance of interest. When a user selects a chemical name, the computer stores the associated chemical in the appropriate data field. A search function allows the user to locate chemicals by typing the first few letters of the desired name.

This feedback system minimizes the use of different names for the same chemical by basing chemical name selections on chemical and substance names in previous event reports and names supplied by ATSDR. This increases chemical name standardization. Since users are more likely to select names from the menu, this method reduces the workload of ATSDR staff and increases the consistency of chemical categorization.

Fitting Square Text Into Round Computer Holes -- An Approach to Standardizing Textual Responses Using Computer-Assisted Data Entry

Richard D. Wendt, Irene Hall, Patricia Price-Green, V. Ramana Dhara, and Wendy E. Kaye, Agency for Toxic Substances and Disease Registry

Introduction

Since 1990, the Agency for Toxic Substances and Disease Registry (ATSDR) has maintained the Hazardous Substances Emergency Events Surveillance (HSEES) system. This epidemiologic surveillance system currently tracks hazardous substance releases in 14 states. HSEES allows health officials to evaluate both the nature and extent of hazardous releases (both threatened and actual) and their effects on public health. The HSEES system is an active surveillance system. Participating state health departments use a variety of reporting sources (for example, individuals, state environmental protection agencies, newspapers, police, fire departments, and hospitals) to collect HSEES information on a data collection form. Information from the data collection form is entered into a computerized data entry system that is a simulation of the paper data collection form. Participating state health departments transmit this information to ATSDR quarterly.

Although most data are categorical in nature and easy to code uniformly, there are some textual responses that require special treatment, including descriptions of the type of industry, responses indicating "other," and chemical names. In this regard, standardizing chemical names presents the greatest challenge.

HSEES defines hazardous substances emergency events as uncontrolled or illegal releases or threatened releases of substances or their hazardous by-products. From 1990 through 1992, reportable substances included the 200 chemicals identified as most hazardous at Superfund sites. Also included were insecticides, herbicides, chlorine, hydrochloric acid, sodium hydroxide, nitric acid, phosphoric acid, acrylic acid, and hydrofluoric acid (*Federal Register*, 1988). Since 1993, all hazardous substances (except petroleum products) have been included in the HSEES definition.

Events are reported if the substance(s) must be removed, cleaned up, or neutralized according to Federal, state, or local law. Additionally, a potential release is reported if it involves one of the designated substances and if it results in an action (for example, an evacuation) to protect public health (Hall et al., 1994). Presently, ATSDR maintains a database of over 11,000 hazardous substance spills and over 13,000 chemical data records.

With so many chemical names recorded in one database, the problems associated with standardizing chemical or substance names are very large. In addition, varied chemical names, trade names, and mixtures add to the difficulty of establishing a uniform naming convention. Existing naming conventions or coding schemes such as Chemical Abstract Service (CAS) Registry numbers, Department of Transportation (DOT) numbers, Chemical Hazards Risk Information System (CHRIS), or United Nations (UN) numbers are often too complicated for use by data entry personnel. Additionally, for many chemicals no codes exist. As an

example, trichloroethane can be listed four ways depending on the reporting source; as TCA, as 1,1,1 trichloroethane, as 1,1,1-trichloroethane, or as trichloroethane. All of these responses are correct names, but the text fields represent completely different answers to a computerized statistical analysis system. To the nonchemist, the choices between proper naming conventions is overwhelming.

Data Entry Method

Originally, the HSEES staff addressed this problem by creating a chemical pick-list that consisted of the 36 most reported chemical names and 2 "Other" fields (one for pesticides and another for all remaining chemicals). While this approach reduced the number of user-defined names, it was not completely successful. The problem of unique chemical naming conventions still remained a major data analysis concern.

This problem hampered efforts to analyze events by chemical substance and persisted for three reasons. First, almost one- third of the spills that were reported to the HSEES system were chemical mixtures. Second, many spills consisted of pesticides and herbicides (which may be made from complex mixtures and compounds). Third, and most significantly, the system itself was being used by ATSDR and the state health departments for two different purposes.

ATSDR staff must classify chemical names and substance names into a standardized format for analysis as part of their data processing procedures. This assists them in disseminating the public health consequences of the release events. The data entry personnel at the state health departments use chemical names as descriptions of events for other state agencies, such as emergency responders. For emergency responders, there are major differences between a pure ammonia release and a 1-percent ammonia solution release. Both spills involve the same chemical, but the level of protective measures used, the issuance of evacuation orders, the use of in-place sheltering, and the level and extent of clean-up are very different.

To address these problems, ATSDR has created a semiautomated chemical selection system that combines chemical names previously supplied by the users and chemical names supplied by ATSDR. The system incorporates data from previous years into a database file that contains a set of selected chemical names.

When users reach the chemical information data entry screen, the chemical name database file is opened. The chemical and substance names in this database are then displayed in a browse screen format. Users scroll, page, or search through the window containing the list of chemical names and select the substance of interest. At this point the computer program enters the chemical name in the appropriate data entry field. A search function allows the user to locate chemical records by typing the first few letters of the desired name.

Presently, most chemicals names in the list are selected by ATSDR staff for correct syntax, but the variety of these chemicals and substances are based on all previously reported releases. Data entry personnel are not allowed to modify these predetermined chemical names, but may still select "Other" and edit that name. As an added incentive for selecting predefined names, all CAS, DOT, CHRIS, and UN chemical codes are automatically entered and are saved to the chemical name database file. These codes are then automatically retrieved each time the user selects a predetermined chemical name.

Conclusion

This pseudo-feedback system should reduce the addition of different names for the same chemical because many selections are based on previous event reports and chemical names supplied by ATSDR. It should also increase chemical name standardization. Finally, because users are more likely to select names from the menu, this method should reduce the workload of ATSDR staff and increase consistency of chemical categorization.

The full potential of this approach should be seen when it is used as a full-fledged feedback system. By incorporating all user defined chemical names into the system and translating this feedback into standardized chemical names during quarterly data processing at ATSDR, the addition of new chemical names by data entry personnel can be reduced to a minimum. While this approach will require an intense amount of programming at first, as time passes the maintenance effort for the translation program will be greatly reduced. As seen from examination of previous data submissions, many hazardous substance releases are repetitive. The problem has been that the chemical name choices do not adequately reflect the idiosyncratic naming habits of each user By tailoring the data entry system to each user's response, the use of "Other" as a free-form text input selection should be reduced.

The main source of nonstandard chemical names will then come from either truly unique hazardous releases or data input from new states as they are added to the surveillance system. HSEES staff are currently evaluating the need for implementing a fully functioning feedback system and will decide on its development once sufficient data has been collected.

References

Federal Register (1988). Hazardous Substances Priority List, 53, 41280-41285.

Hall, I. H., Dhara, R. V., Kaye, W. E., and Price-Green, P. (1994). Surveillance of Emergency Events Involving Hazardous Substances -- United States, 1990-1992. *Morbidity and Mortality Weekly Review*, 43 (suppl S-2), 1-6. ■

Chapter 15

Edit Authoring Techniques

Chair: Jim O'Reilly, Research Triangle Institute

Shirley Dolan

J. Tebbel ◆ T. Rawson

Robert F. Teitel

Methods of Reusing Edit Specifications Across Collection and Capture Modes and Systems

Shirley Dolan, Statistics Canada

15 Chapter

Abstract

As multi-mode approaches become a viable means of collecting and capturing data, the issue of reusing edit specifications across modes and software systems becomes an important issue. Where more than one system has to be used in order to offer the respondent a choice of reporting methods, statistical agencies are faced with the problem of having to develop and maintain more than one set of edit specifications, that is, one for each mode or system used.

This presentation explores some of the methods currently being used and others which are being discussed at Statistics Canada, as solutions to the problem of reusable edit code. Among the ideas are:

- multi-mode options offered by DC2, Statistics Canada's primary collection and capture software, which includes reusable edits, and
- the potential of using DC2's editing engine with other systems used at Statistics Canada to collect data using laptops and electronic questionnaires.

Exploiting the increasingly automated traditional collection and capture modes as well as the emerging methods either singly or in combination promises to deliver savings in resources and improve timeliness and data quality. This promise is unfortunately offset by the increased difficulties in developing and maintaining different versions of the edit specifications, when more than one system is used within a survey or where there is a requirement to apply edits at different stages of the process. This paper first explores the various collection and capture methods available today followed by a discussion of typical automated editing procedures. Finally, some existing and potential solutions to the problem of maintaining multiple versions of the edit specification are presented.

Methods of Reusing Edit Specifications Across Collection and Capture Modes and Systems

Shirley Dolan, Statistics Canada

Multi-Mode Collection and Capture

In the mid-1980's, multi-mode (or mixed-mode) collection and capture was typified by "surveys which combine the use of telephone, mail, and/or face to face interview procedures to collect data for a single survey project" (Dillman and Tarnai, 1988). This is still the most popular view, but our approaches have become much more automated. Telephone surveys are frequently done with the assistance of a computer. Personal interviews can now be conducted with the aid of laptops. And newer mechanisms for processing paper questionnaires, such as Intelligent Character Recognition (ICR) technology, promise to improve the timeliness and quality of data collected by mail. In addition, other modes are beginning to receive increased attention.

Perhaps the most interesting of these is the idea of the respondent extracting data from their MIS and sending it in electronic format to the statistical agency. In fact, this method is not new and has been used at Statistics Canada since at least the early 1970's. As larger respondents (such as provincial governments) invested in computer automation, it became feasible (and desirable) to receive data on a tape destined for the mainframe. There were a number of problems associated with this method. It took considerable time to negotiate a file format and content with each reporter. Despite the best efforts to standardize, it was often impossible or not cost effective for the respondent to conform exactly to the proposed formats and code sets. This added to the development and maintenance burden at the Bureau when custom programs had to be built to read and process the differing formats. Processes tended to be batch-oriented with data being stored on flat files and this added to the effort and time needed to prepare the data tapes. However, despite these constraints, electronic reports did not fall out of favour and new technologies such as improved programming languages, database management systems and advanced communication methods have significantly improved the potential usefulness of this mode of reporting.

Another mode which has been used at several statistical agencies is the electronic questionnaire. Although there are many variations on this theme, this approach generally consists of the development (and maintenance) of an interactive questionnaire with on-line help and edits which is copied to diskette and mailed to the respondent. Data is entered using the software and then returned via the diskette. There are many advantages to this method. The paper questionnaire is eliminated, edits are performed in the presence of the respondent, and the data arrives at the statistical agency already in electronic format and, at least to some degree, edited. In some cases, incentives to use this method of reporting are built into the software, such as data manipulation and reporting features which are attractive to the respondent. The disadvantages to be considered are the development and maintenance costs,

which can be significant particularly when added-value features are built in, the additional processing requirements (diskette generation, reception and archiving of diskettes, decryption and virus scanning to name a few) as well as potential liabilities resulting from respondent expectations for the extra features included in the software.

A variant of the electronic questionnaire is the notion of providing a questionnaire to the respondent over the Internet. This can be done in any number of ways. For example, it is possible to develop a questionnaire using Hypertext Markup Language (HTML). This type of questionnaire or form is accessed by the respondent using the World Wide Web services. The data is sent to the statistical agency using E-mail. There is at least one commercial product called Decisive Survey, from Technology Corp. (Chrisholm, 1995) which also uses the E-mail/Internet approach. This product boasts a drag and drop method to questionnaire development, runs on Windows 3.1 or Windows 95 and interfaces with several E-mail systems.

Although there are many other methods of collecting and capturing data such as pen-based computers and touch-tone technology, this analysis will focus on those mentioned above.

Data Editing

Data editing is a vast topic which can include not only the basic editing which takes place during collection and capture, but also during other stages of the survey process such as inter-case editing for imputation purposes or editing to detect outliers. Editing can also be described as either manual or automatic. For the purposes of this discussion, automated editing is assumed; that is editing which takes place either interactively or in batch, in the context of a computer program or system. Also, editing refers to the checking of data applied during the collection and capture phase. These are typically:

- ***Preliminary Edits*** -- usually used to detect gross reporting or keying errors at the field level. Validation of format types, range checks and verification of simple code sets are commonly included in this category.

- ***Consistency Edits*** -- inter-field value checking, may include computations to ascertain conformity.

- ***Historical Edits*** -- a variant of the consistency edit, where the values for the most recent report are compared to those of previous reports. This typically involves comparison of gross differences against tolerance tables. For example, a warning may be issued if the value reported for number of employees this month differs from what was reported for last month by more than 10 percent.

The requirements for data editing can be different depending on the mode and stage of the collection and capture process. Consider the following illustrations.

First, editing which is applied during a personal interview using a laptop may not include complex consistency or historical checks. These may be applied at a central site where reference files and previously reported data is stored and where more powerful processing equipment is available. It is not

unusual to re-apply the preliminary edits done in the field following the application of the more complex and subjective edits to ensure that these have not compromised the basic edit rules. In this scenario, separate versions of the preliminary edits are required for the laptop application and the centralized editing process.

Second, offering respondents a choice of reporting modes increases the potential for the need to develop separate versions of the edits. A large economic survey could, for example, have respondents who report by paper questionnaire, by telephone and by electronic questionnaire. The paper questionnaires are processed using a system designed to accommodate rapid data entry. A second system, offering Computer-Assisted Telephone Interviewing (CATI) features such as call scheduling, call outcome coding and calling protocols is used to collect data from respondents preferring to report by telephone. The applications built using these two systems would have their own separate versions of the edit specifications. A third version of the edits would be included in the electronic questionnaire.

A third scenario is one in which respondents transmit data extracted from their database. The edits applied to this type of response should be minimal but it may be necessary to apply consistency or historical edits. Typically, electronic data is edited (in a batch process) with edit exceptions being addressed interactively. There is the potential to require different versions of the edits for the batch stage and the interactive stage.

The preceding examples illustrate scenarios in which the potential exists for different versions of the editing specifications to be developed. Edit specifications are an important part of any collection and capture process and their development and maintenance consume a significant percentage of the resources expended in building and testing an application. This is of course multiplied when it must be done for each mode or stage of editing used. Further complexity is added when the systems used differ in their support for specifying an edit. It may be that an edit written in one language or syntax cannot be represented to the same extent in the other system used. DC2, Statistics Canada's generalized collection and capture system, offers a attractive solution to these problems.

Reusing Edit Specifications in DC2

DC2 provides support for mixed mode processing, that is, data reported by questionnaire, by telephone and in electronic format. With DC2, edits may be specified once and reused within a survey to validate data, regardless of the method used to report the data. And there is considerable flexibility in the application of an edit. Although the same piece of code will be called to edit a value (or set of values), there are choices which can be made in how the edit is applied and actioned. In other words, the edit behaviour can be tailored to the individual needs of the particular mode being used without the need to have a separate version of the edits for each variation. The following example will illustrate this.

Data reported on questionnaires can be captured in at least two ways:

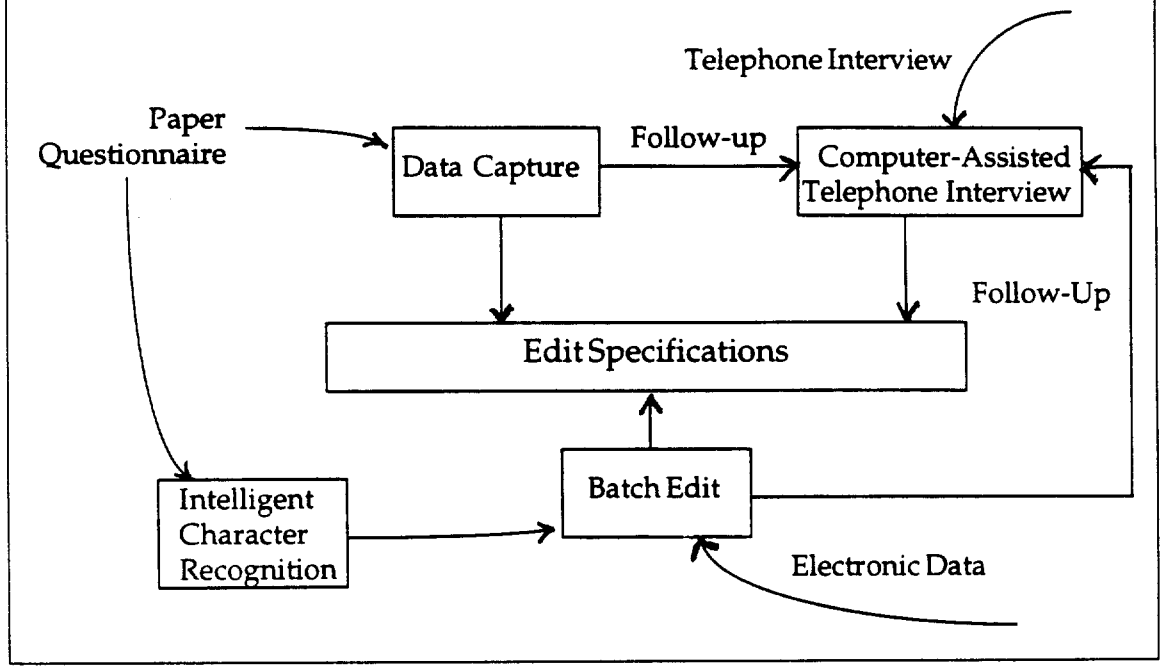
- ***Heads-Down Keying*** -- experienced key operators capture the data from the questionnaires. The term heads-down refers to the practice of keeping the eyes on the questionnaire while keying the data. The emphasis is on high key-stroke rates with low keying errors. Consequently, the capture operation is normally only interrupted when a potential keying error (identified by an edit) is detected. Reporting errors are typically corrected at a second stage by editors who are familiar with the subject matter.

- ***Intelligent Character Recognition*** -- the questionnaire data and image are captured by machine. Basic format edits are applied to detect interpretation errors. As with the operator-captured data, the subject matter corrections are normally done by experienced editors.

With DC2, the suite of edits is applied interactively and the results (passed/failed) are stored as the questionnaires are key-entered. ***However, the key operator is advised only of those errors which indicate a keying error***. The machine-captured data is loaded into the DC2 database and edited in batch, again using the same suite of edits applied during the heads-down keying exercise. Correction of errors identified from either mode would typically be done through a computer-assisted telephone interview. The previously identified edit failures would be listed for the interviewer, corrections would be made as indicated by the respondent and the edits would be reapplied, interactively, as changes were made to the data. ***At this stage, the interviewer is advised of all edit violations***. So, although the same edits are used, the behaviour surrounding their application is tailored to the mode at hand. It is even possible to build a generalized edit which uses different reference files (or some other variation) depending on the collection/capture mode.

This example, depicted in Figure 1, illustrates the flexibility of the editing facilities in combination with the desirable feature of being able to reuse the edits across modes and various stages of the collection and capture process.

Figure 1.--Reusable Edits Across Modes and Processing Steps



Many variations on the above theme are possible:

- A survey in which initial collection is shared by mail-out/mail-back and CATI. CATI is used for follow up.

- A survey in which some respondents report on an electronic questionnaire and some by mail. The electronic questionnaire data is treated as an electronic source. It is loaded into DC2 and edited in batch.

- A survey in which personal interviews are conducted mainly using laptops with some exceptional personal interviews done with paper-and-pencil method. As in the previous example, the output from the Computer-Assisted Personal Interview (CAPI) is edited in batch.

In the last two scenarios, DC2 can improve consistency of editing approach between the paper questionnaire and the electronic report by having the standard set of questionnaire edits applied to the electronic source. This would identify inconsistencies in the two sets of edits (the field edits and those applied using DC2) and would provide a centralized repository for any complex edits not included in the field edits.

DC2's mixed-mode support has, to a large degree, met the challenge of reusing edit specifications across collection and capture modes. This is certainly true for the more popular modes (paper questionnaire and CATI). The ability to process electronic data extends the reusability of the edits to other modes, if not at the point of capture, at least at the back-end as a centralized repository, editing and follow-up facility. However, the potential exists to extend the mechanism. It is conceivable that the editing facilities could be extracted from DC2 and made available to other external collection and capture systems. Before exploring this possibility, it is worthwhile looking at the DC2 editing environment.

The DC2 Editing Environment

The DC2 editing environment (Statistics Canada, 1992), shown in Figure 2, includes the following three main components:

- A specification language
- A compiler
- A run-time engine.

The ***specification language***, known locally as the Edit Specification Language (ESL), is based on the Prolog programming language. Some extensions have been made to accommodate the special case of editing statistical data. For the following reasons, Prolog is an ideal paradigm on which to base an editing language:

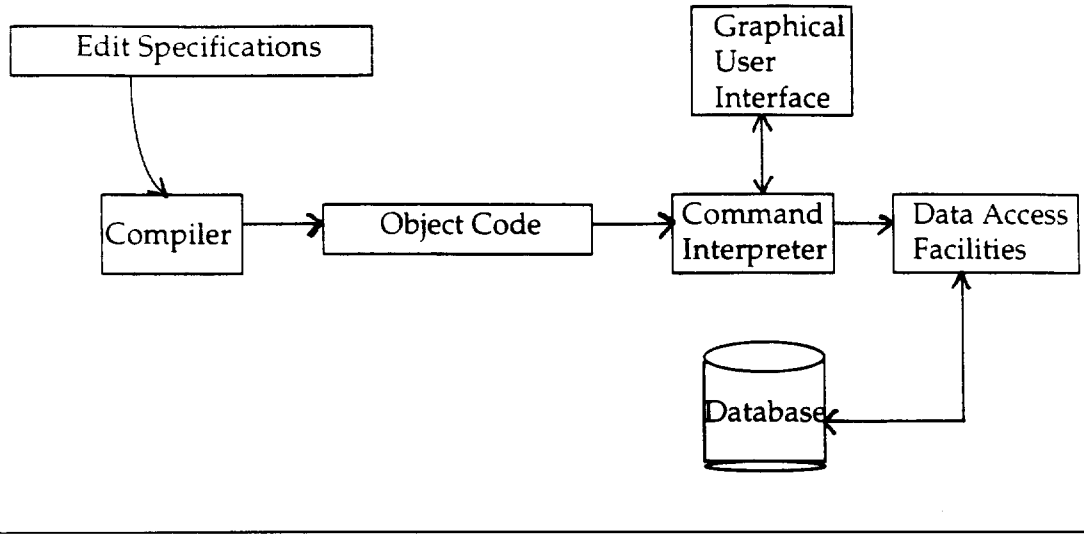
- It is a rule-based language with built-in pattern matching and backtracking.

- Prolog's concept of failure matches well with the concept of edit failure and exception handing.

- With a small set of extensions, numeric edits can be easily expressed.

- Prolog code can be compiled using a technology known as the Warren Abstract Machine (Ait-Kaci, 1991) that is in the public domain.

Figure 2.--The DC2 Editing Environment



The ***compiler*** is a Prolog program, developed at Statistics Canada, which is currently running under ALS Prolog from Applied Logic Systems. The compiler reads the ESL source code and generates code which can be executed by the run-time engine.

The ***run-time engine***, also called the command interpreter, is an in-house program which is a close approximation of the Warren Abstract Machine. It functions as a virtual computer that reads, interprets and acts on the instructions in the form of assembled (and compiled) ESL. The program is written in ANSI C and is, therefore, potentially portable to any computing platform. Within the DC2 system, it operates in two ways:

- As part of the overall production engine: receiving data via the capture instrument or the database, applying the specified rules, and reporting the results back to the calling program which subsequently stores them (and the data) in the database and/or displays them on the screen, and

- As part of a testing/debugging tool: receiving data from a programmer via the command line, applying the specified rules, and reporting the results back to the programmer.

The ESL will, of course, execute anywhere the command interpreter is available. So it becomes a matter of making it available where needed. Could this editing facility be incorporated into other collection and capture systems? This idea is visited next.

Using the DC2 Editing Engine in Other Systems

Naturally, the question of whether the DC2 editing facility can be incorporated into other collection and capture systems depends on the potential of the editing engine to interface with these systems plus the ability of these systems to incorporate a foreign editing mechanism. For this discussion, we will address systems running on laptops and those used to develop electronic questionnaires.

Could other systems dynamically access the command interpreter at run-time? The system in question would have to be designed to allow user exits or some similar mechanism to interface with external code. Most systems today allow some type of user exit -- this is commonly an interface to one or more established programming languages. Two scenarios for integrating the DC2 editing environment into a commercial product come to mind.

In the first instance, the system could communicate directly with the editing engine. This would require having the third party product extended so that it could interface with the DC2 command interpreter. This might not be attractive to the supplier of the product given that the applicability of the mechanism would be limited and of interest only to Statistics Canada. A second and more attractive approach would be to have the third party product communicate with the command interpreter through a non-proprietary interface such as the C programming language or Visual Basic. Many software products offer user exits to C or some other equally suitable programming language. It would work something like this (see Figure 3).

- A collection and capture application form is developed for a laptop or electronic questionnaire using a commercial product.

- At certain points during the capture exercise (when a field value is entered or values for a collection of fields are available), the application issues a user call to an external routine.

- The external routine accepts the data value or values and passes them to the DC2 editing engine which applies the ESL code for the field or fields in question.

- The data is validated and the results are returned to the application via the external routine.

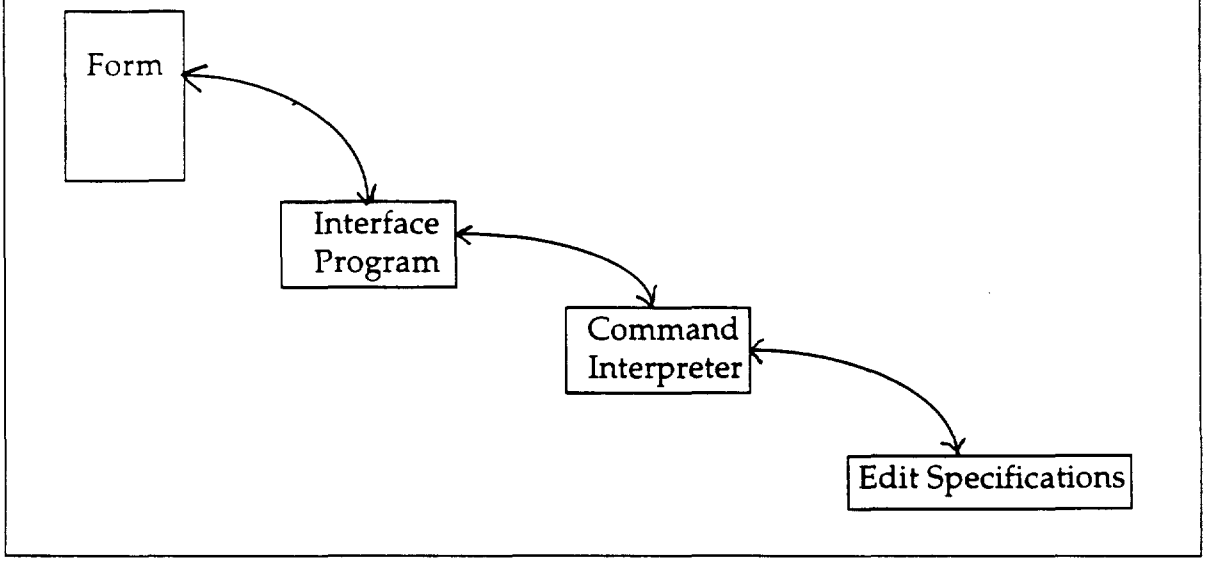
An interesting variation to this approach is the idea of accessing the editing engine to perform checks on data entered into an HTML or Java electronic form. Data entered into the form are processed using a technique called Common Gateway Interface (CGI). A CGI program could call the DC2 editing engine (and the ESL code) and return any edit violations to the user. For reasons mentioned below, this is likely the most feasible use of the DC2 editing environment.

The other side of this equation is the potential of the editing engine to be incorporated into or called from external systems. Some considerations are:

- The ESL can be executed wherever the command interpreter exists.

- The command interpreter is designed to be part of a system; that is, it is designed to be called (accessed dynamically at run-time).

Figure 3.--Accessing the DC2 Editing Engine from External Systems



- The interface requirements are simple and well defined. Essentially, there are two types of edit interfaces: one for single field editing and one for editing two or more fields.

- Data access methods (to an Oracle database) are restricted to discrete routines which could be replaced if required to access some other data storage mechanism.

- The most likely computing platform for laptop applications and electronic questionnaires is Intel-based and running some flavour of Microsoft Windows or Windows NT. Written in ANSI C, the command interpreter could be ported to these platforms. However, this should be considered a non-trivial exercise. The interpreter was developed for UNIX and there are some major differences which would require careful retooling. For example, the UNIX and Windows TIME routines operate differently. TIME is an important element in any statistical editing mechanism. Byte ordering is also different on these two platforms. These types of restrictions suggest that the DC2 editing facilities could be more easily and more cheaply adopted for wider use on UNIX platforms.

Conclusions

For the majority of mixed-mode applications, the DC2 system offers a solution to the problem of having to write and maintain separate versions of edit specifications. Whether the method of collection/capture is by paper questionnaire, telephone interview or electronic data report, an edit can be defined and maintained at the survey level and used seamlessly across reporting modes.

The ability of reusing the edit specification does not restrict the possibilities for choosing relevant behaviour based on the mode of collection. For example, error messages can be reported to the user when appropriate.

It is technically feasible to use the DC2 editing engine in other external collection and capture systems. However, feasibility should not be confused with desirability. The task of porting the engine to a non-UNIX platform may not be cost-justifiable. Commercial systems may not be sufficiently "open" to incorporate DC2's editing facilities and the editing facilities may have platform dependencies which make it difficult to move from the UNIX environment to the Windows platform. Statistics Canada will be evaluating these possibilities and others in its pursuit of methods and techniques for reducing the instances of edit specifications required by applications.

References

Ait-Kaci, Hassan (1991). Warren's Abstract Machine, Massachusetts Institute of Technology.

Chrisholm, John (1995). Surveys by E-Mail and Internet, *Unix Review*, pp. 11-16.

Dillman, Don A. and Tarnai, John (1988). Administrative Issues in Mixed Mode Surveys, *Telephone Survey Methodology*, John Wiley & Sons, Inc., pp. 509-528.

Statistics Canada (1992). ***DC2** Edit Specification Language Reference Manual.* ■

CDC Edits: Tools for Writing Portable Edits

J. Tebbel and T. Rawson,
U. S. Centers for Disease Control and Prevention

15
Chapter

Abstract

The CDC EDITS Project has produced a system for writing collections of executable data edits, which can be distributed as part of a public standard. These collections of edits can be used by interactive data entry programs to achieve real-time field-by-field validation or in batch processes for data already collected.

EditWriter is a complete menu-driven development environment for creating, maintaining, testing, and documenting data edits. Individual edit checks are written in the EDITS language, a C-like language with simplifications and extensions for the editing task.

EditWriter is capable of creating and manipulating all the structures needed to test data: code snippets, data dictionaries, record layouts, and reference tables. The output of EditWriter is an object called the "Metafile."

The EDITS Engine is an interpreter that processes the Metafile when called by an application program to test a field or record of data. It is supplied as C-language source code that can be compiled and linked on a variety of computing platforms or as a Dynamic Link Library (DLL) for use with most database packages in the Windows environment.

EDITS Metafiles have been used since 1993 to improve the quality and efficiency of processing in CDC's national Behavioral Risk Factor Surveillance System. National standard-setting organizations for cancer registry data have adopted EDITS and are currently distributing Metafiles.

CDC Edits: Tools for Writing Portable Edits

J. Tebbel and T. Rawson, U. S. Centers for Disease Control and Prevention

Introduction

The Centers for Disease Control and Prevention (CDC) created the EDITS system to improve the quality of data collected by cancer registries. CDC's Division of Cancer Prevention and Control administers the National Program of Cancer Registries (NPCR) authorized by Public Law 102-515, which was passed in 1992; the program's goal is to help establish new state cancer registries and to update existing ones. Some of the existing registries have collected many years of data in a variety of existing systems, but there is no agreed-upon standard for data checking. This shortcoming impairs the use of the data, a real concern as researchers seek more information about cancer and how to prevent or control it.

CDC's EDITS is a collection of computer programs and data objects. These software tools were intended to encourage independent authorities, who sometimes have competing interests, to contribute to and accept voluntary, shared public standards for data quality. EDITS was also intended to provide the means for efficient development, testing, documentation, and publication of standard data checks in an executable form. The system is neither cancer- nor health-specific; it can be used for any type of data on a variety of hardware platforms and in diverse operating system environments.

Edits as Quality Assurance

Even when data collectors intend to adhere to a standard, the details of field-by-field checking often vary according to the decisions made by individual programmers. The EDITS system eliminates this source of variability by producing a portable, executable version of data-checking logic as specified by the authority for a standard. The data object, which contains an expression of the validation rules, can then be distributed for direct execution upon files and records of data in a variety of processing scenarios. The same edits can be applied at different points in the flow of data through a system; data already collected can be checked in batch mode, and new data can be tested as they are being entered. This feature makes it easy to integrate EDITS into existing systems; processors can apply existing standards in batch mode with very little cost or disruption to operations. When the correction of errors will be costly, identical edit logic can be attached to data entry programs to catch mistakes when they are most readily corrected.

Setting and implementing data standards is not without potential problems, as detailed in Figure 1. These problems are not necessarily completely solvable with software, but portable edits provide a good beginning.

Figure 1.--Some Advantages to the EDITS System

Problem	How EDITS Addresses
Multiple organizations may set and revise conflicting or overlapping standards for the same data item.	Supports consensus-building among standard setters and enables collaboration by showing differences.
Programmers interpret each standard and render it into code for individual system	Standards are directly executable and portable across languages and platforms to avoid reinterpretation.
Data collected differently became different data and may not be comparable.	Data checked by the same edits may be comparable. Difference in edits serves as documentation of how data differ.
Adopted standards are sometimes adapted for local needs or ease of processing.	Availability of standard in executable form may eliminate need for adaptation.
Only simple edits implemented in interactive mode, more complex cross-field checks saved for batch mode.	Same edits can be used for interactive and batch mode.
Standard sometimes buried deep in source code. Documentation sometimes missing, or out-of-sync with edit logic.	Standard separated from application source code so it can be developed, tested, and maintained separately. Documentation can be kept with logic. Can generate reports with logic and documentation.

Cancer Registry Specifics

Cancer registries may report data to one ot more of the following: the NCI SEER program, the North American Association of Central Cancer Registries (NAACCR), or the National Cancer Data Base (NCDB). The data are submitted in a standardized format, but data users need assurance that each field was collected uniformly across all data collection activities. Currently, cancer registry applications are in use on multiple platforms, including MS-DOS, Windows, UNIX, and VMS. As many of these applications implement (at least partially) existing standards, the solution is not necessarily to have additional standards. As a user of data from all of these sources, CDC facilitated the development of EDITS to provide a better means of expressing and using data standards, with the ultimate goal of improving data quality.

Development of EDITS System

The EDITS System was developed with input from competing cancer registry software providers.

In 1991, development began at CDC with a rapid prototype of a linkable C language interpreter module. The concept of having portable edits that could be used anywhere the C interpreter could be compiled was tested and proven with exploratory programming. Performance during this early test was adequate for interactive processing of a record or a few fields at a time but required enhancement for batch processing of large files. Over time, the language evolved slightly from C to make the edit logic

more readable and add extensions specific to data editing. The final design addresses performance issues by replacing the C interpreter with a compiler and p-code interpreter for faster edit execution and by adding indexes for faster lookup table access.

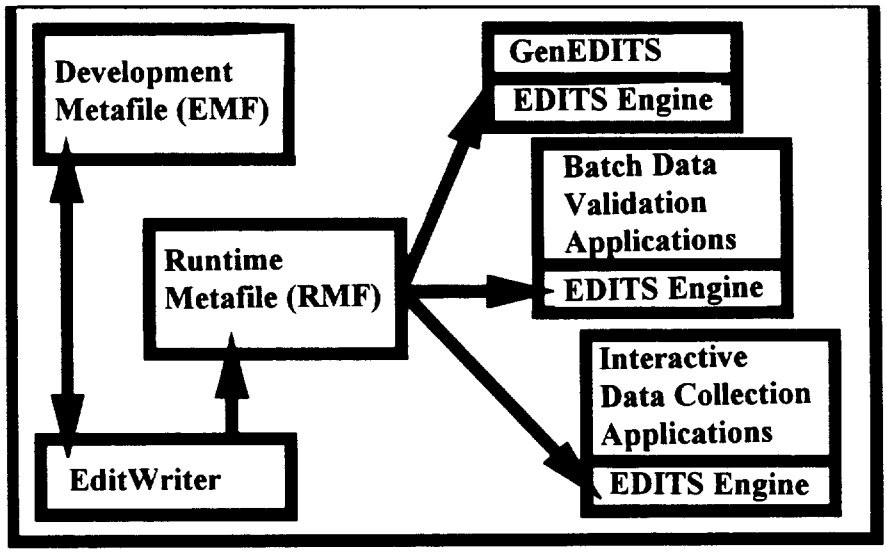
Components of EDITS System

The EDITS Metafile is a single file that encapsulates internal databases. There are tables for the data dictionary, record layouts, edits, and error messages; there can also be any number of lookup tables for table-driven edits. The Metafile has both development and compiled run-time formats; the run-time Metafile is portable across architectures.

The EditWriter is an integrated development environment for maintaining a Metafile that can create and maintain data dictionaries, define record layouts, write and interactively test edits, and create and import tables. The present version of EditWriter is an MS-DOS application written in C and FoxPro.

Edits are written in the EDITS language, which is based loosely on C with a rich function library for editing data. It is compiled to p-code to obtain a mix between speed and portability. Edits may be thousands of lines in length, or just two or three.

The EDITS Engine accesses run-time Metafiles to execute edits and return error messages. It is called via the EDITS Application Program Interface (API) and is callable from C for MS-DOS and other platforms and as a Windows DLL for use by any Windows language. There are options available for edit execution, including SKIPFAIL for skipping multi-field edits where any single field has failed and SKIPEMPTY, which skips edits where any field is blank.



This diagram shows the relationship of the different parts of the EDITS system. EditWriter is used to maintain a Metafile and compile it to a run-time Metafile.The EDITS Engine is then called by batch and interactive applications to perform edits from the run-time Metafile. Multiple applications can use the same run-time Metafile.

Driver Programs Incorporate the EDITS API

An EDITS driver is any program that incorporates the EDITS API. In batch mode, a driver program can detect existing bad data. In interactive mode, edits can prevent bad data. EDITS Drivers can be written in C or in any Windows language

GenEDITS is a generic EDITS driver that works in batch mode only. It produces a report of errors encountered, including the name of the edit, error message, and a list of fields referenced in the edit. GenEDITS also includes a summary report of failure count by edit. It can be used for recoding or reformatting data and to calculate simple frequencies.

Current Uses

Most EDITS users are using MS-DOS platform in batch mode. CDC's Behavioral Risk Factor Surveillance System has been using EDITS at both the state level prior to data submission and at CDC since 1993. The NCI SEER program is now maintaining its edits both in the original COBOL and in EDITS, which gives it a portable solution. The EDITS system consistently gives the same results. NAACCR recently released a Metafile of cancer edits incorporating standards from SEER, the American College of Surgeons, and others.

EDITS is available at no charge and may be obtained by downloading via anonymous ftp (address: ***ftp.cdc.gov, path: /pub/Software/EDITS***), World Wide Web (address: ***http://www.naaccr.org***), or by contacting the authors. ■

Chapter 15

Skip Patterns and Response Bases: Graph Manipulation in Survey Processing

Robert F. Teitel, Abt Associates and George Washington University

Abstract

By incorporating explicit indications of skip instructions (or GOTOs) in a survey description language (which also includes the usual facilities for the question identification, question text, response values, recodes, etc.), it is possible to build an acyclic directed GRAPH of the data collection instrument.

The survey-derived graph -- with questions as nodes and skips as edges -- may be manipulated using graph algorithms to prepare the response bases for each question for display in the printed codebook, and to prepare the control information for a skip pattern editor for use while processing the actual data.

This talk will describe these processes and illustrate some of the results. ■

16
Chapter

Software Demonstrations

Software Demonstrations

[Alphabetically, by first author]

> ***Editors Note:*** Many of the Workshop presentations were accompanied by software demonstrations. Also, in some cases, the presenters chose to do only an exposition. All of the demonstrations are listed below. If a presentation was also included as part of the regular sessions, please see the appropriate chapter for the abstract or paper. If only a demonstration was given or if the presenters provided a separate paper to describe the exposition, that material is included in this chapter.

Richard J. Bennof, M. Marge Machen, and Ronald L. Meeks -- Data Editing Software for NSF Surveys

Ronald S. Biggar -- Generic Editor Developed with Powerbuilder and a Relational Database

Joel Bissonnette -- Generalized Edit and Imputation System for Numeric Data

Dale Bodzer -- PEDRO

Richard Esposito and Kevin Tidemann -- Extensions of ARIES at the BLS

Glen Ferri and Tom Ondra -- Towards a Unified System of Editing International Data

Robert Hood -- Improving the Quality of Survey Data Through an Interactive Data Analysis System

Michael Horrigan, Polly Phipps, and Sharon Stang -- On-Line Edits in the Survey of Employer-Provided Training

Givol Israel -- Automatic Transferring from Paper to ASCII Code

Mary Kelly -- Integrated Data Capture: A System for All Office of Compensation and Working Conditions Surveys

Stanley E. Legum -- A Computer-Assisted Coding and Editing System for Non-Numeric Educational Transcript Data

Sharon Mowry and Jason Bulson -- Distributed EDDS Editing Project (DEEP)

Mark Pierzchala, Roberta Pense, and Arnie Wilcox -- The 1995 June Area Frame Instrument for CAPI and Interactive Editing

Anne Rhodes, Kishau Smith, and Peter Goldstein -- Electronic Data Collection: The Virginia Uniform Reporting Sytstem

Demonstration Abstracts

❒ Generic Editor Developed with Powerbuilder and a Relational Database -- ***Ronald S. Biggar, National Center for Health Statistics***

An interactive demonstration of both editing National Ambulatory Care Survey data and imputation of nonresponse. The software was developed for use by statisticians and may be modified as survey requirements change. Powerbuilder was chosen as the development tool so that statisticians could interact with the data from their PC's, while the data are stored on a file server. In addition, the Windows-based graphical user interface improves the user friendliness of the process.

❒ PEDRO -- ***Dale Bodzer, U.S. Energy Information Administration***

PEDRO is an electronic data collection product that facilitates the fast, accurate, and efficient transmission of data from the respondent's remote site to the EIA computer facility. Using a PC for data entry, PEDRO provides the user with an image of a printed survey form. Users can enter information through the keyboard or by importing data from another computer system. PEDRO performs numerous quality checks comparing the data entered with established ranges, lists of accepted values, or criteria derived from data entered in the past. PEDRO automatically transmits the information via modem to EIA' computer facility. Security of the transmission is protected by passwords and all data are encrypted. Accuracy is ensured by several levels of error detection. PEDRO is available to respondents.

❒ Improving the Quality of Survey Data Through an Interactive Data Analysis System -- ***Robert Hood, National Agricultural Statistics Service***

The National Agricultural Statistics Service (NASS), an agency of the U. S. Department of Agriculture, conducts surveys in order to provide accurate and reliable agricultural forecasts and estimates for a variety of commodities. NASS has recently begun the implementation of an interactive data analysis system based on SAS/AF and SAS/EIS software to ensure the quality of its survey data.

Currently, a lot of time is spent editing incoming data with little time devoted to analysis before data are summarized. Present analysis tools are limited to data listings and outlier printouts. While useful, they are somewhat limited in the problems they flag, and resolution of problems generally involves time-consuming review of paper questionnaires or data files. The time between data collection and summarization is very limited and must be used as efficiently as possible.

The authors have developed a SAS-based application to interactively analyze survey data. This system identifies potential "risky records" during the data collection period. Users are able to more efficiently analyze the data and resolve problems in a more timely manner. This Interactive Data Analysis System (IDAS) is an easy to use mouse-driven system that requires little knowledge of the SAS system. Pushbuttons, icons, list menus, and program entries provide easy selection of options. This paper gives an overview of IDAS and its development.

❒ On-Line Edits in the Survey of Employer-Provided Training -- ***Michael Horrigan, Polly Phipps, and Sharon Stang, U.S. Bureau of Labor Statistics***

The Survey of Employer-Provided Training 2 (SEPT2) was conducted from May to October, 1995 by the Division of Special Studies of the Bureau of Labor Statistics (BLS). The survey data were collected by BLS regional field economists using laptop computers during a personal visit to establishments. Employer and employee representatives were interviewed for SEPT2, and several instruments were administered to each respondent. Due to multiple survey instruments, the decentralized nature of the survey and the desire to avoid telephone callbacks to respondents to clarify inconsistencies, on-line edits with edit-error messages and summaries were included in the SEPT2 instruments. In addition, the SEPT2 laptop system included real-time reports on the status of establishment cases, a function for leaving case notes, and it incorporated Windows standards, such as a graphical-user interface and help system.

Edits involved logical, range or other consistency checks within an instrument.

During an interview when an answer to a question triggered an edit error, a pop-up window appeared with a message describing the error, what questions it involved and how to resolve it. The field economist could move on to other questions, but needed to resolve all errors before the establishment could be transmitted as complete. To resolve an error in situations where data items were not available from the respondent (missing) or when the datum was correct, field economists could select a comment for the question, such as data not available, verified by respondent or employment growth and provide case notes to describe the inconsistencies. When exiting an instrument,an edit error summary was automatically displayed, listing the remaining edit errors by question and the respective comment code. The field economist could simply double click on the question to return to it and resolve the edit error before exiting the instrument.

❒ Distributed EDDS Editing Project (DEEP) -- ***Sharon Mowry and Jason Bulson, Federal Reserve Board***

A key resource supporting the monetary policy-making and open market operations of the Federal Reserve is data representing the daily balances of deposits, borrowings, and reserves for the largest 7,700 depository institutions in the United States. The data require a high level of confidence, and identifying and resolving errors in quality must be done under increasingly stringent deadlines.

The Distributed EDDS Editing Project (DEEP) was initiated with the objective to improve the data analysis effort through the utilization of graphical user interface tools in conjunction with the presentation of statistically significant data edits. The DEEP system is a Windows-based client/server application designed to provide analysts with a sophisticated tool to access both raw data and data that have fallen outside of the data model forecasts based upon five different types of data edits or forecasts. Our presentation will detail the manner in which features of the DEEP application are employed for the editing and analysis of these critical data.

❒ Imputing Numeric and Qualitative Variables Simultaneously -- ***Sylvie Rivest and Mike Bankier, Statistics Canada***

At the Bureau of the Census 1996 Annual Research Conference, a presentation entitled "Imputing Numeric and Qualitative Variables Simultaneously" will be given by Mike Bankier. It describes the New Imputation Methodology (NIM) that will be used in the 1996 Canadian Census to impute the basic

demographic variables: age, sex, marital status and relationship to person 1. The NIM allows, for the first time, minimum change hot deck imputation of numeric and qualitative variables simultaneously.

As a follow-up to the presentation at the Annual Research Conference, it is proposed to give a software exhibit of the mainframe implementation of the NIM which will be used in the 1996 Canadian Census. Mike Bankier, the Senior Methodologist responsible for NIM and Sylvie Rivest, the System Analyst, who implemented it, will be present. They will give short presentations during the 3 hour slot given to demonstrate the software. A brief outline of the presentation is given below.

A small slide show (PC-based) will explain the generalized nature of the NIM program and the input data required. The slide show will also present the User Edit Interface used by NIM in the mainframe environment (3-5 minutes).

A PC version of the NIM program will then be used to demonstrate the functionality of the imputation engine (5-10 minutes). The methodology supporting NIM can be demonstrated interactively as the program is being executed.

Finally, a short (3-5 minutes) slide show will demonstrate the effort of applying the NIM methodology on a large volume of data. Statistics from the NIM testing done so far will be used to show the improvements in Data Quality that the NIM methodology offers on a large scale basis. These statistics will open informal discussions with the other participants at this Software Exhibit.

❒ Editing Occupations from Income Tax Returns -- ***Peter Sailer and Terry Nuriddin, Internal Revenue Service, and Gary Teper, Information Spectrum, Inc.***

Special tabulations, such as selected tax return data classified by Standard Occupational Classification (SOC) codes, are produced by the Statistics of Income (SOI) Division of the Internal Revenue Service. To facilitate coding, a computerized occupation-coding dictionary has been developed. However, because the manual updating process is time-consuming and increases errors, a computer utility program was created to automatically research and edit or replace the occupational entries.

The program compares occupational titles (from tax returns) to similar occupational titles already used in the dictionary. If similar entries are suggested, the user can replace the original entry with the best suggested entry. If no entries are suggested, the user can edit the original entry and retry or determine the record uncodable and bypass. Display will include a computer with the editing software installed; hands-on experience will be possible.

Extensions of ARIES at the Bureau of Labor Statistics

Richard Esposito and Kevin Tidemann,
U. S. Bureau of Labor Statistics

Abstract: A PC-based demo of prototypes that build on and extend the existing ARIES graphical approach to editing Current Employment Statistics Data will be shown. These prototypes include an updated version of ARIES, which incorporates visual representations of statistical measures of sample standard deviations to be used as aids to outlier detection and a DOS version of ARIES adapted to statistical exploration of universe data.

ARIES (Automated Review of Industry Employment Statistics) is a graphical and query PC-based data review system which has significantly enhanced the sample screening and estimation review procedures in the Current Employment Statistics (CES) program of the U. S. Bureau of Labor Statistics. During the time that ARIES has been in use, a number of areas of possible improvement have been suggested by the industry analysts who are responsible for reviewing the CES sample data and resulting estimates. As a result of studying these areas of possible improvement, we have developed two further prototypes to extend the capabilities of ARIES, which were shown at the Data Editing Conference, and outline these prototypes in this paper. Fuller treatments of ARIES and the first prototype are given in the *Journal of Computational and Graphical Statistics*, June 1994, and in the ASA 1994 *Proceedings of the Section on Statistical Graphics.*

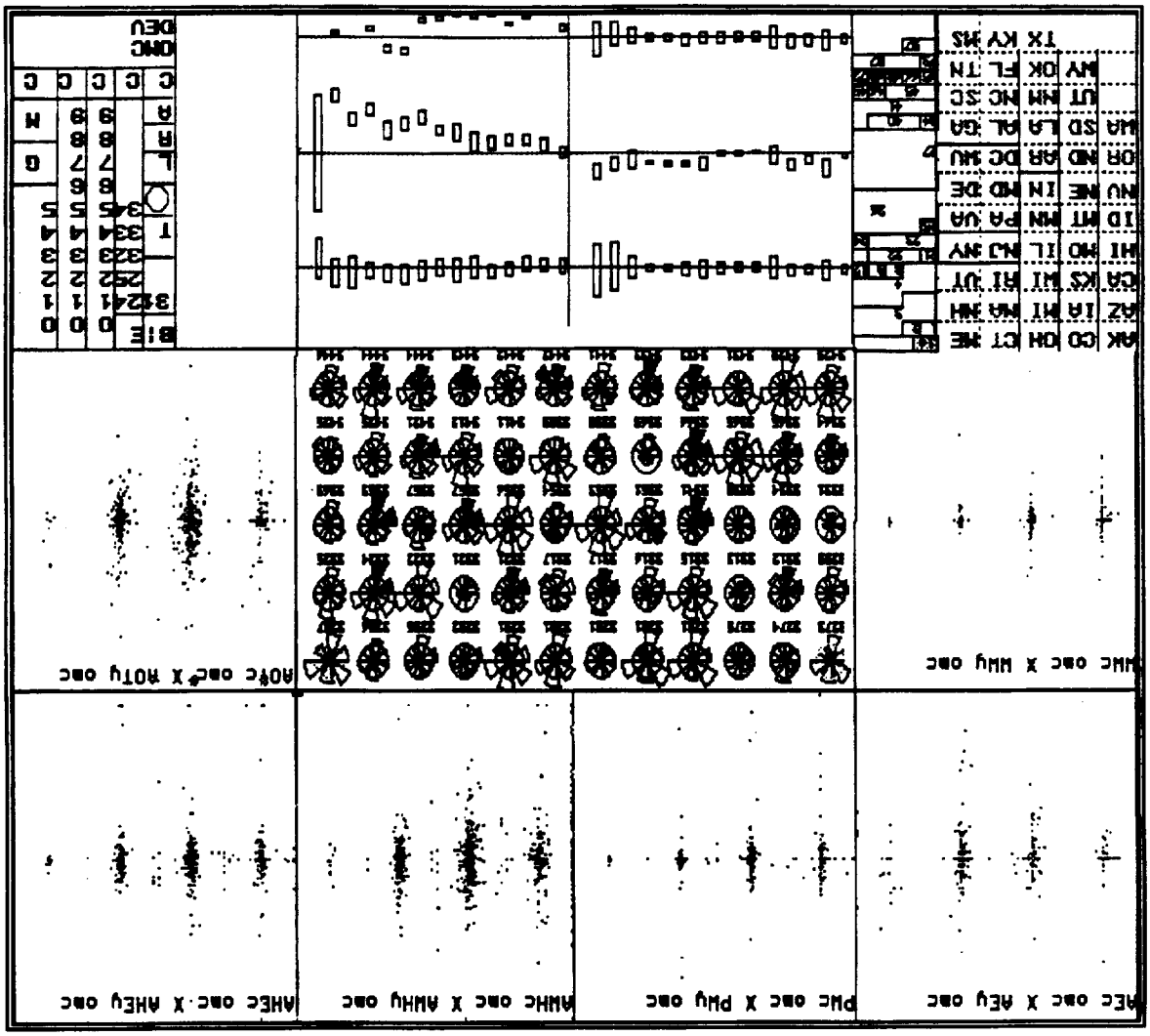
The prototypes have been designed to meet the specific needs of the CES data review process. As the prototypes move closer to production use, it is to be expected that their design and underlying principles may be suitably modified. Perhaps the most important principle evident in the prototypes is "put as much information on a single screen as reasonably possible." Figure 1, on the following page -- in which ARIES has been adapted to simultaneously show all 6 data variables -- shows the results of this idea.

Underlying this principle is the desire to have more layers of information immediately accessible, without having to page through an extensive hierarchy of screens. In Figure 1, rather than showing the scattergrams of the different variables on separate screens, all six variables are shown simultaneously (the upper row and left and right center row boxes). In this way, cross-comparisons among data variables are facilitated. Additionally, in the option shown in Figure 1, each scattergram has been subdivided by size of establishment (4 strata), so that the relative importance of sample establishments can be immediately seen. The actual PC screen uses colors liberally, as indicators of various characteristics and measures, including, somewhat redundantly, size of establishments.

Just as in ARIES, when one selects points in the scattergrams using a mouse, the corresponding establishment information will appear (not shown here), overriding the center and bottom row screens.

A second principle followed in this prototype is to make every element on the screens both informative and interactive. That means that each object on the screen should be designed to provide statistical information, as well as be used as an index to further information or further actions. This is true of the scattergrams, which both show statistical movement of the sample, as well as allowing retrieval of detailed individual or groups of individual establishment information.

Figure 1.--Scattergrams for 6 Variables, Daisy Variations, and Rough Tukey Box Plots



O.K., what are the strange objects in the center of Figure 1? First, each "daisy" of the array of 60 daisies in the center of the screen functions as an index to a specific SIC industry. When a specific daisy is selected by the mouse, the scattergrams corresponding to that industry immediately appear. In addition, rough tukey box plots corresponding to each of the six variables appear at the bottom row center. Each sequence of 15 vertical bars represents a monthly time-series of the sample standard deviations for that variable, with the current month's sample standard deviation represented by the rightmost bar within each of the six sequences. In our review, we concentrate on the new current month sample data, since we presume that we've already done an O.K. job cleaning up bad data for previous months. One possible indicator of outliers in the current month might be an especially large standard deviation for the current month as compared to previous months. Such a case is shown in Figure 1 for several of the variables.

Following the 2nd principle above, besides using the daisies as indexes to select statistical information, we have also made the daisies statistically informative: each of the 6 petals of each daisy represents one of the six variables we produce monthly estimates for in the CES, and the shape of each daisy

functions as an informative index to the rough tukey box plots just mentioned. If the size of a petal is larger than normal, that represents that the current month's standard deviation for that variable is larger than the average standard deviation of the previous 14 months. That average point is represented by the circle or "flowerpot" within each daisy. Each analyst will have approximately 250 industries to review, so by selecting those daisies that flop outside the circle, the analyst can pinpoint industries with unusually large standard deviations, and immediately show the associated scattergrams and box plots.

The following prototype shows ARIES adapted for universe data and for sampling and estimation capabilities:

Figure 2.--One Variable, 11 Months, with Establishments Selected in Scattergrams and Corresponding Sample Establishment Time-series Shown in the Center

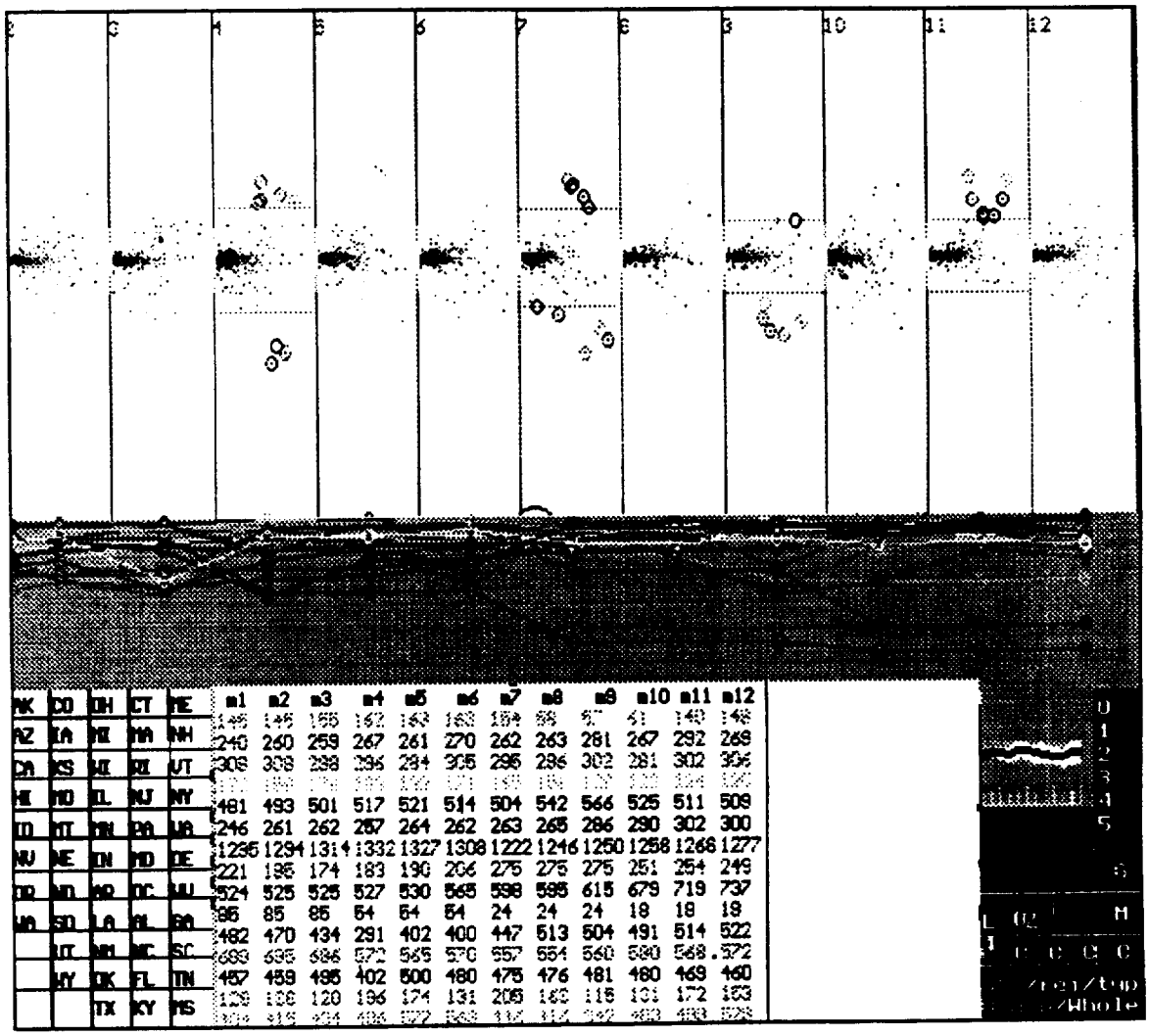
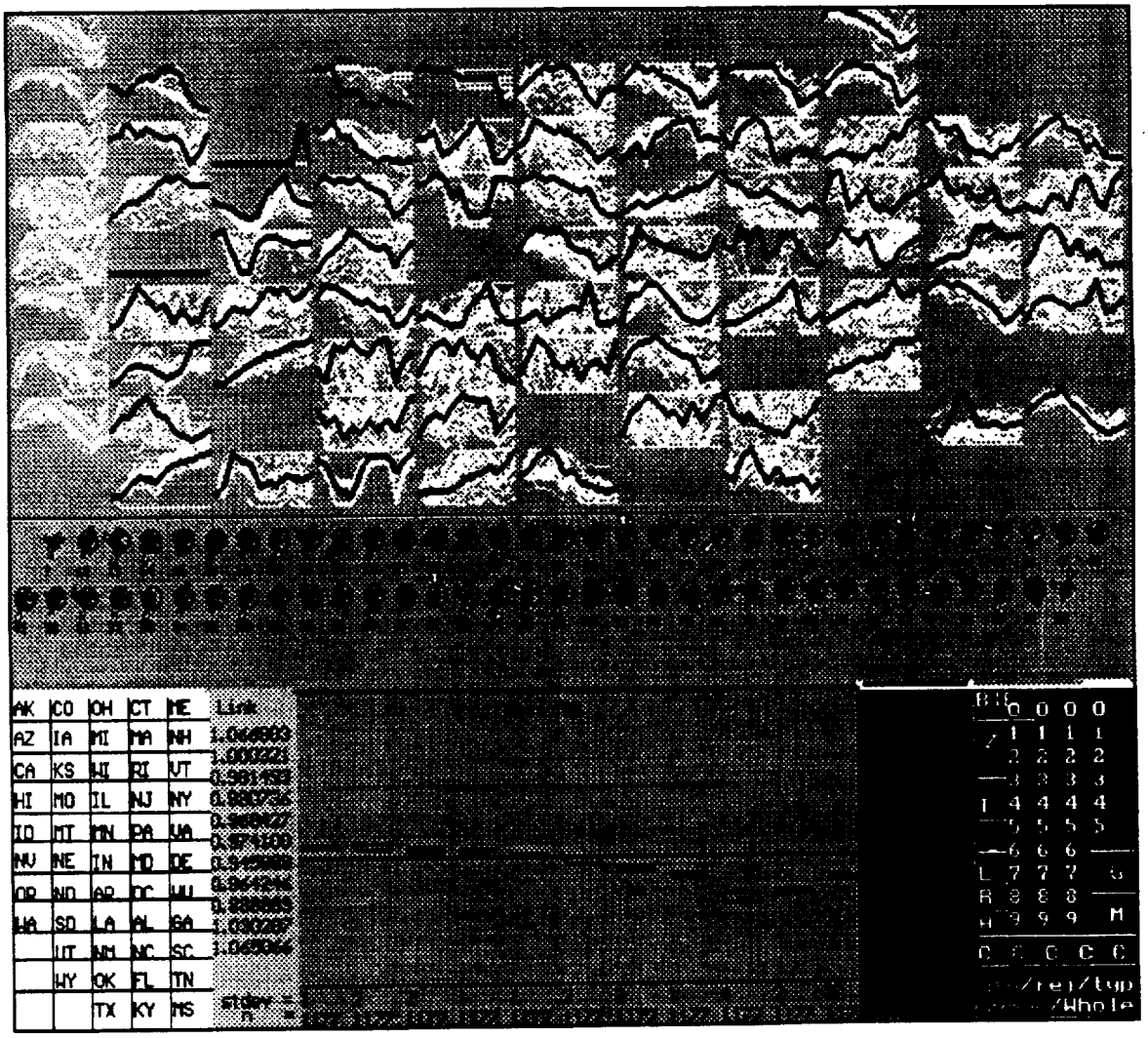


Figure 2 represents a simple adaptation to handle 11 months of a single data variable's sample change from month to month, rather than the six variables of Figure 1; we view more months but viewer variables. In this case, however, we show universe data rather than sample data, so this prototype is suitable for viewing and editing the frame.

Figure 3 moves the whole process in the direction of sampling, editing and estimation research. Figure 3 is an option of Figure 2. The middle row contains reduced size 2-digit SIC daisies for the 12 months of universe data for the industry employment variable, and the daisies work analogously to those of Figure 1. The top window of the screen graphically represents the results of selecting 30 random samples from the universe data for each 2-digit SIC industry, and computing the 30 independent time-series of monthly estimates for each 2-digit industry. The actual universe time-series is shown as the solid black line, with each distinct sample-based time-series as a white line. One can immediately see how well the estimates fared for each 2-digit industry. The left-most column represents the sum (i.e., the 1-digit SIC estimates) of each row.

Figure 3.--Estimates for 30 Random Samples as White Time-series, Universe as Solid Black



This type of comparison between estimates and universe can of course only be done once one has universe data, which in our program are fortunately available some months after we produce our sample-based estimates. However, with the capability to instantaneously select many samples, compute estimates and visually (and also numerically) portray the results, we can test different editing practices, different sampling techniques, and different stratification patterns, to determine an optimum combination of techniques to improve the estimates we produce.

References

Esposito, Richard; Fox, Lin; and Tidemann, Kevin (1994). ARIES: A Visual Path in the Investigation of Statistical Data, *Journal of Computational and Statistical Graphics*, Volume 3, Number 2, 113-125.

Esposito, Richard; Fox, Lin; and Tidemann, Kevin (1994). ARIES: Visual Techniques for Statistical Data Investigation at the Bureau of Labor Statistics, American Statistical Association, *Proceedings of the Section on Statistical Graphics*, 7-10. ■

Chapter 17

Appendix

DATA EDITING WORKSHOP AND EXPOSITION
BUREAU OF LABOR STATISTICS
MARCH 22, 1996

IN MEMORIAM: MARIA ELENA GONZALEZ (1932-1996)

SPONSORS
Bureau of Labor Statistics
Federal Reserve Board
Internal Revenue Service
National Center for Education Statistics

SPONSORS
Federal Committee on Statistical Methodology
Joint Program in Survey Methodology
National Agricultural Statistics Service
Washington Statistical Society

CONFERENCE SESSION I:
TIME: 9:00am-10:15am

OVERVIEWS

LOCATION: CONFERENCE ROOM #2

Chair: Fred Vogel, National Agricultural Statistics Service

A Paradigm for Data Editing
Linda M. Ball, U.S. Bureau of the Census

The New View on Editing
Leopold Granquist, Statistics Sweden

Data Editing at the National Center for Health Statistics
Kenneth W. Harris, National Center for Heath Statistics

FELLEGI-HOLT SYSTEMS

LOCATION: CONFERENCE ROOM #7

Chair: John Kovar, Statistics Canada

DISCRETE, A Fellegi-Holt System for Demographic Data
William E. Winkler, U.S. Bureau of the Census

Generalized Edit and Imputation System for Numeric Data **[Separate DEMO]**
Joel Bissonnette, Statistics Canada

The New SPEER Edit System
William E. Winkler, U.S. Bureau of the Census

ON-SITE DATA CAPTURE

LOCATION: CONFERENCE ROOM #9

Chair: George Hanuschak, National Agricultural Statistics Service

Data Collection by EDI: Some Quality Aspects
Wouter J. Keller, Statistics Netherlands

PERQS (Personalized Electronic Reporting Questionnaire System) **[Separate DEMO]**
Janet Sear and Peter Garneau, Statistics Canada

Electronic Data Collection: the Virginia Uniform Reporting System **[Separate DEMO]**
Anne Rhodes and Kishau Smith, Virginia Commonwealth University, and Peter Goldstein, NI-STAR Data Systems

CASE STUDIES

LOCATION: CONFERENCE ROOM #1

Chair: Leda Kydoniefs, Bureau of Labor Statistics

Towards a Unified System of Editing International Data **[Separate DEMO]**
Glen Ferri and Tom Ondra, U.S. Bureau of the Census

Data Editing Software for NSF Surveys **[Separate DEMO]**
Richard J. Bennof, Marge Machen, and Ronald L. Meeks, National Science Foundation

CONFERENCE SESSION II:
TIME: 10:45am-12:00pm

CENSUSES

LOCATION: CONFERENCE ROOM #2

Chair: Clyde Tucker, Bureau of Labor Statistics

Automated Record Linkage and Editing: Essential Supporting Components in the Data Capture Process
[Separate DEMO]
Givol Israel and Olivia Blum, Central Bureau of Statistics, Israel

Editing and Imputation Research for the 2001 Census in the United Kingdom
David Thorogood, Office of Population Censuses and Surveys, United Kingdom

A Priority Index for Macro-Editing the Netherlands Foreign Trade Statistics
Frank van de Pol and Bert Diederen, Statistics Netherlands

GRAPHICAL/INTERACTIVE SYSTEMS

LOCATION: CONFERENCE ROOM #7

Chair: Cynthia Z. F. Clark, National Agricultural Statistics Service

Experiences on Changing to PC-based Visual Editing in the Current Employment Statistics Program
[Separate DEMO]
Richard Esposito, Cynthia Engel, Laura Freeman, Bill Goodman, and Mike Murphy, U.S. Bureau of Labor Statistics

Graphical Editing and Query System (GEAQS)
[Separate DEMO]
Paula Weir, U.S. Energy Information Administration

Integrated Data Capture: A System for All Office of Compensation and Working Conditions Surveys
[Separate DEMO]
Mary Kelly, U.S. Bureau of Labor Statistics

CATI-CAPI TECHNICAL

LOCATION: CONFERENCE ROOM #9

Chair: Fred Wensing, Australian Bureau of Statistics

Questionnaire Programming Language (QPL)
Kevin Dooley, U.S. General Accounting Office

Using a Parallel CATI Instrument to Edit Call Record Information and Removal of Incorrect Interview Data
Timothy Triplett and Beth Webb
University of Maryland - College Park

A Computer-Assisted Coding and Editing System for Non-Numeric Educational Transcript Data
[Separate DEMO]
Stanley E. Legum, Westat, Inc.

STATISTICAL TECHNIQUES

LOCATION: CONFERENCE ROOM #1

Chair: Alan Estes, Federal Reserve Board

Rethinking the Editing Algorithm for the Survey of Employment Payrolls and Hours
Michael Scrim, Statistics Canada

A Statistical Edit for Livestock Slaughter Data
[Separate DEMO]
Linda Simpson, Henry Chiang, and Cathy Tomczak, National Agricultural Statistics Service

A CSFII Data User's Principal Components Analysis for Outlier Detection
Adeline J. Wilcox, Beltsville Agricultural Research Center

CONFERENCE SESSION III:
TIME: 1:00pm-2:15pm

CASE STUDIES

LOCATION: CONFERENCE ROOM #2

Chair: Ann Hardy, Centers for Disease Control

Third-Party Documents: The Line Between Editing and Imputation
Clifford Adelman, U.S. Department of Education

Sampling Design and Estimation Properties of a Study of Perinatal Substance Exposure in California
Jimmy Hwang, National University (San Diego)

The Processing and Editing System of the National Health Interview Survey: the Old and New
Susan S. Jack, National Center for Health Statistics

NEURAL NETWORKS

LOCATION: CONFERENCE ROOM #7

Chair: Ray Board, Federal Reserve Board

Data Editing Using Neural Networks
L. H. Roddick, Statistics Canada

Editing and Imputation by Means of Neural Networks
Svein Nordbotten, University of Bergen, Norway

CATI-CAPI CONCEPTUAL

LOCATION: CONFERENCE ROOM #9

Chair: Denice Myers, National Agricultural Statistics Service

Statistics Canada's Experience in Moving to CAI from Paper and Pencil
R. Jamieson, Statistics Canada

Developing an On-line Editing System for Respondents Reporting via Touchtone Data Entry
David O'Connell, U.S. Bureau of Labor Statistics

CAI and Interactive Editing in One System for a Survey in a Multimode Environment **[Separate DEMO]**
Mark Pierzchala, National Agricultural Statistics Service

STATISTICAL TECHNIQUES

LOCATION: CONFERENCE ROOM #1

Chair: Linda Stinson, Bureau of Labor Statistics

Time Series and Cross Section Edits
David Pierce and Laura Gillis, Federal Reserve Board

Inflation Factors for Stratified Samples with Control Information
Peter Ochshorn, State University of New York, Albany

Empirical Data Review: Objective Detection of Unusual Patterns of Data
James Kennedy, U.S. Bureau of Labor Statistics

CONFERENCE SESSION IV:
TIME: 2:45pm-4:00pm

CASE STUDIES

LOCATION: CONFERENCE ROOM #2

Chair: Charles Day, National Agricultural Statistics Services

Time-Series Editing of Quarterly Deposits Data
Anusha Fernando Dharmasena, Federal Reserve Board

Experiences in Re-engineering the Approach to Editing and Imputing Canadian Imports Data
Clancy Barrett and Francois Laflamme, Statistics Canada

Data Editing in an Automated Environment: A Practical Retrospective -- The CPS Experience
Gregory D. Weyland, U.S. Bureau of the Census

STATISTICAL TECHNIQUES

LOCATION: CONFERENCE ROOM #7

Chair: Sylvia Kay Fisher, Bureau of Labor Statistics

Statistical Analysis of Textual Information
Thierry Delbecque, Sid Laxson, Nathalie Millot, INFOWARE, Inc.

The Impact of Ratio Weighting
Jai Choi, National Center for Health Statistics

Fitting Square Text into Round Computer Holes - An Approach to Categorizing Textual Responses Using Computer-Assisted Data Entry
Richard Wendt, I. Hall, P. Price-Green, R. Dahara, W. Kaye, U.S. Centers for Disease Control

EDIT AUTHORING TECHNIQUES

LOCATION: CONFERENCE ROOM #9

Chair: Jim O'Reilly, Research Triangle Institute

Methods of Reusing Edit Specifications Across Collection and Capture Modes and Systems
Shirley Dolan, Statistics Canada

CDC EDITS: Tools for Writing Portable Edits
[Separate DEMO]
J. Tebbel and T. Rawson, U.S. Centers for Disease Control

Skip Patterns and Response Bases: Graph Manipulation in Survey Processing
Robert F. Teitel, Abt Associates and George Washington University

SOFTWARE DEMONSTRATIONS SESSION I:
TIME: 9:00am-12:00pm

SOFTWARE DEMONSTRATIONS

LOCATION: CONFERENCE ROOM #4

Editing Occupations from Income Tax Returns
Pete Sailer and Terry Nuriddin, Internal Revenue Service, and Gary Teper, Information Spectrum, Inc.

On-Line Edits in the Survey of Employer-Provided Training
Michael Horrigan, Polly Phipps, and Sharon Stang, U.S. Bureau of Labor Statistics

SOFTWARE DEMONSTRATIONS SESSION I (Continued): TIME: 9:00am-12:00pm

SOFTWARE DEMONSTRATIONS

LOCATION: CONFERENCE ROOM #5

Generic Editor Developed with Powerbuilder and a Relational Database
Ronald S. Biggar, National Center for Health Statistics

Improving the Quality of Survey Data Through an Interactive Data Analysis System
Robert Hood, National Agricultural Statistics Service

A Computer-Assisted Coding and Editing System for Non-Numeric Educational Transcript Data*
Stanley E. Legum, Westat, Inc.

CDC EDITS: Tools for Writing Portable Edits*
J. Tebbel and T. Rawson, U.S. Centers for Disease Control

**Indicates that an oral presentation accompanies this software demonstration*

SOFTWARE DEMONSTRATIONS

LOCATION: CONFERENCE ROOM #6

Distributed EDDS Editing Project (DEEP)
Sharon Mowry and Jason Bulson, Federal Reserve Board

Imputing Numeric and Qualitative Variables Simultaneously
Sylvie Rivest and Mike Bankier, Statistics Canada

PEDRO
Dale Bodzer, U.S. Energy Information Administration

Automatic Transferring from Paper to ASCII Code*
Givol Israel, Central Bureau of Statistics, Israel

**Indicates that an oral presentation accompanies this software demonstration*

SOFTWARE DEMONSTRATIONS SESSION II: TIME: 1:00pm-4:00pm

SOFTWARE DEMONSTRATIONS

LOCATION: CONFERENCE ROOM #4

The 1995 June Area Frame Instrument for CAPI and Interactive Editing*
Mark Pierzchala, Roberta Pense, and Arnie Wilcox, National Agricultural Statistics Service

PERQS (Personalized Electronic Reporting Questionnaire System)*
Janet Sear and Peter Garneau, Statistics Canada

**Indicates that an oral presentation accompanies this software demonstration*

SOFTWARE DEMONSTRATIONS

LOCATION: CONFERENCE ROOM #5

Integrated Data Capture: A System for All Office of Compensation and Working Conditions Surveys*
Mary Kelly, U.S. Bureau of Labor Statistics

Graphical Editing and Query System (GEAQS)*
Paula Weir, U.S. Energy Information Administration

Data Editing Software for NSF Surveys*
Richard J. Bennof, Marge Machen, and Ronald L. Meeks, National Science Foundation

Generalized Edit and Imputation System for Numeric Data*
Joel Bissonnette, Statistics Canada

**Indicates that an oral presentation accompanies this software demonstration*

SOFTWARE DEMONSTRATIONS

LOCATION: CONFERENCE ROOM #6

Extensions of ARIES at the BLS*
Richard Esposito and Kevin Tidemann, U.S. Bureau of Labor Statistics

A Statistical Edit for Livestock Slaughter Data*
Linda Simpson, National Agricultural Statistics Service

Towards a Unified System of Editing International Data*
Glen Ferri and Tom Ondra, U.S. Bureau of the Census

Electronic Data Collection: the Virginia Uniform Reporting System*
Anne Rhodes and Kishau Smith, Virginia Commonwealth University, and Peter Goldstein, NI-STAR Data Systems

**Indicates that an oral presentation accompanies this software demonstration*

ACKNOWLEDGEMENTS

A conference of this magnitude requires many contributions from many people in addition to the members of the organizing committee listed on this page. Special thanks to the sponsoring organizations and agencies, and to Statistics Canada, the Bureau of the Census, and the Energy Information Administration for additional assistance in making this a successful conference.

Thanks also to the session chairs, and especially to the presenters of talks and software demos for sharing their work.

Finally, behind the scenes are many additional people who contributed their time and energy to ensure the success of this venture, some of whom are: John Bosley, Monica Dashen, Leda Kydoniefs, Deborah Stone and Daphne Van Buren, Bureau of Labor Statistics; Mia Johnson, Melissa Walsh and Nancy Thomas, Federal Reserve Board; Wendy Alvey, Internal Revenue Service; and Wendy Waters, National Agricultural Statistics Service.

DATA EDITING WORKSHOP AND EXPOSITION: ORGANIZING COMMITTEE

Dave Pierce, Chair
Federal Reserve Board

Mark Pierzchala, Co-Chair
National Agricultural Statistics Service

Yahia Ahmed
U.S. Internal Revenue Service

Frances Chevarley
National Center for Health Statistics

Charles Day
National Agricultural Statistics Service

Rich Esposito
U.S. Bureau of Labor Statistics

Sylvia Kay Fisher
U.S. Bureau of Labor Statistics

Laura Bauer Gillis
Federal Reserve Board

Maria Elena Gonzalez
U.S. Office of Management and Budget

Robert Groves
Joint Program in Survey Methodology

Ken Harris
National Center for Health Statistics

David McDonell
National Agricultural Statistics Service

Renee Miller
U.S. Energy Information Administration

Denice McCormick Myers
National Agricultural Statistics Service

Jeff Owings
National Center for Education Statistics

Tom Petska
U.S. Internal Revenue Service

Linda Stinson
U.S. Bureau of Labor Statistics

Paula Weir
U.S. Energy Information Administration

Bill Winkler
U.S. Bureau of the Census

List of Attendees at the Data Editing Workshop and Exposition

Margret Adams
The Arbitron Company

Clifford Adelman
Department of Education

Yahia Ahmed
Internal Revenue Service

Jaspreet Ahuja
Agricultural Research Service

Tony Alvarez
Federal Reserve Board of NY

Darrell Anderson
Bureau of the Census

Gary Anderson
Federal Reserve Board

Craig Anderson
National Agricultural Statistics Service

Mark Apodaca
National Agricultural Statistics Service

Maribel Aponte
Bureau of the Census

Catherine Armington
Consultant

Mary Brightwell Arnold

Joy Aso
Bureau of the Census

Donna Atkinson
Birch & Davis Assoc., Inc.

Dale Atkinson
National Agricultural Statistics Service

R. Clifton Bailey
Health Care Financing Admin.

Linda M. Ball
Bureau of the Census

D. Catherine Baltzell
General Accounting Office

Scott Banker
The Arbitron Company

Mike Bankier
Statistics Canada

Peggy Barker
National Center for Health Statistics

Betty Barlow
Energy Information Admin.

Oscar Barnhardt
Federal Reserve Board

Clancy Barrett
Statistics Canada

Carl Barsky
Bureau of Labor Statistics

Lisa Bastian
Defense Manpower Data Ctr.

Stephanie Battles
Energy Information Admin.

Karen Beauregard
Agency for Health Care Policy and Research

Bernard Bell
Bureau of Labor Statistics

Michael E. Bellow
National Agricultural Statistics Service

Millie Bendl
Westat, Inc.

Frank Benford
National Agricultural Statistics Service

Rich Bennof
National Science Foundation

Suzan Benz
National Agricultural Statistics Service

Sherry Beri
Energy Information Admin.

Martha Bethea
Federal Reserve Board

Karil Bialostosky
National Center for Health Statistics

Zelia Bianchini
IBGE-Diretoria de Pesquises,Brazil

Ronald S. Biggar
National Center for Health Statistics

Yvonne M. Bishop
Energy Information Admin.

Joel Bissonnette
Statistics Canada

Dan Blanchette
University North Carolina, Chapel Hill

Olivia Blum
Central Bureau of Statistics, Israel

Ray Board
Federal Reserve Board

Janice E. Bodner
Agricultural Research Service

Dale Bodzer
Energy Information Admin.

Thomas P. Bonczar
Bureau of Justic Statistics

Lisa Bradburn
Bureau of Economic Analysis

Douglas Braddock
Bureau of Labor Statistics

Bradley Braden
Health Care Financing Admin.

Raphail Branch
Bureau of Labor Statistics

Laura Branden
Westat, Inc.

Russ Bredbenner
Bureau of the Census

Gary Breslau
Abacus Technology

Beth Bridgeman
Westat, Inc

Debra Brody
National Center for Health Statistics

Thomas Broene
Energy Information Admin.

Camilla A. Brooks
CBQ

Stephen P. Broughman
National Center for Education Statistics

Elizabeth R. Brown
Colorado Dept. of Public Health
and the Environment

Prudy Brown
National Research Council

Janet A. Brunelle
National Institutes of Health

Bill Brykczynski
Institute for Defense Analyses

Linda Buckles
National Agricultural Statistics Service

Paul Buckley
Abt Associates, Inc.

Andrew Bullivant
Statistics New Zealand

Jason Bulson
Federal Reserve Board

Eugene M. Burns
Energy Information Admin.

Jim Burt
National Agricultural Statistics Service

Henry R. Burt
The World Bank

Gerald Bushee
American College of Radiology

Michael Buso
Bureau of Labor Statistics

Shail Butani
Bureau of Labor Statistics

John K. Butler, Jr
Bureau of the Census

Fe Caces
CSR, Incorporated

Karen Campbell
CSR, Inc.

Brian S. Canham
U.S. Department of Labor

John C. Cannon
Bureau of the Census

Susan Capella
Statistics Canada

Margaret Carroll
National Center for Health Statistics

William Chan
American College of Radiology

Chen-ning Chan
New Jersery Dept of Agriculture

Frances Chevarley
National Center for Health Statistics

Henry Chiang
National Agricultural Statistics Service

Pei Lu Chiu
National Center for Health Statistics

Jai Choi
National Center for Health Statistics

Paul Cichello
Bureau of Labor Statistics

Cynthia Z. F. Clark
National Agricultural Statistics Service

Nicole Close-Zimmerman

Michael P. Cohen
National Center for Education Statistics

Mary Colbert
National Center for Health Statistics

Gloria Colclough
Centers for Disease Control

Richard Coles
National Center for Health Statistics

Elizabeth Cologer
Bureau of Economic Analysis

Nancy Connelly
Cornell University

Dan Conti
Bureau of Labor Statistics

Timothy M. Copeland
Centers for Disease Control

Rosalee Copeland
Aspen Systems Corp.

Jim Cotter
National Agricultural Statistics Service

Barbara Cox
Bureau of Labor Statistics

R. Dahara
Centers for Disease Control

Donald R. Dalzell
Bureau of the Census

Lance Daugherty
Bureau of Economic Analysis

Jeanette Davis
Bureau of Labor Statistics

Jeannie Davis
Westat, Inc.

Joelle Davis
DOE

Trenita R. Davis
National Institute of Dental Research

Jennifer Day
Bureau of the Census

Charles Day
National Agricultural Statistics Service

Ron DeCarlo
Bureau of Labor Statistics

Thierry Delbecque
INFOWARE, Inc.

Maria Deloria
National Institute for Allegy
and Infectious Diseases

Sonia Demers
Statistics Canada

Raoul Depoutot
Eurostat, Luxembourg

Dennis DeRycke
Westat, Inc.

Anusha Dharmasena
Federal Reserve Board

Dave Dickerson
Bureau of the Census

Bert Diederen
Statistics Netherlands

Richard Dietz
Bureau of Labor Statistics

Gregg J. Diffendal
Bureau of the Census

Cathryn Dippo
Bureau of Labor Statistics

Barbara N. Diskin
Bureau of the Census

Shirley Dolan
Statistics Canada

Kevin Dooley
General Accounting Office

Pat Doyle
Agency for Health Care Policy and Research

Wanda Dreslin
Federal Reserve Board

Patricia W. Dunham
National Center for Health Statistics

Linnea Efner
Westat, Inc.

Bruce Eklund
National Agricultural Statistics Service

Nabil El-Khorazaty
Research Triangle Institute

Bob Emery
SAIC

Rich Esposito
Bureau of Labor Statistics

Alan Estes
Federal Reserve Board

James T. Fagan
Bureau of the Census

Kenneth H. Falter
Centers for Disease Control

Tracy E. Farmer
Bureau of Labor Statistics

Ron Fecso
National Agricultural Statistics Service

Sara Fein
Food & Drug Administration

Dania Ferguson
U.S. Department of Agriculture

Glenn Ferri
Bureau of the Census

Gary Feuerberg
Defense Manpower Data Center

Sylvia Fisher
Bureau of Labor Statistics

Tom Flood
Bureau of the Census

Cristina Ford
Food & Drug Administration

Ann Forquer
Bureau of Labor Statistics

Henry Foster
National Agricultural Statistics Service

Gabriel B. Fosu
Univ. of Maryland, Baltimore

Jean Fowler
Agricultural Research Service

Alan Fox
U.S. Department of Housing and Urban Development

Howard B. Fredrick
Energy Information Admin.

Fred Freme
Energy Information Admin.

Carol French
Energy Information Admin.

Madeleine Friedlander
Bureau of Labor Statistics

Gerhard Fries
Federal Reserve Board

Nancy Gagne
National Center for Health Statistics

David Gailer
Bureau of Economic Analysis

John S. Gardenier
National Center for Health Statistics

Peter Garneau
Statistics Canada

Nick Gerbino
Federal Reserve Board

Kimberly Giesbrecht
Bureau of Census

Laura Bauer Gillis
Federal Reserve Board

Israel Givol
Central Bureau of Statistics, Israel

Edmund Glad
Bureau of Labor Statistics

Margaret E. Goldsmith
Bureau of the Census

Peter Goldstein
NiStar Data Systems Inc.

Z-F Gosselin
Bureau of Labor Statistics

Phil Graffunder
Centers for Disease Control

Richard W. Graham
Bureau of the Census

Leopold Granquist
Statistics Sweden

Donna Gray
Food and Drug Administration

Amy Green
Westat, Inc.

Brian V. Greenberg
Bureau of the Census

Richard Greene
Temple University

Dennis Griffin
Bureau of Labor Statistics

James Grounds
Bureau of Labor Statistics

Sioux Groves
Bureau of Labor Statistics

Robert Groves
JPSM -- University of Maryland

Kerry J. Gruber
National Center for Education Statistics

Patricia M. Guenthen
Agricultural Research Service

Leverett Lynn Guess
Research Triangle Institute

Louise Guey-Lee
Energy Information Admin.

Tanya J Guthrie
Defense Manpower Data Center

Etta Susanne Haggerty
U.S. Department of Agiculture

I. Hall
Agency for Toxic Substances and Disease Registry

Theresa Hallquist
Energy Information Admin.

George Hanuschak
National Agricultural Statistics Service

Linda Hardy
National Science Foundation

Ann Hardy
National Center for Health Statistics

Glenn Harke
Westat, Inc.

Jane L. Harman
National Center for Health Statistics

Ken Harris
National Center for Health Statistics

Rachel Harter
National Opinion Research Center

Anita Hartke
Federal Reserve Board

Mohammad Hasan
UTA

Barbara Haupt
National Center for Health Statistics

Mary Sue Hay
Defense Manpower Data Center

Katherine Heck
National Center for Health Statistics

Kelly Heilman
MD Health Resources Planning Commission

Robert J. Hemming
Bureau of the Census

Peter Henderson
National Research Council

Julie Heneberry
The Arbitron Company

Tammy Heppner
Energy Information Admin.

Doug Herold
Federal Reserve Board

Mary D. Herr
Delaware Div. of Public Health

John S. Hilton
Bureau of the Census

Jacy Y. Hobson
Federal Reserve Board

Anani K. Hoegnifioh
IPA, Inc.

Sandra Hofferth
Institute for Social Research

William Holman
Westat, Inc.

Ann Marit K. Holmoy
Statistics Norway

Robert Hood
National Agricultural Statistics Service

Michael Horrigan
Bureau of Labor Statistics

Easley Hoy
Bureau of the Census

Paul L. Hsen
Bureau of Labor Statistics

Kevin J. Hudson
Equal Employment Opportunity Commission

Cheryl Hurt
Westat, Inc.

Wenke Hwang
Johns Hopkins University

Jimmy Hwang
University of California, San Diego

Ho-Ling Hwang
Oak Ridge National Laboratory

Linda Ingwersin
U.S. Department of Agriculture

Cathleen Irish
Consumer Product Safety Commission

Bill Iwig
National Agricultural Statistics Service

Susan S. Jack
National Center for Health Statistics

KA Jagannathan
Admin. for Children & Families

R. Jamieson
Statistics Canada

Willis Jefferson
Bureau of the Census

Donna M. Jewell
Research Triangle Institute

Fabian Jimenez
Bureau of Labor Statistics

Mia Johnson
Federal Reserve Board

Gerald A. Joireman
U.S. Department of Agriculture

Dalia Kahane
Westat, Inc.

Kelly Kang
National Science Foundation

Sam Yong Kang
Bureau of Labor Statistics

Roy Kass
Energy Information Admin.

Devi Katikineni
Social & Scientific Systems, Inc.

SK Katti
University of Missouri

Irvin Katz
Bureau of Labor Statistics

W. Kaye
Agency for Toxic Substances and Disease Registry

Yonnas Kefle
Bureau of Labor Statistics

Wouter J. Keller
Statistics Netherland

Mary Kelly
Bureau of Labor Statistics

James Kennedy
Bureau of Labor Statistics

Arthur B. Kennickell
Federal Reserve Board

Eileen Kessler
Consumer Product Safety Commission

Rick Kestle
National Agricultural StatisticsService

Meena Khare
National Center for Health Statistics

Kay Lee Khuu
Internal Revenue Service

Nancy Kieffer
Social & Scientific Systems, Inc.

Nora Kincaid
Bureau of Labor Statistics

Carol S. King
Bureau of the Census

Jim Knaub
Energy Information Admin.

Dave Knopf
National Agricultural Statistics Service

Carol M. Knowles
Centers for Disease Control

Corey Koenig
Federal Reserve Board, Kansas City

Michael Kornbau
Bureau of the Census

John G. Kovar
Statistics Canada

Robert S. Krasowski
National Center for Health Statistics

Nancy A. Krauss
ACOG

Mary Kuta
Highway Loss Data Institute

John Laffman
Bureau of Economic Analysis

Francois Laflamme
Statistics Canada

Lauchland Lake
Acting Chief Statistician, St. Johns, Antigo

Danielle Lalande
Statistics Canada

Ron Lambrecht
Energy Information Admin.

K. Patrick Lampani
Federal Reserve Board

James LaVern
Bureau of Labor Statistics

Sid Laxson
INFOWARE, Inc.

Douglas Lee
Bureau of the Census

Soo Lee
National Center for Health Statistics

Stanley E. Legum
Westat, Inc.

Ibo Levent
Development Data Group

Ibrahim Levent
The World Bank

Virginia Lewis
Federal Reserve Board

Fred Licari
Temple University

Larry Lie
Bureau of Labor Statistics

Alice Lippert
Energy Information Admin.

Stephen Litavecz
Westat, Inc.

Yan Liu
George Washington Univ.

Baiming Liu
Cornell University

Barbara Livington
Internal Revenue Service

Nancy Loester
Bureau of Labor Statistics

Tecla C. Loup
University of Michigan

Ruey-Pyng Lu
Energy Information Admin.

Richard Lutyk
Bureau of Economic Analysis

Don Lutz
U.S.Department of Agriculture

Marge Machen
National Science Foundation

Jim MacIntosh
Bureau of Labor Statistics

Mary Madden
Orkard Corp.

Naoki Makita
Statistics Bureau of Japan

Robert Malin
Bureau of Labor Statistics

Vilas Mandlekar
The World Bank

Asa Manning,
National Agricultural Statistics Service

Richard Mantovani
Macro International

Angelita A. Manuel
Social & Scientific Systems, Inc.

Stephen Marcus
National Institute of Dental Research

John S. Martin
Merck Medco Managed Care, Inc.

Antoinette Ware Martin
Energy Information Admin.

Donald A. Mathews
Aspen Systems Corp.

Marilyn Mattson
Westat, Inc.

Betty D. Maxfield
Defense Manpower Data Center

Joseph Mbu
CPCS

Mary McCarthy
Bureau of Labor Statistics

Patrick McCarthy
Food and Drug Administration

David McDonell
National Agricultural Statistics Service

Megan McKee

Andrew McKeen
SAIC

Don McKinnie
National Agricultural Statistics Service

Sarah McLaughlin
George Washington Univ.

Tommy McLemore
National Center for Health Statistics

Kathy McMahon
Merck Medco Managed Care, Inc.

John McPeek
National Center for Health Statistics

Gary McQuown
Macro International

Paul Medzerian
Bureau of Economic Analysis

Ronald L. Meeks
National Science Foundation

Jeanine Mellem
Centers for Disease Control

Egee Mengestu

Nathalie Millot
INFOWARE, Inc.

Susan Mitchell
National Research Council

Nash J. Monsour
Bureau of the Census

Wesley Montgomery
Bureau of Labor Statistics

Ron Monticone
Consumer Product Safety Commission

Kevin Moore
Federal Reserve Board

Mary T. Moore
Fish & Wildlife Service

Sharon Mowry
Federal Reserve Board

Tom Mullen
Birch Davis Associates, Inc.

M. Denice McCormick Myers
National Agricultural Statistics Service

Steve Nalley
Abacus Technology

Gad Nathan
Central Bureau of Statistics, Israel

Jack Nealon
National Agricultural Statistics Service

Elizabeth Nelson
Internal Revenue Service

Thomas Nephew
CSR, Inc.

Clara Nilles
Bureau of Labor Statistics

Svein Nordbotten
University of Bergen, Norway

Antonia H. Nowell

Terry Nuriddin
Internal Revenue Service

Benita O'Colmain
Macro International

D. O'Connell
Bureau of Labor Statistics

Jim O'Reilly
Research Triangle Institute

Nancy O'Reilly
ACOG

Peter Ochshorn
SUNY at Albany

Carol Odbert,
National Institutes of Health

Kenneth P. Offord
Mayo Clinic

Fred L. Olson
SCORE

Tom Ondra
Bureau of the Census

Mauricio Ortiz
Bureau of Economic Analysis

Jeff Owings
National Center for Education Statistics

Margaret Pacious
Westat, Inc.

Mike Pallesen
National Agricultural Statistics Service

Thomas Palumbo
Bureau of the Census

Young J. Park
Fu Associates

Christina H. Park
National Center for Health Statistics

Mary Parran
Bureau of Labor Statistics

Charles P. Pautler, Jr
Bureau of the Census

Guy Pense
National Agricultural Statistics Service

Roberta Pense
National Agricultural Statistics Service

Tom Petska
Internal Revenue Service

Tai Phan
National Center for Education Statistics

Polly Phipps
Bureau of Labor Statistics

David Pierce
Federal Reserve Board

Daphne Pierre
National Lead Information Center

Mark Pierzchala
Westat, Inc.

Steve Pinney
University of Michigan

Val Pisacane

Tim Pivetz
Bureau of Labor Statistics

Thomas J. Plewes
Bureau of Labor Statistics

Lorraine Porcellini
Temple University

D.E.B. Potter
Agency for Health Care Policy and Research

Jeffrey J. Potts
Bureau of Labor Statistics

P. Price-Green
Agency for Toxic Substances and Disease Registry

Bela Prigly
Statistics Canada

Laura Prizzi
Internal Revenue Service

Joe Prusacki
National Agricultural Statistics Service

Jiahe Qian
National Opinion Research Center

Magdalena Ramos
Bureau of the Census

Linda Ramsey
Statistics Canada

Nancy Raper
Agricultural Research Service

Tom Rawson
Centers for Disease Control

Elizabeth J. Reed
Westat, Inc.

Bill Regina
Bureau of the Census

Christine Reid
Federal Reserve Board

Linda M. Reiff
Consumer Product Safety Commission

Roy Reitz
Bureau of Labor Statistics

Jim Reitz
Bureau of Labor Statistics

Anne Rhodes
Survey Research Laboratory

Patricia Rhoton
Ohio State University

Jose G. Rigau
Centers for Disease Control and Prevention, Puerto Rico

Jean Ritzer
Statistics Netherlands

Emilda B. Rivers
Energy Information Admin.

Sylvie Rivest
Statistics Canada

Lee Robeson
Response Analysis

Hugh Roddick
Statistics Canada

Jane P. Rollow
Oak Ridge National Laboratory

Allison Rose
Research Triangle Institute

Wendy Rotz
Internal Revenue Service

Jay Ryan
Bureau of Labor Statistics

Don Saboe
National Agricultural Statistics Service

Pete Sailer
Internal Revenue Service

Laurie Salmon
DOAS

Gordon Sande
Sande & Associates

Ronny Schaul
Social Security Administration

Fritz Scheuren
George Washington Univ.

John Schmidt
Federal Reserve Board

Mary Lynn Schmidt
Bureau of Labor Statistics

Richard Schroeder
Bureau of Labor Statistics

Kim Schultz
Federal Reserve Board

Richard Schumann
Bureau of Labor Statistics

Sid Schwartz
U.S. Postal Service

Jimmie B. Scott
Bureau of the Census

Elizabeth Scott
Energy Information Admin.

Chester Scott
National Center for Health Statistics

Michael Scrim
Statistics Canada

Janet Sear
Statistics Canada

Barbara Sedivi
Bureau of the Census

Janet Shapiro
Bureau of the Census

Peter Shin
George Washington Univ.

Jim Shippey

Richard J. Shute
Merck-Medco Managed Care, Inc.

Diane Sickles
Westat, Inc.

Richard Sigman
Bureau of the Census

Adriana Silberstein
Bureau of Labor Statistics

Sam Simmens
George Washington Univ.
Medical Center

Linda Simpson
National Agricultural Statistics Service

Irving L. Skinner
Bureau of Economic Analysis

Sara Slater
American Speech-Language-
Hearing Assoc.

Sam Slowinski
Federal Reserve Board

Kishau Smith
Virginia Commonwealth Univ.

Valerie Smith
IPA, Inc.

Kemi Somuyiwa
Fish & Wildlife Service

Sandra Sperry
Westat, Inc

Reina Sprankle
Westat, Inc.

Sharon Stang
Bureau of Labor Statistics

Martha Starr-McCluer
Federal Reserve Board

Philip M. Steel
Bureau of the Census

Richard Sterner
Bureau of the Census

William S. Stewart
The Tarrance Group

Sin Stormer
Statistics Norway

Anne K. Stratton
National Center for Health Statistics

Debra A. Street
Food and Drug Administration

Joanne Streeter
National Science Foundation

Michael T. Stroot
Bureau of the Census

Annika Sunden
Federal Reserve Board

Jonathan Sunshine
American College of Radiology

Nittaya Suppapanya
Merck-Medco Managed Care, Inc.

Priya Suresh
Research Triangle Institute

Brian J. Surette
Federal Reserve Board

Rebecca Sutterlin
American Association of
Retired Persons

Kathy Sykes
U.S. Department of Agriculture

Joyce Tabor
University of North Carolina, Chapel Hill

Tim Taccardi
SAIC

Junko Tamaki
Agricultural Research Service

Charles Tardif
Statistics Canada

Jim Tebbel
Centers for Disease Control

Robert F. Teitel
Abt Associates & GWU

Gary Teper
Information Spectrum, Inc.

Montie Tesky
National Agricultural Statistics Service

Katherine Thompson
Bureau of the Census

David Thorogood
Office of Population Census & Surveys,
U.K.

Kevin Tidemann
Bureau of Labor Statistics

Babgaleh B. Timbo
Food & Drug Administration

Terri Tinker
The Arbitron Company

Mark Tokarski
Federal Reserve Board

Cathy Tomczak
National Agricultural Statistics Service

Amy Tong
Agricultural Research Service

Sharon Tossey
CSR, Inc.

Alan I. Trachtenberg
National Institutes of Health

Timothy Triplet
JPSM -- University of Maryland

Richard Troiano
National Center for Health Statistics

Mary N. Troxel
University of Maryland, Baltimore
County

Daniel Troy
ASTD

Sovan Tun
Equal Employment Opportunity
Commission

Linda Unger
Bureau of Labor Statistics

Frank van de Pol
Statistics Netherlands

David A. Vanderbroucke
U.S. Department of Housing and
Urban Development

Clemencia Vargas
National Center Health Statistics

Jose Villar
Bureau of Labor Statistics

Joan Vogel
General Accounting Office

Fred Vogel
National Agricultural Statistics Service

Gail Vossler
Westat, Inc.

Dennis Wagner
Bureau of the Census

Veronica R. Walgamotte
SAS Institute, Inc.

John M. Walker
SAIC

Sue Wallace
Federal Bureau of Prisons

Katherine Wallman
Office of Management and Budget

Melissa Walsh
Federal Reserve Board

Chongfang Wang
Delaware Div. of Public Health

Martina Wasmer
ZUMA

Kevin K. Watanabe
Westat, Inc.

Teresa Watkins
National Center for Health Statistic

Kimberly W. Webb
National Institute of Dental Research

Beth Webb
University of Maryland,
College Park

Paula Weir
Energy Information Admin.

Richard Wendt
Agency for Toxic Substances and
Disease Registry

Fred Wensing
Australian Bureau of Statistics

Gregory D. Weyland
Bureau of the Census

Bob White
National Agricultural Statistics Service

Diane Whitmore

William J. Wiatrowski
Bureau of Labor Statistics

Max Wigbout
Statistics New Zealand

Adeline J. Wilcox
Beltsville Agricultural Research Center

Arnie Wilcox
National Agricultural Statistics Service

Thomas G. Wilcox
Food and Drug Administration

George Wilkie
SAIC

Sheri S. Williams
Fish & Wildlife Service

Janet Williams
Bureau of Labor Statistics

Greg Wilson
Bureau of Labor Statistics

Janet Wilson
Internal Revenue Service

Alice Winkler
Bureau of Labor Statistics

Bill Winkler
Bureau of the Census

Pam Wisor
Westat, Inc.

Sandie Wodlinger
Westat, Inc.

Linda Wohlford
Bureau of Labor Statistics

Lily Wong
Bureau of the Census

Rebecca Wood
Texas Department of Health

Suzanne Worth

Laverne Wright
Defense Manpower Data Center

Jacqueline Wright
National Center for Health Statistics

Akihito Yamauchi
Statistics Bureau of Japan

Arthur K. Yao
Bureau of Labor Statistics

Frank Yu
Australian Bureau of Statistics

Julia Zachary
George Washington Univ.

Sandra Zak

Pamela Zorich
Maryland National Capital Park
and Planning Commission

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 Documents Sales, PB86-230570/AS)
9. *Contracting for Surveys* (NTIS Documents 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 Documents Sales, PB86-139730)
14. *Workshop on Statistical Uses of Microcomputers in Federal Agencies* (NTIS Document Sales, PB87-166393)
15. *Quality on 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)
20. *Seminar on the Quality of Federal Data* (NTIS Document Sales, PB91-142414)
21. *Indirect Estimators in Federal Programs* (NTIS Document Sales, PB93-209294)
22. *Report on Statistical Disclosure Limitation Methodology* (NTIS Document Sales, PB94-165305)
23. *Seminar on New Directions in Statistical Methodology* (NTIS Document Sales, PB95-182978)
24. *Electronic Dissemination of Statistical Data* (NTIS Document Sales, PB96-121629)
25. *Data Editing Workshop and Exposition* (NTIS Document Sales, PB97-104624) ■

Copies of these working papers may be ordered from NTIS Document Sales, 5285 Port Royal Road, Springfield, VA 22161; telephone: (703) 487-4650. The Statistical Policy Working Paper series is also available electronically through the Bureau of Transportation Statistics World Wide Web home page (**http: //www.bts.gov/fcsm/methodology**).