UNITED NATIONS STATISTICAL COMMISSION
and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS
STATISTICAL STANDARDS AND STUDIES - No. 48

STATISTICAL DATA EDITING

Volume No. 2

METHODS AND TECHNIQUES

UNITED NATIONS
New York and Geneva, 1997

CONTENTS

Page

PREFACE vii

INTRODUCTION ix

Chapter 1 WHAT TO DO WHEN AN EDIT FAILS 1

Foreword - *John Kovar, Statistics Canada* 1

Impact of editing on the salary statistics for employees in county councils - *Kajsa Lindell, Statistics Sweden* 2

Selective editing in the Netherlands annual construction survey - *Frank van de Pol, Statistics Netherlands* 8

Development of a cost-effective edit and follow-up process: the Canadian survey of employment experience - *Michel Latouche, Marcel Bureau, James Croal, Statistics Canada* 13

Simulation experiments for hot deck imputation - *Peter Verboon, Erik Schulte Nordholt, Statistics Netherlands* 22

Imputing numeric and qualitative variables simultaneously - *Mike Bankier, Jean-Marc Fillion, Manchi Luc, Christian Nadeau, Statistics Canada* 30

Chapter 2 DESIGNING SETS OF EDITS 39

Foreword - *Giulio Barcaroli, National Statistical Institute, Italy* 39

DAISY (Design, Analysis and Imputation System): structure, methodology and first applications - *Giulio Barcaroli, Marina Venturi, National Statistical Institute, Italy* 40

The SPEER edit system - *William E. Winkler, Lisa R. Draper, Bureau of the Census, USA* 51

The DISCRETE edit system - *William E. Winkler, Thomas F. Petkunas, Bureau of the Census, USA* 56

Generation of instruments for data collection and interactive editing - *Mark Pierczhala, National Agricultural Statistics Service, USA* 62

Survey of generalized systems used for edit specifications *Giulio Barcaroli, National Statistical Institute, Italy* 68

Chapter 3 **GRAPHICAL EDITING** **75**

Foreword - ***Ron James, Electronic Data Systems Ltd, United Kingdom*** **75**

Improving outlier detection in two establishment surveys - ***Julia L. Bienias, David M. Lassman, Scott A Scheleur, Howard Hogan, Bureau of the Census, USA*** **76**

The ARIES review system in the BLS current employment statistics program - ***Richard Esposito, Dong-yow Lin, Kevin Tidemann, Bureau of Labour Statistics, USA*** **84**

A description of a graphical macro-editing application - ***Per Engström, Christer Ängsved, Statistics Sweden*** **92**

The graphical editing analysis query system - ***Paula Weir, Department of Energy, Robert Emery, John Walker, Science Applications International Corporation, USA*** **96**

Chapter 4 **EVALUATION OF THE DATA EDITING PROCESS**

Foreword - ***Leopold Granquist, Statistics Sweden*** **105**

Statistical measurement and monitoring of data editing and imputation in the 2001 United Kingdom Census of Population - ***Jan Thomas, Census Division, Office for National Statistics, United Kingdom*** **106**

Selected issues of data editing - ***Bogdan Stefanovicz, Warsaw School of Economics, Poland*** **109**

An overview of methods of evaluating data editing procedures - ***Leopold Granquist, Statistics Sweden*** **112**

Evaluating data editing process, using survey- and register based data - ***Marius Ejby Poulsen, Statistics Denmark*** **123**

Chapter 5 **IMPACT OF NEW TECHNOLOGY ON DATA EDITING** **133**

Foreword - ***Ron James, Electronic Data Systems Ltd, United Kingdom*** **133**

Evaluating data quality with computer-assisted personal interviewing - ***Tom Pordugal, Roberta Pense, National Agricultural Statistics Service, USA*** **134**

Trend in technology for data collection and editing at Statistics Sweden - ***Evert Blom, Statistics Sweden*** **139**

Technology infrastructure used for data editing at Slovenia Statistics - ***Milan Katić, Statistical Office, Slovenia*** **146**

Electronic Data Interchange for Statistical Data Collection - ***Wouter J. Keller, Winfried F. H. Ypma, Statistics Netherlands*** **152**

New developments in automated data collection: Electronic Data Interchange and the World Wide Web - ***Richard L. Clayton, Tony M. Gomes, Louis J. Harrell jr., Bureau of Labour Statistics, USA*** **159**

Quality of optical reading the Census '91 in Croatia - ***Srdan Dumičić, Central Bureau of Statistics, Ksenija Dumičić, Faculty of Economics, Croatia*** **168**

Chapter 6 **AUTOMATED CODING** **173**

Foreword - *Pascal Rivière, Institute Nationale de la Statistique et des Etudes Economiques (INSEE), France* 173

Outline of a theory of automated coding - *Pascal Rivière, Institute Nationale de la Statistique et des Etudes Economiques (INSEE), France* 174

Trigram coding in the family expenditure survey in Statistics Netherlands - *Martje Roessingh, Jelke Bethlehem, Statistics Netherlands* 180

The 1991 Canadian Census of Population experience with automated coding - *Jocelyn Y. Tourigny, Joanne Moloney, Statistics Canada* 186

Automatic coding and text processing using N-grams - *Alois Haslinger, Central Statistical Office, Austria* 199

Automated coding in the Census '91 in Croatia - *Srdan Dumičić, Central Bureau of Statistics, Ksenija Dumičić, Faculty of Economics, Damir Kalpić, Vedran Mornar, Faculty of Electrical Engineering and Computing, Croatia* 209

Automatic coding of diagnosis expressions - *Lars Age Johansson, Statistics Sweden* 216

SICORE - general automatic coding system - *Pascal Rivière, Institute Nationale de la Statistique et des Etudes Economiques (INSEE), France* 222

GLOSSARY 233

PREFACE

This publication, Data Editing Methods and Techniques Volume 2, is the logical continuation of the first part of the series, which defined statistical data editing and presented associated methods and software. Volume 2 deals with *how* to solve individual data editing tasks, focusing on efficient techniques for data editing operations and evaluating the whole process. The aim of the publication is to assist National Statistical Offices in their efforts to improve and economise data editing processes.

The material was prepared in the framework of the project on Statistical Data Editing in the programme of work of the Conference of European Statisticians. It was reviewed at the work session on Statistical Data Editing in Athens (November, 1995) and compiled and edited by the Statistical Division of the United Nations Economic Commission for Europe. It represents an extensive voluntary effort on the part of the authors.

The editors express their thankfulness to all authors who contributed to the publication. Special thanks and appreciation goes to the members of the Steering Group, namely Giulio Barcaroli *(ISTAT, Italy)*, Dania Ferguson *(National Agricultural Statistics Service, USA)*, Leopold Granquist *(Statistics Sweden)*, Ron James *(Electronic Data Systems Ltd, United Kingdom)*, John Kovar *(Statistics Canada)*, Pascal Rivière *(INSEE, France)* and Bill Winkler *(U.S. Bureau of the Census, USA)*, for coordinating and introducing the individual chapters, and for helping to compile and edit the material. Their joint efforts contributed significantly to the preparation of the publication as a whole.

This book consists of an Introduction, 6 chapters and a Glossary. The Glossary is an update to the Glossary prepared by the Joint Group on Data Editing established inside the Statistical Computing Project of the United Nations Development Programme and the Economic Commission (ECE/UNDP/SCP/H.2).

INTRODUCTION

Definitions of data editing vary widely. In this Volume we concentrate on activities related to respondent contact, actual data collection, respondent follow-up, and manual and machine verification of rules. Let us use the Volume 1 definition of editing: "an activity aimed at the acquirement of data which meet certain requirements...". And more specifically, define editing as the procedure for detecting, by means of edit rules, and for adjusting, manually or automatically, errors resulting from data collection or data capture. Editing aimed at ensuring validity and consistency of individual data records is generally referred to as micro editing. By contrast, approaches which ensure the reasonableness of data aggregates are often referred to as macro editing, and procedures which target only some of the micro data items or records for review, by prioritizing the manual work and establishing appropriate and efficient process and edit boundaries, are termed selective editing methods. Chapter 1 deals with this topic in greater depth.

Furthermore, it is useful to distinguish between fatal edits and query edits. That is, respectively, edits which identify errors with certainty, and edits which point to suspicious data items. It is generally agreed that fatal errors must be removed, and that editing is very well-suited for this task. Chapter 2 deals in detail with the task of designing fatal edit sets, in particular as they relate to a number of automatic data editing software systems. Use of such systems can reduce the costs to some extent. Once in place, that is, once the infrastructure costs, implementation costs and feasibility testing costs have been absorbed, individual applications can be processed at lower marginal costs. In addition, numerous hidden benefits emerge: uniformity of approach, reproducibility, defensibility and consistency of approach, audit trail capabilities, etc.

Designing query edit sets, on the other hand, is unfortunately much more elusive. Exacerbating the situation is the fact that the query type edits are exactly those responsible for the high costs of editing. How tight should these rules be? How many? How often are they to be applied? While Chapter 1 illustrates how this has been done in several specific situations, it does not answer the question in general. There is no simple, single way of designing query edits. Each situation is different. One must study the process, learn from past experiences. Hit rates of individual edits must be evaluated, sources of error tracked down and corrected. It has to be understood that editing cannot possibly be expected to find all the errors, and that editing can in fact be counterproductive: not only because of the time delays caused but because, beyond a certain point, further editing is likely to introduce more errors than will be removed.

It is very probable that many editing strategies must be rethought. Adding more rules to an already cumbersome system must be avoided. Innovative approaches must be tried and carefully evaluated. Chapter 3 provides a handful of excellent examples of new and imaginative approaches to what is often a tedious task. The reproduced examples are intended to be only illustrations of what can be done, and as such constitute only a sampling of possibilities.

The goals of editing have been said to be threefold: to provide information about the quality of the data, to provide the basics for the (future) improvement of the survey vehicle, and to tidy up the data. Many studies suggest that a disproportionate amount of resources is concentrated on the third objective of "cleaning up of the data". Ample evidence also exists demonstrating the dangers of overediting, and the limitations of the activity. Chapter 1 deals with the topic of overediting in more detail. However, the primary goals of editing must be those of providing information about the process and the quality of the data. Properly maintained meta-databases, efficient use of the many available administrative sources and survey registers, are the tools to use to improve subsequent data releases. Learning from the process is paramount. Evaluating the edit system is thus crucial and is addressed here in Chapter 4.

Reducing errors in survey data is a question of doing it right the first time, rather than cleaning up at the end. To this end, moving the editing process to the early stages of the survey cycle, preferably while the respondent is still available, or in fact all the way to the respondent when possible, will go a long way in helping prevent errors. Computer-assisted data collection methods with appropriately designed edit sets will help this effort immensely, as is discussed in more detail in Chapter 5. An implicit message of this chapter is the warning that edit strategies must be adapted when moving from the paper and pencil environment to the computer-assisted world. One cannot simply move on-line what used to be done in batch mode. Furthermore, one should take advantage of the fact that many new opportunities present themselves because the respondent is available during the editing phase.

Finally in Chapter 6 we deal with a related topic of coding. As this activity often takes place during the

interview, or at the data capture stage, it is often intimately related to the editing task. A number of successful approaches to the automated coding are presented and the role of the computer is elaborated. The Volume ends with a brief glossary of terms used throughout this book.

How does a statistical agency put all this together? First, and likely foremost, is appropriate training of staff. Making all concerned aware of the potential usefulness as well as the dangers of editing is essential. Public discussion of best practices as well as unsuccessful strategies, by means of seminars, lectures, demonstrations and written documentation, is essential. Secondly, a number of successful organizations have found that a team approach to designing an edit strategy works best. Entrusting a team of specialists with the task, including statisticians, methodologists, informaticians as well as field operations personnel, can prove to be most efficient when evaluating alternate methods, investigating new approaches, designing new editing systems and seeking efficiencies. Of course, in order for the team approach to work, members must agree to respect each other's professional opinion. Communication is vital. Planned rotation of middle managers throughout the organization also helps, as it tends to spread out expertise and helps build bridges as old colleagues reunite under different circumstances.

Finally, while some editing will always be essential in eliminating gross errors and embarrassing inconsistencies, a much more productive role of editing must lie in its ability to provide the agency with information about the data quality, sources of errors, and areas for potential improvement. Editing results can be used wisely in the sharpening of survey concepts and definitions, improving the survey questionnaire be it paper or electronic, and optimizing data collection procedures. More resources must be spent on preventing errors rather than attempting to correct them at the end. As a result, careful monitoring and appropriate use of quality assurance and control of the editing process itself is crucial.

Chapter 1

WHAT TO DO WHEN AN EDIT FAILS

FOREWORD

by John Kovar, Statistics Canada

Data editing is one of the most expensive activities of the survey cycle. Numerous studies suggest that editing consumes between 20 to 40% of the total survey budget. Furthermore, ample evidence now exists indicating that the impact and usefulness of editing is limited. The hit rate of editing, that is the proportion of warnings that point to true errors, has been estimated to be of the order of 20 to 30%. Conversely, many true errors go undetected. In fact, commonly used edits cannot possibly detect small, systematic errors reported consistently in repeated surveys. It is also apparent that, especially in cases of quantitative survey data, very few errors contribute to the majority of the changes. Lindell's paper in this chapter demonstrates this highly differential impact of errors in the case of the Swedish study on Salary Statistics for Employees in County Councils.

What is responsible for the unacceptably high costs of editing? One can distinguish between fatal and query edits, and correspondingly between fatal and query errors. The fatal errors must be removed in order to retain the statistical office's credibility, and to simplify further automated processing. Editing is well-suited for this task, and relatively inexpensive as it can be easily automated. The inordinate cost of editing is due to the query edits which result in manual procedures, respondent follow-up, and often little or no change to the data. It is the latter type of editing that must be rationalized.

Not only is the usefulness of editing limited, it can in fact be counterproductive. There exists a point in time during the editing process when just as many errors are introduced as are corrected. Survey statisticians are now beginning to realize that most survey data are overedited, and that some form of selective editing should be considered.

Given the high cost of editing and its limited or even counterproductive effect on data quality, coupled with the highly differential impact of errors, it is clear that selective editing promises to have a potential for large savings. These are not only monetary savings, but savings due to reduced response burden and improved data timeliness and relevance. As a result, many statistical offices around the world have undertaken experimental studies that would allow them to reap some of these savings. Among the numerous successful methods and approaches, two more recent papers are reproduced in this chapter, those written by van de Pol, and Latouche, Bureau and Croal.

Van de Pol demonstrates clearly that there is a tremendous potential for savings in the Dutch Annual Construction Survey. The novel point in his study is the use of statistical confidence intervals to determine whether further editing would significantly change the estimates - a long overdue development. Latouche et al. demonstrate one further important point to note: in implementing selective editing methods, the whole survey process must be reconsidered in order to optimize the timing of the various steps, and in order to eliminate duplication of effort.

Most applications of selective editing recognize that fatal errors must be reconciled, but that such reconciliation must be done as expediently and as quickly as possible. To this end, automatic imputation systems, particularly generalized, re-usable software, can be of great assistance, as evidenced by Latouche et al.

The general effectiveness of imputation in producing good first order estimates (e.g. means and totals) is demonstrated in the Verboon and Schulte-Nordholt paper. They also underscore the usefulness of donor imputation methods, which make it easier for the imputer to ensure that the imputed values are valid and plausible. In order to improve the donor methods, Bankier et al. describe a search algorithm which takes into account the donor availability. The algorithm's second virtue is its ability to deal with qualitative and quantitative data simultaneously.

Finally, Verboon and Schulte-Nordholt also provide a word of caution concerning second order estimates (e.g. variances and correlations) derived from imputed data. Most imputation methods add a certain

amount of noise (i.e. variability) to the estimates. The decomposition of the total variance shows clearly the role played by the variance due to imputation component. Numerous studies exist showing the effect of ignoring the latter variance component and providing corrections to the usual variance estimator. What must be stressed here is that it is the total variance of the estimates that should be of interest. In other words, once imputation was performed, and estimates using the imputed data were produced, it is the variance of these estimates that is of interest, not the variance of estimates that would have been obtained had there been no non-response, and hence no need for imputation!

So what to do when an edit fails? The underlying message of this chapter is that while editing can serve a useful purpose in tidying up some of the data, its much more useful role derives from its ability to provide information about the survey process. Editing must be considered as an integral part of the data collection process by gathering intelligence about the process. In this role, editing can be invaluable in sharpening definitions, improving the survey vehicle, evaluating the quality of the data, identifying non-sampling error sources, serving as the basis of future improvement of the whole survey process, and in feeding the continuous learning cycle. The reliance on editing to fix problems after the fact must diminish significantly. While some editing is essential, its scope must be reduced, and its purpose redirected. Learning from the editing process must gain in prominence.

IMPACT OF EDITING ON THE SALARY STATISTICS FOR EMPLOYEES IN COUNTY COUNCIL

by Kajsa Lindell, Statistics Sweden

ABSTRACT

The purpose of this evaluation is to give information on the impact of editing for variables and domains of interest, by calculating the difference between the edited and the unedited data for each individual. These differences are then analyzed in the form of diagrams and tables in order to find out whether the editing process can be rationalized by excluding unnecessary editing. The results can form a basis from which measures can be taken to improve the editing and survey design, which can lead to a considerable rationalization of the production process.

Keywords: adjustments; corrections; editing measure; editing process; evaluation.

1. INTRODUCTION

Editing is considered to be an important item in the production of statistics, but it often takes a lot of time and is very often rather expensive. The purpose of editing is to minimize the number of errors in statistical data in order to improve the quality of the estimates. In general, little is known what effect it has on the survey results. Studies of the editing process that has been done at Statistics Sweden and elsewhere, indicate that the adjustments of data through editing do not have much of an effect on the quality of the estimates. This implies that greater resources are spent on the editing process than really are necessary. If instead it would be possible to limit the editing to find and adjust the errors that have an effect on the estimates, the same quality might be obtained, but with less resources.

This evaluation of the editing process in the survey on salary statistics for employees in county councils in 1992 is done according to [1]. The purpose of the evaluation is to give information on the **impact of editing on variables and domains of interest**. The method consists of **comparisons between edited and unedited data**. For each individual the difference between the edited and the unedited data is calculated. These differences are analyzed in form of diagrams and tables for each item and domain in order to find out whether the editing process can be rationalized by excluding *unnecessary* editing. This kind of evaluation does not make it possible to distinguish between *adjustments* (i.e. changing the faulty value into another value, which may be more accurate, less accurate, or even correct) and *corrections* (i.e. changing a faulty value into a correct, "true", value). In order to do this re-interview studies are necessary.

The results of this evaluation can form a basis for taking measures to improve the editing and the survey

design, which can lead to a considerable rationalization of the production process.

2. BRIEF DESCRIPTION OF THE PRODUCTION OF SALARY STATISTICS FOR EMPLOYEES IN COUNTY COUNCILS

The survey of employees in county councils is carried out in co-operation with the Swedish Federation of County Councils, Statistics Sweden and the trade union organizations concerned, viz. the Swedish Municipal Workers' Union (SKAF), the Swedish Confederation of Professional Associations (SACO) and the Swedish Central Organization of Salaried Employees (TCO).

The main variable is the average monthly salary. It is defined as the monthly salary before any deductions have been made and where all fixed and variable increments are included. For part-time employees the monthly salary is calculated in the following way.

$$Monthly\ salary = \frac{\sum_{i=1}^{N} L_i}{\sum_{i=1}^{N} Q_i}$$

where

L_i = actual salary for individual *I, I = 1, ... , N*

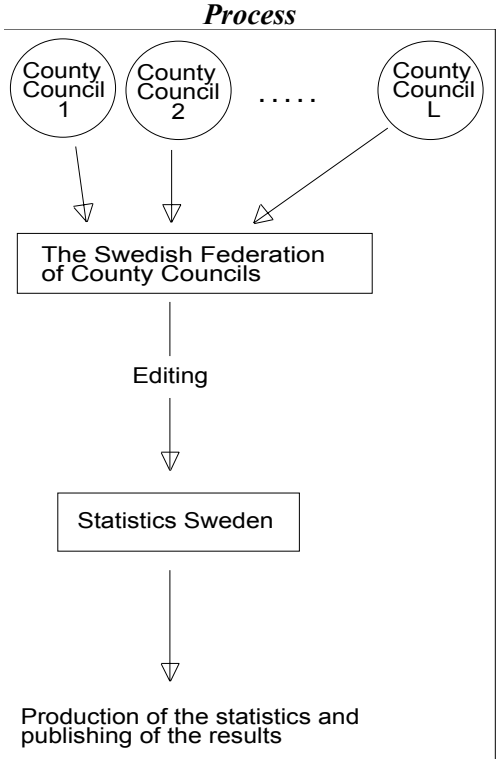
Q_i = extent of employment, expressed as percentages of full-time for individual *I*

The salary is reported in different ways, either according to a grade, which corresponds to a certain amount of salary, or precisely the actual salary.

The survey population consists of all employees in county councils between 18 and 65 years of age. The survey is an enterprise census and is based on individual data for every employee.

The editing is done by the Swedish Federation of County Councils. The file with unedited data therefore consisted of the data that were delivered to the Swedish Federation of County Councils from the different county councils all over Sweden, and the edited data file is the file that Statistics Sweden received from the Swedish Federation of County Councils.

Fig. 1 - A General Outline of the Production Process



3. EVALUATION OF THE EDITING PROCESS

There were 226 702 employees included in this evaluation of the statistics in 1992. As can be seen in the Table 1, the number of employees differs somewhat when comparing the edited and the unedited data, whereas the average monthly salary is not affected that much. A great part of the editing work consists of removing duplicates from the files.

3.1 Selected Parameter

The parameter selected to be studied is average monthly salary including all the increments (e.g. bonus for inconvenient working hours, remuneration for on-duty time, stand-by remuneration etc.) for

- full-time employees
- part-time employees
- full-time and part-time employees

Table 1 - Overall Comparison of Edited and Unedited Data. Number of Employees, Average Monthly Salary in Swedish Kroner (SEK)

	Men		Women	
	Number	Average monthly salary	Number	Average monthly salary
Full-time employees				
Edited value	28 967	17 184	77 576	13 541
Unedited value	28 746	17 228	76 910	13 538
Edited value - Unedited value	221	-44	666	3
Full-time and part-time employees				
Edited value	33 729	16 795	132 147	13 502
Unedited value	33 522	16 801	131 453	13 502
Edited value - Unedited value	207	-6	694	0

3.2 Selected Domains

The domains were chosen according to two classifications.

a) *Major domains,* which cover the whole population:

- Employees belonging to SKAF
- Employees belonging to SACO or TCO
- Total (1+2)

b) *Minor domains*, which cover a subset of the population:

- County council of Stockholm
- County council of Blekinge
- County council of Norrbotten
- County council of Malmöhus
- County council of Östergötland
- Health and medical services
- Assistant nurses
- Doctors

The minor domains *health and medical services* (which is the greatest field of activity among the county councils), *assistant nurses* and *doctors* are intended to illustrate the need for editing of the official salary statistics as well as of the commissions of trust.

3.3 Mode of Procedure

The file with unedited data was matched with file with the edited data, so that each row consisted of information on one individual. Since it is necessary to have an unedited and an edited value for each individual, only employees included in both files could be matched. The results of the analysis are presented in Table 2.

In the editing process the salary was adjusted for 16.2 per cent of all employees. The corresponding value for full-time employees was 15.7 per cent and for part-time employees 16.8 per cent. For most (99 per cent) of the employees who have had their salaries adjusted, the salary has been adjusted with more than 10 SEK (Swedish Kroner).

3.4 Definition of Editing Measure 1

The difference between the edited and the unedited data is calculated for each individual for all the *N* employees that can be found in both the unedited and the edited data. By summing the greatest absolute differences and dividing the sum by the total absolute difference it is possible to get information for example whether a small number of editing changes contribute to a great share of the total editing change. If it is possible to identify these individuals in advance there is a considerable potential of rationalizing the editing process.

Table 2 - Employees by Domain. Number of Employees, per cent

	Number of employees		Percentage of
	total	whose salary has been adjusted	employees whose salary has been adjusted (%)
	(N)	*(M)*	*(100*M/N)*
Major domains			
Full-time and part-time employees			
Employees belonging to SKAF	161 875	27 545	17.0
Employees belonging to SACO or TCO	64 827	9 258	14.3
Total	226 702	36 803	16.2
Full-time employees			
Employees belonging to SKAF	74 693	13 821	18.5
Employees belonging to SACO or TCO	43 671	4 808	11.0
Total	118 364	18 629	15.7
Part-time employees			
Employees belonging to SKAF	87 182	13 724	15.7
Employees belonging to SACO or TCO	21 156	4 450	21.0
Total	108 338	18 174	16.8
Minor domains			
Full-time and part-time employees			
County council of Stockholm	35 004	3 783	10.8
County council of Blekinge	4 241	255	6.0
County council of Norrbotten	7 176	559	7.8
County council of Malmöhus	21 395	10 878	50.8
County council of Östergötland	12 088	1 108	9.2
Health and medical services	177 334	28 660	16.2
Assistant nurses	44 912	8 292	18.5
Doctors	8 598	1 074	12.5

The absolute differences are

$$|d_i| = |y_{ig} - y_{iog}|$$

where

y_{ig} = edited value for variable *y*, individual *I*

y_{iog} = unedited value for variable *y*, individual *I*, $I = 1,..., N$

The total absolute difference is

$$D = \sum_{i=1}^{N} |d_i| = \sum_{i=1}^{M} |d_i|$$

since all $d_i = 0$, except for the *M* adjusted values.

The *M* differences are ordered in descending magnitude, that is

$$|d_1| \geq |d_2| \geq \ldots \geq |d_M|$$

The total absolute difference *D* is the total adjustment and d_I / D represents the share of the *i*:th biggest adjustment to the total adjustment. The share of *D* that consists of the *j* biggest adjustments can be written as

$$\sum_{i=1}^{j} d_i / D$$

Accordingly, the percentage of the differences put in relation to the total difference is

$$q_j = \left(\sum_{i=1}^{j} |d_i| / D \right) \cdot 100\% \quad \text{for } j = 1,..., M$$

This measure is plotted in a diagram against

$$p_j = j / M \cdot 100\% \quad \text{for } j = 1,..., M$$

Where p_j= percentage of j individuals with j biggest adjustments.

The following diagram indicates the share of the total adjustment received as a result of a given percentage of the greatest adjustments. Such diagrams have been produced for all the major and minor domains.

In Figure 2 it can be seen that 90 per cent of the total adjustment can be carried out by eliminating about 20 per cent of the greatest errors in the editing process. That equals about 51 adjustments, which can be compared to the 255 adjustments that have been made for the county council of Blekinge.

3.5 Definition of Editing Measure 2

Another way of illustrating the effects of editing is to relate the editing changes to the estimate being published. The purpose is to find out whether some changes are insignificant compared with the estimate. In this case it may be possible to refrain from editing for these individuals without affecting the quality of the estimate. The precision demand and the effect of other error sources constitute a basis for what should be considered insignificant in this connection.

The difference is

$$d_i = y_{ig} - y_{iog}$$

where

y_{ig} = edited value for variable y, individual I

y_{iog} = unedited value for variable y, individual I, $I = 1,..., N$

and

$$D = \sum_{i=1}^{N} d_i = \sum_{i=1}^{M} d_i$$

since all $d_i = 0$, except for the M adjusted values.

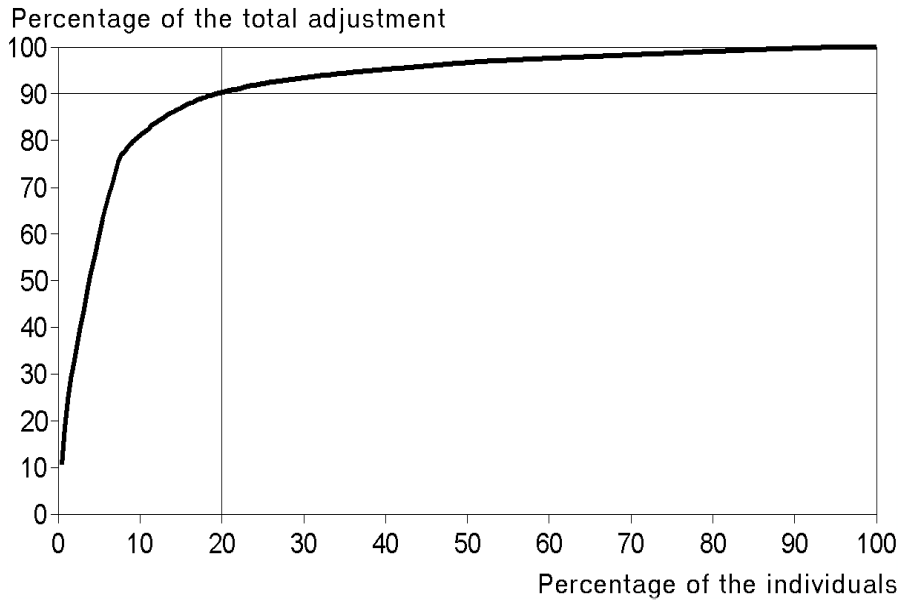
The average monthly salary, based on edited data, is

$$\bar{y}_g = \frac{\sum_{i=1}^{N} y_{ig}}{N}$$

and the corresponding value, based on unedited data, is

$$\bar{y}_{og} = \frac{\sum_{i=1}^{N} y_{iog}}{N}$$

Figure 2 : Editing Measure 1 for the County Council of Blekinge



The total measure of error is defined as the ratio between the estimate based on unedited data and the estimate based on edited data, that is:

$$R = \frac{\bar{y}_{og}}{\bar{y}_g} = \frac{\sum_{i=1}^{N} y_{iog}}{\sum_{i=1}^{N} y_{ig}} = \frac{\sum_{i=1}^{N} y_{iog}}{\sum_{i=1}^{N} y_{iog} - \sum_{i=1}^{M} d_i}$$

The M differences d_I are ordered in descending absolute magnitude as for editing measure 1, but this editing measure takes regard to the sign of the change. Thus the editing measure is defined:

$$r_j = \frac{\sum_{i}^{j} d_i \sum_{i}^{N} y_{iog}}{\sum_{i}^{M} d_i \sum_{i}^{N} y_{iog}} \cdot 100\,\% = \frac{\sum_{i}^{j} d_i / N \, \bar{y}_{og}}{\bar{y}_g} \cdot 100\,\%$$

It illustrates how the adjustments (in descending magnitude) affect the estimate $\bar{y}_g$.

This measure is plotted in a diagram against

$$p_j = j / M \cdot 100\% \qquad \text{for } j = 1,..., M$$

Figure 3 shows how the graph approaches the estimate after the j biggest differences have been adjusted (instead of all the differences). r_o shows the impact of the editing on the estimate, and $r_j = 100$ indicates that the following changes have no impact on the estimate. In contrast to editing measure 1, regard is taken to the sign of the difference. This is the reason why the diagram looks oscillating when it approaches the final estimate.

Diagrams have been produced for all the major and minor domains. In the diagram below it can be seen that the graph is almost on a level with the estimate before any adjustments have been made. This implies that there is a considerable potential of rationalizing the editing process.

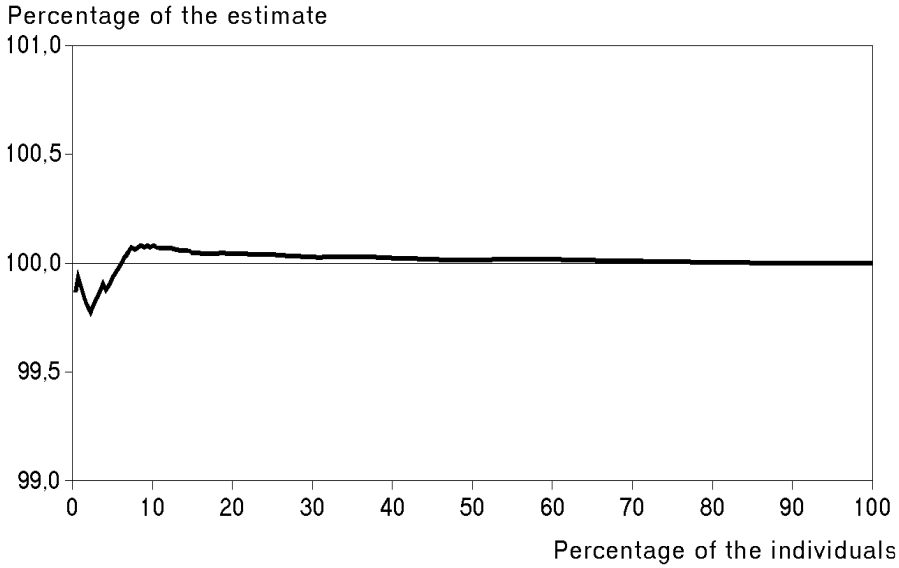
REFERENCES

[1] Forsman, Gösta. *Guide for Evaluation of the Editing Process*, Editing Memo 24, 1991, (in Swedish).

[2] Garås, Tomas. *Evaluation of the Editing Process in the Annual Statistics of Government Employees*, Statistics Sweden, Bakgrundsfakta till arbetsmarknads- och utbildningsstatistiken, 1993, (in Swedish).

[3] Wahlström, Charlotte. *The Effects of Editing: A Study of The Financial Statistics at Statistics Sweden*, Statistics Sweden, F-metod 27, 1990, (in Swedish).

Figure 3 : Editing Measure 2 for the County Council of Blekinge



SELECTIVE EDITING IN THE NETHERLANDS ANNUAL CONSTRUCTION SURVEY

Frank van de Pol, Statistics Netherlands

ABSTRACT

For the Annual Construction Survey some key results under selective editing are compared with the same results under extensive editing. For two key variables the applied type of selective editing turns out to affect estimates plus or minus +%, compared to extensive editing. This is a small difference in comparison to the confidence interval, which is 3% for this survey. Plots of cumulative edit effects are used to fine-tune the score variable, which summarizes error messages and the firm's importance for population estimates.

Keywords: selective editing; symmetric distribution; graphical editing.

1. INTRODUCTION

Statistics Netherlands has Annual Manufacturing Surveys for all fields of economic activity. One of these is the Annual Construction Survey. For this survey a sample of 6000 contractors, plasterers, painters, glaziers and decorators is drawn from the population of 18 000 firms. About 4000 firms return a completed form, a response of 65%. There are several forms, a concise one with 55 questions for small firms and more extensive ones with up to 120 questions for larger firms.

Data entry, data editing and the production of tables takes about 8 man-years for the Annual Construction Survey (ACS). Most of this time, about 5 man-years, is spent on data editing. This medium-sized survey was chosen for experiments with selective editing, not only because relatively much data editing is being done here, but also because the ACS team is a very cooperative one, with a positive attitude toward change. Furthermore, a spin-off is expected that will benefit the other Annual Manufacturing Surveys which are being held by Statistics Netherlands.

Data entry and data editing of the paper forms is done with a Blaise data entry machine (Bethlehem et al.,1989-1993), which runs on 6 network-PC's. Here Blaise applies 96 on-line edits to each form when it is entered. Entries may be dirty, suspect or acceptable. (We will not call acceptable entries 'clean', because this would suggest absence of inconsistencies, which cannot be guaranteed.) When an entry is dirty or suspect the subject specialist may alter the entry, but the status can be altered also without changing the entry. Changes are sometimes made right away at data entry ('heads up'), but sometimes also at a later stage.

When all entries of a contractor's form are acceptable the complete record will have Blaise status 'acceptable'. Otherwise the record is still 'dirty' or 'suspect'. Editing could be limited to those records that are dirty or suspect. However, subject specialists rightfully claim that even in acceptable records errors can be found, especially when comparisons are made with data from previous years. Therefore they still follow the long-standing instruction that each record should be checked most carefully. For this checking not only the Blaise machine is used, but always also the paper forms from previous years and, if necessary, inquiries by telephone. After this thorough investigation the subject specialist declares a record 'clean'.

Looking at these exhausting data editing activities one wonders whether approximately the same outcomes could be obtained with less efforts. Some errors might be too small to be worth correcting, other errors occur with firms too small to contribute noticeably to the overall mean. In the present paper we start investigating what the consequences of selective editing are for ACS publications. We will look at two central variables, production and labour costs. If the consequences of selective editing are small, the present Blaise approach could be used more efficiently, skipping the traditional way of integral editing.

2. DESCRIPTION OF APPROACH

We want to assess to what extent we can expect a discontinuity in our time series if selective editing will be applied. Therefore we will list the difference between averages over firms estimated with several

Table 1. Error message ratios for production and labour costs

v	error message ratio ($r_v=y_v/x_v$)	lower bound	upper bound
1.	production / man-days	dfl.250	dfl.1500
2.	labour costs / man-days	dfl.150	dfl. 500
3.	man-days / # employees	160 days	230 days
4.	sum of sub-costs / total costs	1	1
5.	sum of sub-benefits / total benefits	1	1
6.	total costs / total benefits	1	1

editing schemes:[1]

a) don't edit;
b) edit only when the score variable is too large;
c) edit only when the present Blaise error messages ask for it;
d) complete editing.

Moreover, we will report how many records should be edited with each approach.

Complete editing is the present option. It will be used as a bench-mark. It might be more efficient to edit records only when the present Blaise error messages ask for it. The present Blaise machine is based on 96 error messages. From these we selected 6 error message variables which seem most relevant for our target variables, production and labour costs. The error message variables are ratios which have comparable values for small firms and large firms. An overview of these error message ratios is given in table 1. Similar edits are used in the US Annual Survey of Manufactures [5].

In the present Blaise-system a violation of the boundaries for one or more of these ratios results in an error message (or warning). Ratio 1 is obtained by dividing total production by total number of payed man-days. Ratio 2 is obtained by dividing total labour costs by total number of payed man-days. These ratios would not be a good check of production and labour costs if man-days is erroneous. Therefore also ratio 3, payed man-days divided by number of employees, is evaluated. Ratios 4, 5 and 6 are tests on good book-keeping. The sum of cost-entries should be equal to the total costs, which are also asked for (ratio 4). The same holds true for benefits (ratio 5). Finally costs should balance benefits (ratio 6).

[1] To avoid confusion between totals as sums over variables and totals as sums over firms, we will present our results as averages over firms, reserving the term total for sums of variables.

Another option would be to fine-tune these error messages to be less sensitive for small firms. The general idea of selective editing is that an error in the ratio r_{vf} of a big firm (1,000,000 man-days a year, 500 employees) is much more important than the same error in the r_{vf} of a small firm f' (10,000 man-days a year, 5 employees).

Inspired by Hidiroglou and Berthelot [2] a 'score variable' is defined to prioritize edits. We will first consider the case of only one error message ratio, $r_{vf}=y_{vf}/x_{vf}$, with v the number of this ratio, y_{vf} its numerator and x_{vf} its denominator. Ideally, in case of a random sample we would like to prioritize edits according to the size of error in y_{vf}, to be called $e_{vf}=y_{vf}-Y_{vf}$.[2] After all, the main objective of the ACS is to estimate sums $\Sigma^N_f Y_{vf}$ over all N firms in the population, such as the production of all construction firms together. In the ACS sample, however, smaller firms have a lower probability, π_f, to be included in the sample. A crude estimate of $\Sigma^N_f Y_{vf}$, ignoring selective nonresponse, is then $\Sigma^N_f Y_{vf}/\pi_f$.[3] Therefore, if we knew the size of the error, e_{vf}, we should prioritize edits according to

$$s^*_{vf} = e_{vf}/\pi_f \qquad (1)$$

However, we do not observe error e_{vf}, so we should devise some model to estimate it. This model has to be formulated after standardization to figures which are comparable for all firms, i.e. the model has to be formulated in terms of one of the ratios in table 1. This means we will predict error e_{vf}/x_{vf} in ratio r_{vf}, instead of the error e_{vf} which we are actually interested in. Since we deal with unedited data we should use robust methods [6]. Let us introduce a 'deviance' as a fallible indicator of error in ratio r_{vf}. Deviance d_{vf} is defined as the absolute difference between r_{vf} and the median ratio,

$$d_{vf} = \mathrm{abs}[r_{vf}-\mathrm{median}_f(r_{vf})] \; . \qquad (2)$$

In case of a high deviance d_{vf} an error in the ratio is likely to have occurred, with a size in the same order of magnitude as the deviance. If denominator x_{vf} is correct, then error e_{vf} is indicated by $d_{vf}x_{vf}$. Taking the

ACS design into account, our score variable to prioritize edits for variable y_v is

$$s_{vf} = d_{vf}x_{vf}/\pi_f \quad (3)$$

Both for ratios 1 and 2, production per man-day and labour costs per man-day, the denominator is the number of man-days, $x_{1f}=x_{2f}$. The score variable for these ratios can only be used validly to trace errors if this number of man-days is correct. To test this assumption, ratio 3 relates the number of man-days to the number of employees, $y_{3f}=x_{1f}=x_{2f}$ In the same way ratio 6 tests the denominator of ratio 4. A seventh ratio, $r_7=1/r_6$, was added to do the same with the denominator of ratio 5.

It is more time-saving to skip complete clean records than to skip clean fields only. Therefore we combine all six score variables, $s_{1f}, s_{2f}, \ldots s_{6f}$, into one single score variable. In order to give each score variable s_{vf} about the same weight, it will be standardized, .

$$s^*_{vf} = (s_{vf}-s_{vf})/\text{stdev}_f(s_{vf}) \quad (4)$$

(Standardization with the median absolute deviation (mad) could not be applied since score variables 4, 5 and 6 have mad=0.) When some ratios are deemed more important than others one could make a weighted average score variable,

$$s_f = (1/\Sigma_v w_v)\, \Sigma_v\, w_v\, s^*_{vf} \quad (5)$$

For the present data, however, results are satisfying with weight w_v =1 for all v.

We will make comparisons between unedited data and completely edited ('clean') data. However, when the project started, part of the data had been cleaned already; the set of unedited records had become diluted with an over-representation of clean records. To obtain test-data comparable to unedited data we discarded 75% of the clean records, leaving a sample of 1492 records for which two values of all variables are available: before and after editing.

3. RESULTS

With the 6 tight error messages given in table 1, a large part, 65% of the records has to be edited. For the 573 firms in our sample with less than 10 employees this is 60%, and for the 919 larger firms 68%. With all 96 error messages operational even more records will have to be edited. Therefore we should try to be more selective, and use a score variable.

Before this sort of selective editing can be applied one has to choose a critical value for the score variable, below which no editing will take place. In order to do this we want to see the cumulative effect of various amounts of editing in a graph. We will graph estimates of the average,

$$\Sigma^n_f\, [\, [\delta_f Y_{vf}+(1-\delta_f)y_{vf}]\, (\Sigma_f\, \pi_f\,) / \pi_f\,], \quad (6)$$

for decreasing critical values of the score variable s. For a given critical value $s_{.|}$ all firms with $s_{.f}>s_{.|}$ (δ_f / 1) have y_{vf}, which is suspect, replaced with hand-edited value Y_{vf}, whereas for firms with $s_{.f}\#s_{.|}$ (δ_f / 0) the unedited value y_{vf} is used[2]. Figure 1 shows such graphs for the average production and the average labour costs.

To obtain a symmetric distribution the natural logarithm of the score variable was taken instead of the score variable itself. (A value of .25 had to be added to handle negative values of the score variable.) The vertical size of the graphs is about a 3% interval around the average, which also is the confidence interval for ACS-statistics. It shows that most edits have tiny effects on the estimate of the average; most vertical moves of the line are very small. Edits with a sizeable effect, a steep drop or rise, occur with firms that have a large score variable (in the left part of a graph).

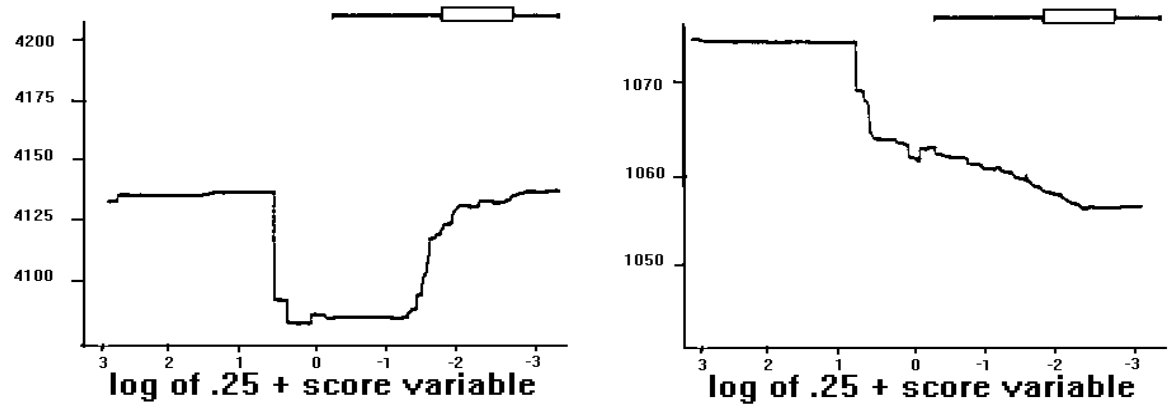
The boxes on top of the graphs mark the region that holds 50% of the edits. Left of these boxes are the 25% high score variable edits that cause relatively large changes in the estimates of the average [3].

We choose the critical value below which no editing should take place to be $\ln(.25+s_{.|})$. -1.6; $s_{.|}$=-.05. With this critical value our estimates will be very close to what we would get from complete editing, and yet a majority of the records can pass without being edited.

[2] Capital Y_{vf} is used for the clean, consistent value and y_{vf} for the unedited, possibly false, value.

[3] The lines beyond the boxes, called the whiskers, are at most 1.5 times as long as the boxes. In our graphs some cases lie left of the whiskers range.

Figure 1: Average estimates of production (left) and labour costs (right) with 0 percent editing on the left to 100 percent editing on the right, ordered by the score variable



Let us now look more closely into the effects of the four editing strategies on the estimated average. Table 2 gives average values of production and labour costs for all edit options. Averages are computed with weights $1/\pi_f$.

A stunning result is in the leftmost column in table 2: (almost) no editing gives reasonable estimates. This notwithstanding that this column was computed with most records (65%) violating one or more of the 6 error messages of table 1, i.e. with ratios often 0 or far too large. Errors seem to cancel out to some extent. It should be noted, however, that data entry was done by subject specialists, who may have removed some of the extreme outliers (1000 times too large) before the leftmost column was computed. Moreover the left-hand side of figure 1 shows that average production is rather accidentally estimated so well without editing (only -0.12% compared to complete editing), since one big error happens to compensate the cumulative effect of all other errors. Average labour costs are estimated 2% higher without editing than with complete editing, an acceptably high deviation. In ACS publications a 3% interval (+/-1.5%) is recommended as confidence interval containing the population average in 95% of the samples.

A second finding is that both types of selective editing give reasonable results. The use of a score variable seems preferable since it takes only 25% of the complete editing effort, whereas editing all records that violate any error message takes more than twice as much editing, 67% compared to complete editing.

The price that has to be paid for the higher efficiency of the score variable method, is a slight deviation of minus -.39% or +.46% from the complete editing average. This will not cause a disruption of our time series. Moreover our benchmark, the completely edited data, may be slightly in error as well.

Table 2. Average values of production and labour costs (in thousand guilders)

variable	1. (almost) no editing	2. edit only if score variable > -.05	3. edit if one of 6 error messages applies	4. complete editing
production	4133 (-.12%)	4120 (-.41%)	4136 (-.04%)	4138
labour costs	1077 (1.94%)	1060 (+.30%)	1058 (+.17%)	1057
percentage edited	0 (?)	26	65	100

Disclaimer: Due to sample selection, results are not representative for the population of firms.

4. DISCUSSION AND PROSPECTS

Our score variable is somewhat different from the one proposed by Hidiroglou and Berthelot [2], although the general idea is the same. We did not copy their deviation (s_i). It seems symmetrical in y and x, but it is not, as is shown by Höglund Davila [3]. In fact, we are not looking for a symmetrical measure of deviation but for an estimate of the error in a target statistic, like average production. Another difference is that H&B devised a method for one target variable, whereas the ACS has many target variables. Therefore we combined several field-specific score variables into a record-specific one. Moreover we skipped their parameter U, which enables more editing of small firms than is necessary from an efficiency viewpoint. Furthermore we propose a graphical method to determine the critical value of the score variable, below which no editing has to take place. This graphical method is especially useful if both unedited data and completely edited data are available. In other cases the Sande criterion suggested by H&B may be preferable. Finally we did not copy the imputation part of Hidiroglou and Berthelot [2].

Our work on selective editing with the ACS is not finished yet. Not only production and labour costs should be analysed, but many more variables. Also more work on the score variable is to be done. Firstly, the critical value of the score variable should be made dependent on the size of the publication cell concerned. Firms from densely populated cells should be edited less heavily than firms from sparsely populated cells. Secondly, we should analyse whether adding more error messages to the score variable improves the performance of selective editing. Thirdly, there should be a way to estimate optimal weights w_v in equation 5.

The ACS has a fairly large nonresponse on some of the more detailed questions. Editing on these questions involves many telephone calls, especially with small firms. This effort will be reduced after implementation of some imputation scheme for the less important firms.

Beside these ACS-specific activities we intend to look into the possibilities of purely graphical editing (using SAS-Insight, [7]), of the Fellegi-Holt approach (using Lince; [4] and of statistical editing [5].

NOTES

1) To avoid confusion between totals as sums over variables and totals as sums over firms, we will present our results as averages over firms, reserving the term total for sums of variables.

2) Capital Y_{vf} is used for the clean, consistent value and y_{vf} for the unedited, possibly false, value.

3) We have the population ordered such that the first n cases are the sampled ones.

4) The lines beyond the boxes, called the whiskers, are at most 1.5 times as long as the boxes. In our graphs some cases lie left of the whiskers range.

REFERENCES

[1] Bethlehem, J. G., A. J. Hundepool, M. H. Schuerhoff and L. F. M. Vermeulen. Blaise version 2 manuals, Automation Department, Statistics Netherlands, 1989-1993.

[2] Hidiroglou, M. A., and J.-M. Berthelot. Statistical editing and imputation for periodic business surveys. Survey Methodology 12, 1986, pp. 73-83.

[3] Höglund Davila, E. A report on a study on the Hidiroglou-Berthelot Method (Statistical Edits) applied to the Swedish Survey of the Delivery and Orderbook Situation in the Swedish Industry, Statistics Sweden, 1989.

[4] Informatica Comunidad de Madrid SA, Lince, Sistema de validación e imputatión automatica de datos estadisticos; manual de usuario, ICM, Madrid, 1993.

[5] Little, R. J. A., and Ph. J. Smith. Editing and imputation for quantitative survey data, Journal of the American Statistical Association 82, 1987, pp. 58-68.

[6] Rousseeuw, P. J., and A. M. Leroy. Robust regression and outlier detection, Wiley, New York, 1987.

[7] SAS, SAS/INSIGHT user's guide, version 6, 2nd ed., SAS Institute, Cary NC, USA, 1993.

DEVELOPMENT OF A COST-EFFECTIVE EDIT AND FOLLOWUP PROCESS: THE CANADIAN SURVEY OF EMPLOYMENT EXPERIENCE

By Michel Latouche, Marcel Bureau and James Croal, Statistics Canada

ABSTRACT

The editing and follow-up strategy currently used in the Survey of Employment Payrolls and Hours (SEPH) is very time consuming, burdensome and expensive. A single questionnaire can be queried up to three times during the production cycle and every rejected questionnaire is followed up with the respondent. In order to improve the situation, SEPH will soon implement a new strategy that is based on a centralization of editing and follow-up approach. This paper focuses on the methods used to develop a cost effective edit and follow-up process through identification of units with large impact on the estimates. A simulation study compares different alternatives for a score function that evaluates the impact on survey estimates of records that contain doubtful responses and determines the units to be followed up. Results confirmed that the new edits in conjunction with the selected score function tends to identify the largest units that have a high impact on the estimates as needing follow-up. The paper ends with some recommendations.

Keywords: score function; simulation study.

1. INTRODUCTION

The Canadian monthly Survey of Employment, Payroll and Hours (SEPH) collects data on payroll employment, weekly earnings, and weekly paid hours. The primary objectives are twofold:

(I) To provide monthly estimates of the total number of paid employees, average weekly earnings, average weekly hours and other related variables at the industry division by province level.

(ii) To provide these estimates for Canada at the three digit Standard Industrial Classification (SIC) level.

Sample design

SEPH covers all industries except agriculture, fishing and trapping, private household services, religious organizations, and military services. It is designed as a stratified sample of establishments with stratification by industry, province or territory, and employment size group. In the new design implemented in March 1994, there are 413 industry-province cross-classifications and 4 size groups for a total of 1652 strata. All establishments, regardless of size, belonging to enterprises having 300 employees or more are in a take-all stratum. The sample is made up of about 30 000 take-all establishments and 10 000 take-some establishments. The survey is complemented by administrative data [2]. However, at the time this study was performed, no administrative data were used and the sample was made up of 35 000 take-all establishments and 22 000 take-some establishments selected from a population of 800 000 units [6].

Questionnaire

The questionnaire is two pages long and has to be filled almost exclusively with numeric data such as dates, number of employees, number of hours worked and amounts. The first page of the questionnaire contains information about the last pay period for the reference month. This information is asked by category of employees (paid by the hour, salaried or other) and for each pay period type used by the establishment. The pay period types are either weekly, every two weeks, semi-monthly, monthly, every 4 weeks or other. The second page of the questionnaire covers special payments, gross payments and periods of absence for the whole reference month. Depending on the number of pay period types used, the respondents have to provide from 20 to 76 answers.

Questionnaires relating to government agencies are the responsibility of the Public Institutions Division of Statistics Canada while all other questionnaires are the responsibility of the Labour Division. This study was based on the latter. Most of these respondents (86%) choose to report their data by mail; 14% use the telephone, while a few send computer reports. The initial mail response rate is around 60%. Most of the non respondent recontacts and follow-ups for error corrections are done by phone. However, answers are written on questionnaires and then the data are captured and edited a second time. SEPH does not yet

use Computer Assisted Telephone Interviewing.

Data editing

There are six major groups of micro edits (i.e., at the establishment level) in SEPH: basic, primary, relational, month-to-month, consistency and special.

The basic edits identify if a record should be processed or dropped. The reason for dropping a record could be: invalid identification number, invalid survey month etc.

The primary edits validate field sizes and check for missing data or invalid data.

The relational edits identify invalid reporting structures or patterns and are performed for each combination of category of employees and pay period type.

The month-to-month edits validate changes in the variables which occur from one month to another.

The consistency edits are used for units which are in the sample for the first time, since there is no data from the previous month for them.

The special edits validate various items like operating days, overtime hours and earnings, hours in the standard work week, special payments, gross taxable earnings etc.

The micro editing is performed in two steps. First, the questionnaires are captured and screened for completeness by the interviewers in the regional offices. Basic, primary and relational edits are performed in the regions. Any rejected questionnaires are followed up with the respondent, while the others are transmitted to head office where the remaining, more sophisticated, edits are applied. At head office also, any rejected record is followed up with the respondent. Once corrected or confirmed, the information is sent to the Labour Division where some macro editing is performed before the publication is released. Here again, follow-ups may be done by subject matter analysts.

The SEPH editing as well as the follow-up strategy are very time consuming, burdensome and expensive. A single questionnaire can be queried up to three different times by three different individuals. Although the strategy should ensure very good quality of the collected data since all suspicious data are followed up, it attempts to"kill a fly with a gun". This approach is very costly and inefficient. In fact, the strategy of a 100% follow-up is not often respected because of time constraints, and some follow-ups are replaced by manual imputation of the data or the edits are ignored as if the respondent confirmed the data. Moreover, respondents feel very frustrated when they are contacted on several occasions for reasons they think to be of little importance. This has a negative impact on response rates and on Statistics Canada's image.

New editing strategy

In order to improve the situation, SEPH will soon implement a new strategy, that takes into account operational constraints, as described by GSFD [4], and Berthelot and Latouche [1] to redesign the data collection, capture and editing process. This strategy is mainly based on (a) centralization of editing and follow-up in the regions, (b) an improvement of edit rules to reduce the editing process and (c) a selective follow-up approach.

(a) The first improvement is to centralize the collection, capture, edit and follow-up process in the regional offices. Thus, all edits are performed before any follow-up is initiated. The respondent is recontacted for follow-up only once and all queries are solved at the same time.

(b) The second improvement, is to revise the edits rules; they have been reduced from 60 to 32 by eliminating inefficient and redundant rules. At the same time, new historical edits have been developed based on the Hidiroglou-Berthelot method [5]. The basic idea of this method is that, it is more important to detect a small change of a large respondent than a large change of a small respondent. Consequently, the edits bounds vary with the size of the respondents. The edit bounds are computed in order to reject (approximately) a fixed proportion of records. A high rejection rate (rigid bound) identifies most of the errors but also rejects many good records. On the contrary, a low rejection rate (tolerant bound) misses most of the errors but accepts most or all of the good records. To be efficient, the bounds must be set such that a large number of errors are detected, while most of the good records are accepted.

(c) As the third element of the redesign, a selective follow-up approach has been developed [1]. This strategy is twofold: first, a systematic recontact or follow-up effort is made for total non respondents to ensure that at least the status of the unit (active, inactive or out of scope) is obtained to perform an

appropriate imputation, if necessary, during a subsequent phase.

Secondly, the rest of the follow-up resources are concentrated on units that have a significant impact on the estimates. If only respondents with the largest impact are recontacted, the impact of the remaining errors will be minor because only marginal gains are realized by following up more respondents [3]. A score function that evaluates the impact on survey estimates of records that contain doubtful responses is used to determine the units with large impact that should be followed up. The remaining questionnaires with minor impact are made to satisfy the edit rules using an automated imputation system. However, a sample of these imputed records are also followed up to make sure that the imputation does not cause any bias.

This paper focuses on the methods used to develop a cost effective edit and follow-up process for SEPH through identification of units with large impact on the estimates. Section 2 provides more information about the trend based score function that is recommended and the reasons why it is preferred to other currently known functions. Section 3 presents two alternatives to that score function: a score function based only on the size of the respondent, and a tolerant edit rules approach. Sections 4 presents the methodology used and the results of a study comparing the 3 options. The paper ends with some recommendations.

2. SCORE FUNCTION

A score function is a mathematical formula that assigns a relative score to each respondent. As input parameters, the function uses those characteristics of the respondent that are related to the respondent's impact on the estimates. A score is calculated for each questionnaire variable and then summed to obtain a global score for the respondent. Respondents with the highest scores are considered to have the greatest impact on the estimates. The main criteria that guide the search for a score function are the size of the responding unit, the size and number of potential errors, as well as practical considerations. The score function that SEPH plans to use is based on the discrepancy between the current reported value and the final released value of the previous cycle.

Let $x_{i,k,t}$ be the value reported by respondent (i = 1,2,...I), for the variable (k = 1,2,...K) at time t;

$x^{)}_{i,k,t\&1}$ be the final value for respondent iI, for variable k at time t-1;

$w_{i,t}$ be the estimation weight (same for all variables);

P_k be the relative importance of variable k: a positive integer;

$Z_{i,k,t}$ be an error flag assigned to respondent i for variable k at time t. It takes the value of 1 when the variable fails one or more edits and 0 otherwise.

The DIFF score function is given by:

$$DIFF_{i,.,t} = \sum_{k=1}^{K} \frac{w_{i,t}\,(x_{i,k,t} - x^{)}_{i,k,t\&1})}{\hat{X}_{.,k,t\&1}} Z_{i,k,t} P_k$$

where $\hat{X}_{.,k,t\&1}$ is the total (at a given level of aggregation) for the variable from the previous reference month.

If $DIFF_{i,.,t}$ is equal to or greater than a pre-defined bound, then the respondent is followed up. For all other records any errors detected are corrected through imputation. This score function was successfully used in the Canadian Annual Retail Trade Survey [7].

Initially, while discussing the design of the new collection and capture system, it was thought that the score function would be computed at the time of capture rather than in batch mode outside the capture process. For such a scenario the DIFF score function has a major drawback. It involves complex programming requiring the system to access current and historical data files as well as parameters required for calculation of DIFF, with potentially long system response time. To avoid this, it was felt that it would be better to either use a simpler score function that does not need such information, or another editing strategy that would lead to similar results, i.e., reduced number of follow-ups without significantly affecting data quality.

3. ALTERNATIVES TO THE DIFF SCORE FUNCTION

PRE-score function

The first alternative to the DIFF score function is a score which can be computed prior to the collection process. For this reason that function is called a PRE score function and is expressed as follows:

$$PRE_{i,.,t} = \sum_{k=1}^{K} \frac{x)_{i,k,t-1} w_{i,t-1}}{\hat{X}_{.,k,t-1}} P_k$$

This function is simply the relative contribution of the respondent to the total of each variable at time t-1. The advantage of this function is its great simplicity. It is considered to be efficient enough in identifying respondents with large impact on estimates. The PRE score function is set to 0 when all variables pass the edits. As noted earlier in the paper, with the use of a score function, the suspicious questionnaires that are not followed up are automatically imputed.

Tolerant edit bounds

As a second alternative, a completely different approach is used. More tolerant edit bounds are used instead of a score function. In this way fewer questionnaires are rejected, thus reducing the amount of follow-ups, although all suspicious questionnaires which do not pass the edits are followed up. In this approach, imputation is required only for total non response and it is highly possible that non-negligible errors would not be detected unless macro editing is performed.

Although one wishes to use the simplest and most practical method to reduce the amount of follow-ups, that should not be at the risk of significantly lowering data quality. Further, it is desirable to decrease the number of follow-ups forced by macro editing. The following comparison study was done to evaluate the impact of these three proposed options on the data quality for SEPH.

4. COMPARISON STUDY

The purpose of the study is to evaluate the impact of the DIFF and PRE score functions and the tolerant edit bounds on major estimates for SEPH. The opportunity is also taken to use the Hidiroglou-Berthelot historical edits. The idea is to simulate the editing process on old raw data using the three options, and then to compare these estimates with the final released estimate.

The data used comes from the March 1991 cycle. At that time, the old design was still used and the sample was much larger than now [6]. About 20 000 questionnaires out of 65 000 were selected for the study. These were the respondents in 4 industries: Manufacturing Durable Goods, Manufacturing Non-Durable Goods, Wholesale Trade and Retail Trade. For the purpose of the study, records that were imputed for total non-response were removed, and because only historical edits are applied, only respondents who provided data for February and March were considered. A set of 8 edits on 15 variables was used in the simulation. (See appendix A for the detailed list of the edits considered.)

The edit bounds were first developed based on the variation of February and March 1991 final files, and then applied to the raw data file of March and the final data of February. (In actual practice the edit bounds will be developed based on data for the previous year. In this study that was not possible because of practical considerations.) It has to be mentioned that the raw data are not completely "raw" because the edits in the regions are already done and some corrections performed. However, the new month-to-month edits which are not applied in the regions could be considered and tested using this data. It should be noted that the month-to-month edits are not totally independent of the other edits and that fixing some other edits may prevent some of the month-to-month edits from failing.

For the DIFF and the PRE score function a set of rigid edit bounds that reject about 12% of final data was used. The relaxed bounds used in the third scenario with tolerant edit rules were adjusted so that a 6% rate of rejects of final data was achieved. The bounds were developed for every combination of 4 regions (East, Québec, Ontario, West and British Columbia) and the four industries considered except when the number of units was small. In those cases, the edit bounds were developed at the Canada by industry level. Collapsing of regions was necessary in order to ensure a minimum number of records to build the bounds.

To simulate the follow-up activity, the final data were used to replace the raw data. To simulate the partial imputation for scenarios 1 and 2, the current raw data were replaced by the previous final value from the same respondent. It should be noted that the imputation system in the current process is much more sophisticated than this since it uses historical trends and relationships between variables in addition to a variable-mean method for new respondents.

The estimates from the current process are compared to the estimates from each of the three

methods: use of the DIFF and PRE score functions and use of tolerant edit bounds. In fact the study compares an almost 100% follow-up scenario (the current process) to the three scenarios described above with (simulated) follow-up and partial imputation. The estimates are simple weighted totals of each of the variables at the cell level, i.e., province by subgroup of industry (two digit SIC) by Size (less than 200 employees and 200 employees or more).

5. RESULTS AND DISCUSSION

The comparison was mainly based on the following 4 aspects:

(i) Number of queries: queries are edit failures that lead to either follow-up or imputation. The strategy is to reduce the number of queries, while still detecting important errors.

(ii) Number of changes made by the current process, that are detected: these changes result from follow-up or manual correction. These indicate that a real error is detected as opposed to a confirmation of the original data. The number of these changes detected by the edit option is a good indication of its efficiency.

(iii) Number of macro edit failures detected: these edits are performed by subject matter analysts very late in the survey process. One would like these failures to be detected and corrected by the new micro editing and follow-up strategy.

(iv) The impact on estimates for each of the 15 base variables.

Number of queries

Table 1 below compares the number of records rejected as well as the number of follow-ups and errors detected between the current and the new proposed approaches. It can be noted that almost two thirds of the questionnaires are rejected in the current process and followed up. This is a good example of over-editing and justifies the current redesign.

The rigid bounds used in conjunction with either the DIFF and the PRE score function were adjusted to reject about 12-15% of final data records for every edit considered. When these bounds were applied to the raw data file the rate of rejects went up to 20-30% of records for every edit considered. Table 1 also shows the overall rejection rate at the record level (i.e., one or more edits failing for a given record) for the three options.

Table 1: Comparison between current and new processes

APPROACH	CURRENT PROCESS	NEW PROCESS	
ACTION (# of records)		RIGID BOUNDS FOR SCORE FUNCTION	TOLERANT BOUNDS
ACCEPTED	7 400 (37%)	8 561 (43%)	14 121 (71%)
REJECTED	12 436 (63%)	11 275 (57%)	5 715 (29%)
FOLLOWED-UP	12 436	2 254	5 715
ERRORS DETECTED	4 432	924	2 572

When using the rigid bounds, a total of 11 275 (57%) records failed at least one edit. This is comparable to the number of rejects in the current process. 90% of the records that were rejected in the current process were identified by the rigid edit bounds. The remaining records were low impact records that are accepted. As currently defined, these bounds produce very similar results compared to the current process.

One objective for re-computing a new set of bounds with the score function was to reduce the number of

interventions: either follow-ups or imputation. These bounds do not achieve this objective. Since the bounds are developed on each variable separately, it is difficult to predict what the end result of the combined failures will be in terms of number of records rejected. At least, the use of the score function will permit a reduction in the number of follow-ups identified.

Tolerant bounds were developed for the third scenario. These bounds reject 4% - 6% of each base variable. When the results of the edit failures of all variables were combined, 29% of the records were identified for follow-up, representing half the rejection rate with the rigid bounds.

For both score function strategies used in this simulation, all the rejected records are either followed up or imputed depending upon the priority assigned by the function. A rate of follow-up is arbitrarily chosen and the remaining records are identified for imputation. A 20% rate of follow-up was used for the simulation. The rate of follow-up can easily be changed (augmented or decreased), but since the total number of records identified by the edit bounds remains the same, the rate of imputation is directly related to the global number of queries identified. Unfortunately, this makes the final data too dependent on the imputation model used.

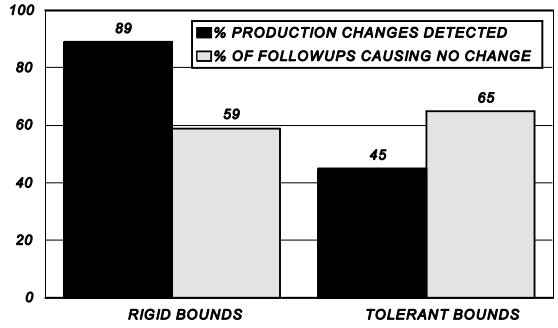
Number of changes made by the current process, that were detected.

In the current process, the 12 436 follow-ups lead to only 4 432 changes in the data. The main reason for this situation is the poor efficiency of the edit rules and the over-editing strategy that detects too many suspicious data items. Following up the respondents often only confirms that their data are correct. Another explanation is that, because of time constraints, it is impossible to perform all follow-ups and the data are accepted as they are.

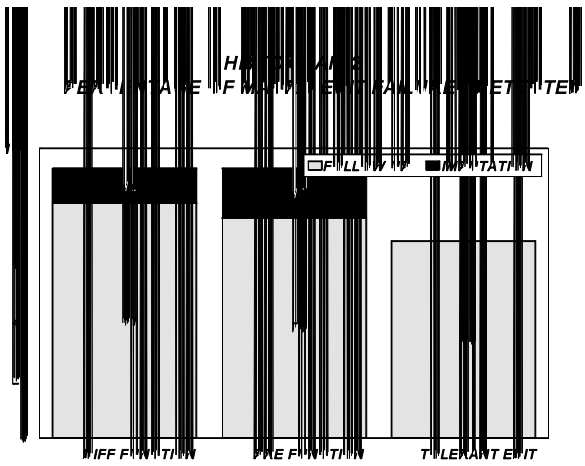
Histogram 1 shows the proportion of questionnaires that were changed in the current process (with 100% follow-up of all edit failures) that were identified by the new edit bounds used in this study. This comparison permits the assessment of the performance of each option in identifying the records that are indeed changed in the current process. The rigid bounds identify 89% of the units that are changed in the current process. On the other hand, the tolerant bounds only catch 45% of them. This result indicates that the tolerant edit bounds are less efficient than the others in detecting questionnaires that need corrections.

Histogram 1 also shows that with the score

HISTOGRAM 1
DETECTION OF PRODUCTION CHANGES
BY THE RIGID AND TOLERANT EDITS BOUNDS



functions, approximately 59% of the follow-ups result in no change in the data. This is similar to the 65% resulting from the tolerant edits. Unfortunately, it is impossible to determine if the absence of change is caused by respondent confirmation or because the respondents were not really followed up in the current process.



Agreement with macro editing.

As a result of macro editing, 226 corrections take place in the current process. Histogram 2 shows the number of these edit failures identified by the three options. The DIFF and PRE scores are the best at catching the macro edit errors; they identify about 93% of the macro edit errors, while the tolerant edits identify only 68%. The little difference between the two score functions is in the number of the questionnaires that are followed up. The DIFF function requires follow-up of 81% of the questionnaires that fail macro edit against 74% for the PRE function.

Impact on the estimates.

Finally, the main criterion for comparing the estimates obtained under the three scenarios is the absolute value of the relative pseudo-bias (ARPB) which is the difference between the current estimate and the estimate produced by any of the options as a percentage of the current estimate.

$$ARPB \text{ ' } 100 = \frac{OPTION\ ESTIMATE \text{ \& } CURRENT\ ESTIMATE}{CURRENT\ ESTIMATE}$$

The smaller the value of the ARPB, the better the option. The absolute value of the relative pseudo-bias are analysed with a view to select the edit option that generates the best estimates.

The value of the absolute relative pseudo bias is computed at the Canada, Province, size-group, and two digit SIC by province by size levels. At the Canada level (Table 2), the absolute value of the relative pseudo-bias is less than 2% for all options,

Table 2: Absolute value of the relative pseudo bias (%) at Canada level

VARIABLE	DIFF FUNCTION	PRE FUNCTION	TOLERANT EDIT
HOURLY EMPLOYEES: EMPLOYMENT	0.32	0.60	0.80
REGULAR GROSS PAY	0.90	1.33	0.19
OVERTIME PAY	0.06	1.09	0.25
TOTAL HOURS WORKED	1.21	1.71	1.16
OVERTIME HOURS	1.43	0.08	3.36
IRREGULAR PAY	5.80	25.69	6.09
SALARIED EMPLOYEES: EMPLOYMENT	0.15	0.00	71917.00[1]
REGULAR GROSS PAY	0.37	0.39	1.44
OVERTIME PAY	0.30	1.08	0.51
HOURS IN WORK WEEK	0.10	0.00	0.03
IRREGULAR PAY	9.72	14.22	9.46
OTHER EMPLOYEES; EMPLOYMENT	0.24	2.39	0.32
REGULAR GROSS PAY	0.72	1.29	0.56
IRREGULAR PAY	0.94	18.97	18.40
MONTHLY GROSS PAY	2.29	2.49	5.87

except for the three irregular payments and monthly gross pay. The fact that, the irregular payments are

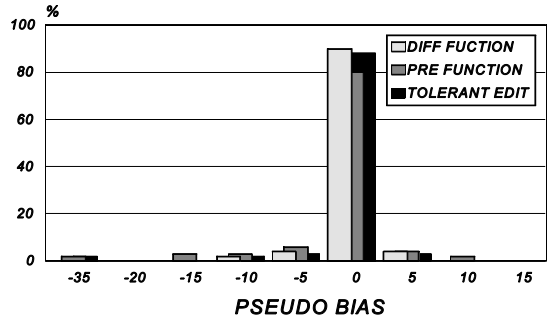
[1] This value was verified for accuracy.

often poorly reported and are included in the monthly gross pay, explains why the ARPB is larger for these variables. In general, the DIFF score function provides the smallest ARPB and seems more stable even for the irregular payments. The PRE score function is found a little superior to the tolerant edit for two reasons. First, the PRE score function causes a zero ARPB for the variables salaried employment and hours in work week. Second, the tolerant edit gives an extreme ARPB value for the salaried employment.

The same pattern is shown at more disaggregated levels although the pseudo bias is higher. At these levels the DIFF score function outperforms the other two edit options and generates estimates that are more concentrated about 0 (no pseudo-bias) as shown by histogram 3. This histogram displays the distribution of the actual relative pseudo-bias (RPB) (not its absolute value), at the two digit SIC x PROVINCE x SIZE cell level for the 602 cells in this study. For each of the three options, the values for the 15 base variables are pooled and then displayed. The criterion for judging the options is the following: the more concentrated the RPB's are about zero or the less the spread about zero, the better is the option. On this basis, the DIFF score function generates the best estimates.

One can ask why the DIFF score function outperforms the PRE score function. For only 13% of the records the scores computed by the two functions are different enough to justify a different action. The records that are identified by the DIFF score function for follow-up but identified for imputation by the PRE score function are records that have many variables in error (4+). On the other hand, the records that are identified for follow-up by the PRE score function and for imputation by the DIFF score function tend to have less variables in error. In fact, half of them had only one. There is no other situation where actions are different. It is possible that the use of better imputation methods would make the PRE function as stable as the DIFF function and consequently provide a good score function for SEPH. However, because the PRE function tends to impute questionnaires with several suspicious data items rather than follow them up, even a good imputation method can lead to poor results because there is insufficient sound data available for partial imputation.

HISTOGRAM 3
DISTRIBUTION OF THE RELATIVE PSEUDO BIAS
AT THE CELL LEVEL FOR ALL VARIABLES



6. CONCLUSION AND RECOMMENDATIONS

Results confirmed that the use of the Hidiroglou-Berthelot bounds in conjunction with the score function tends to identify the largest units that have a high impact on the estimates as needing follow-up. The DIFF score function tends to assign follow-up action flags primarily to large units with several errors. By contrast, the PRE score function concentrates follow-up on the largest units with one error. The records that are identified for imputation by the DIFF and the PRE score functions but are corrected by follow-up in the current process, contribute little to the total estimate and cause only small variations between the estimates using the score functions and the estimate from current process. It seems that the main difference between the DIFF and the PRE score functions is the flexibility in DIFF to assign the error flag and importance to each variable. This feature seems to have a significant impact that favours use of the DIFF score function. Also, use of the difference of the variable between month *t* and month *t-1* in calculating the DIFF score function forces follow-up of questionnaires that have several suspicious data values or that fail the macro edits, again favouring use of the DIFF score function. For these reasons, it is recommended that SEPH use the DIFF score function. Nevertheless, the use of the PRE score function is much better than the current strategy in term of reducing survey costs while still providing estimates of good quality. In fact, a score function similar to PRE has already been implemented in the current SEPH process until the new system takes over. This function is more objective and automates what was often done by the interviewers. However, because there is no partial imputation yet, this function has a follow-up rate of 50% as compared to the 20% of the study.

Considering the potentially large amount of imputation required by the proposed strategy, it is recommended that a separate set of edits and bounds be developed exclusively to identify questionnaires that really need to be imputed. In this way, questionnaires with minor suspicious data would not be followed up or altered during imputation.

APPENDIX A: MONTH-TO-MONTH EDITS APPLIED

The following set of 8 edits on 15 variables was used in the simulation.

- month to month variation of employment, sections A, B and C;
- month to month variation of average weekly hours, section A for education and section B for general;
- month to month variation of regular gross pay, sections A, B and C;
- month to month variation of regular hours, section A;
- month to month variation of overtime gross pay, sections A and B;
- month to month variation of overtime hours, section A;
- month to month variation of irregular payments, sections A, B and C;
- month to month variation of total gross pay.

It should be noted that the new edit strategy also contains some 22 other non month-to-month edits that were not considered in the simulation.

REFERENCES

[1] Berthelot, J. M. , Latouche, M. Improving the Efficiency of Data Collection: A Generic Respondent Follow-up Strategy for Economic Surveys, *Journal of Business and Economy Statistics*, 11:4, 1993, pp. 417-424

[2] Dolson, D. On Using a Very Current Source of Frame Information in Canada's Establishment Based Employment Survey, *Paper presented at the First Eurostat Workshop on Techniques of Enterprise Panel*, Luxembourg, February 21-23,1994.

[3] Greenberg, B., Petkunas, T. An evaluation of Edit and Imputation Procedures Used in the 1982 Economic Censuses in Business Division, in *1982 Economic Censuses and Census of Governments Evaluation Studies*, U.S. Department of Commerce, Washington, 1987, pp. 85-98.

[4] GSFD, Generalized Survey Function Development Team, Methodological and Operational Concepts in the Collection and Capture Module, Technical report, Statistics Canada, Ottawa,1989.

[5] Hidiroglou, M. A., and Berthelot, J.-M. Statistical Editing and Imputation for Periodic Business Surveys, *Survey Methodology*, 12:1, 1986, pp. 73-83.

[6] Schiopu-Kratina, I. and Srinath, K. P. The Methodology of the Survey of Employment, Payroll and Hours, Internal Working Paper, Statistics Canada, Ottawa, 1986.

[7] Latouche, M., and Berthelot, J.-M. Use of a score function to prioritize and limit recontacts in business surveys, *Journal of Official Statistics*, 8:3, 1992, pp. 389-400.

SIMULATION EXPERIMENTS FOR HOT DECK IMPUTATION

By P. Verboon and E. Schulte Nordholt, Statistics Netherlands

ABSTRACT

This report describes two simulation studies of the hot deck method. In the first study data are randomly generated, and various amounts of missing values are then non-randomly 'added' to the data. The hot deck method is used to reconstruct the data in this Monte Carlo experiment. The performance of the method is evaluated for the means, standard deviations and correlation coefficients, and compared with the available case method. In the second study a selection of the data of the Dutch Survey of Living Conditions is perturbed by leaving out specific values on a variable. Again hot deck imputations are used to reconstruct the data. The imputations are then compared with the true values. In both experiments it is concluded that the hot deck method performs generally better than the available case method.

Keywords: hot deck imputation; simulation.

1. INTRODUCTION

While processing statistical data collected through surveys, the statisticians have to deal with missing data. Those non-responses fall into two main categories:

a) **unit non-response** - no data about a statistical unit as a whole.

In order to correct this, the technique of weighting is generally accepted. That is, the remaining data are weighted by making use of known population totals of background variables, such as age, sex, marital status, degree of urbanisation and municipality where the respondent lives. The users of a Statistics Netherlands survey mostly use the weights which are included in the standard file. The researchers have the possibility to use their own weights or to produce an unweighted analysis.

b) **item non-response** - no data about particular items, or impossible combination of data.

A technique to fill in values for these missing values is the hot deck imputation. There are two advantages of using imputation: correction for selective item non-response and use of statistical packages that assume rectangular data matrices without missing values.

The technique of imputation is not as generally accepted as the technique of weighting. Recently much research effort has been put into the study of imputation techniques and the analysis of the results of imputation for data analysis. It is clear that a researcher who uses a file of a statistical office must have the choice to use either imputed variables or non-imputed variables. In Northern America most surveys of the statistical offices are imputed. In many other countries imputation techniques are not used at all. In the Netherlands some surveys like the Survey of Living Conditions use imputations, while others do not.

This paper presents the results of two simulation experiments on the hot deck method. The experiments are designed to resemble empirical situations as close as possible. The objective is to find out whether the hot deck method yields less biased results than methods using only the information available in a variable with missing values.

The paper is organized as follows. Section 2 briefly explains the hot deck method. The first simulation experiment, the Monte Carlo study, is described in section 3 and the results in section 4. The second simulation experiment is a study with data of the Dutch Survey of Living Conditions. More information about this survey and the experiment is given in section 5. Section 6 contains the results of the second experiment. Finally, in section 7 the two experiments are discussed and some conclusions are drawn.

2. THE HOT DECK METHOD

The principle of the hot deck method lies in using the current data (donors) to provide imputed values for records with missing values. A search in the data set is made to find a donor record that matches the record with missing values. Values from the donor record are then used for the imputation. The matching process is carried out using the so called filter variables, which can either be defined by the user or can be found automatically in some optimal way. Records match if they have the same values on the filter variables. When more than one matching record has been found, it is possible to add another filter variable or to select one record at random from the matched records. If all records in the selection have the same value for the variable to be imputed, then this value is imputed for the missing value. We call this an exact match.

A continuous variable has to be divided into categories first. The number of categories should be sufficiently large to obtain a good imputation, but not too large because then the probability increases that no matching records can be found.

Extensive reviews on the hot deck method and its variations can be found in [1], [3] and [4].

3. FIRST EXPERIMENT: MONTE CARLO STUDY

3.1 Background

In a simulation experiment the hot deck method for dealing with item-nonresponse will be compared with the available case method. The latter one computes the parameter estimates for the available data only. The experiment is designed so as to resemble real life situations for survey research as close as possible. The experiment is carried out in two ways. First, a survey design with unequal sampling probabilities is assumed to underlie the data. For computing the required statistics the design weights are used to yield unbiased estimates of these parameters. Secondly, patterns of missing values are generated non-randomly.

3.2 Data

The process starts by generating the data, in this case four variables with scores for n fictitious records (n=200). Two variables are used as criterion variables and two as predictor variables. First, the two predictor variables (x_j ;j=1,2) are independently generated to consist of random scores from the standard normal distribution $\varphi = N(0,1)$. Hence, we have $E(x_j) = 0$ and $E(var(x_j)) = 1$. Next, an additional restriction is imposed on x_1 to mimic a sampling design with sampling probabilities. A certain range of positive values is under-represented. To restore the normal distribution of the records the procedure assigns larger weights to records which are under-represented. The expected percentage of large-weight records is about 5%. Details on how this sampling design is effectuated can be found in [5]. Linear combinations of the generated x-variables are then used to construct the criterion variables:

$$\ddot{y}_1 = (1/\%6)(2x_1 + x_2 + \varepsilon_1),$$
$$\ddot{y}_2 = (1/\%18)(x_1 + x_2 + 4\varepsilon_2),$$

where the n-vector ε is a random component, drawn from a N(0,1) distribution. Henceforth, we shall denote the complete variables by $\ddot{y}_j$. The linear combinations imply that $\ddot{y}_j$ can be predicted rather well by the two x-variables (the expected value of the squared multiple correlation coefficient $R^2(\ddot{y}_1, X)$=.833) and that $\ddot{y}_2$ can be poorly predicted ($R^2(\ddot{y}_2, X)$=.111). The expected values of the variances of the criterion variables is equal to one.

3.3 Erasing data to simulate missing values

The proportion of missing values in one of the criterion variables, denoted by α, is systematically varied: α = 5%, 10%, 20%. The values of α represent small, moderate and large amounts of missing values. The missing values are 'added' to the criterion variables only. A selective mechanism is applied to select the scores to be erased from the data set. The probability that the i^{th} score becomes missing, given its value on $\ddot{y}_j$, is

$$P(\ddot{y}_{ij} \text{ missing} \mid \ddot{y}_{ij} > c_j) = \beta_1$$
$$P(\ddot{y}_{ij} \text{ missing} \mid \ddot{y}_{ij} \# c_j) = \beta_2$$

where c_j is defined as $c_j = a(\ddot{y}_j) + s(\ddot{y}_j)$ with $a(\ddot{y}_j)$ and $s(\ddot{y}_j)$ representing the mean and the standard deviation of $\ddot{y}_j$, respectively.

Furthermore, we impose $\beta_1 >> \beta_2$, so that the probability of being missing is much higher for relatively large scores than for small ones. In this study β_1 is chosen to be approximately twice as large as β_2. The y-variables with missing values are denoted as y_j.

To exclude possible effects of the constructed data which are typical for that data set only, the above process is replicated 50 times, which seems to be sufficient to obtain reliable results. Thus, each time new data with new weights are generated, the missing values are constructed.

3.4 Analysis

In each replication and for all levels of α the means and variances of the criterion variables and five correlation coefficients are computed. The correlations of particular interest are:

$r(y_1, x_1)$, $r(y_2, x_1)$, $r(y_1, x_2)$, $r(y_2, x_2)$ and $r(y_1, y_2)$.

The data set containing the missing values is first analysed using the available case approach. That is, the mean and variance of a variable are computed for the observed records in that particular variable. The correlation between two variables is computed using all available information of these two variables. Next the hot deck method is applied to impute the missing values. The predictor variables are used as filter variables to establish the set of potential donor records.

After imputation of the missing values, the means, variances and correlations are computed from the imputed data set. For the computation of the statistics the design weights are used. Thus, weighted means, weighted variances and weighted correlations are computed from both the available data and from the imputed data.

3.5 Measures of performance

The statistics are compared with their target values: these are the means, variances, and correlations, which are derived from the complete $\ddot{y}_j$'s. The hot deck and the available case method are compared based on the following measures:

(a) the absolute differences between the recovered means and the target means;

(b) the absolute differences between the variances and correlations of the recovered and the corresponding target values.

(c) the averages of those differences across replications.

4. RESULTS OF THE MONTE CARLO STUDY

4.1 Means

First the results for the means are given. The average values of the mean across all replications with α=20% for the available case method is equal to -.110 (y_1) and -.080 (y_2). The hot deck method with α=20% results in averages of -.079 and -.070, respectively. The average target values for the mean are -.058 (y_1) and -.018 (y_2). The standard deviations across the replications for the averages mentioned above are all about .07. Because the expected values of the mean are zero, it follows that the mean is underestimated, even when there are no records with missing values. In Table 1 for both variables the average of the absolute differences between the obtained means and their target values are given.

Table 1. Average Absolute Differences between Target Means and Computed Means

Variable	Available Case			Hot Deck		
	α=.05	α=.10	α=.20	α=.05	α=.10	α=.20
y_1	.020	.029	.071	.023	.032	.060
y_2	.027	.047	.078	.036	.071	.125

There is a clear effect of the proportion of missing values in the data. For both methods the deviation increases with the amount of missing values. The overall results of the hot deck method are clearly better than those of the available case method. With the available case method there is little difference in the results for y_1 and y_2, but the hot deck method yields clearly better results for y_1 than for y_2, which was to be expected because y_1 has a higher correlation with the predictor variables than y_2. For y_2 only with many missing values the hot deck method is performing better than the available case method, but with a small amount of missing values the two methods give similar results.

4.2 Variances

In Table 2 the averages of the absolute differences between the obtained variances and the target variances are given.

Table 2. Average Absolute Differences between Target Means and Computed Variances

Variable	Available Case			Hot Deck		
	α=.05	α=.10	α=.20	α=.05	α=.10	α=.20
y_1	.020	.029	.071	.023	.032	.060
y_2	.027	.047	.078	.036	.071	.125

The results for the standard deviations indicate that the hot deck method is only performing better than the available case method for y_1 and with a large amount of missing values. The expected values of the variance of both variables is 1.00. The target values are 1.030 (.05) and 1.287 (.18), hence a slight overestimation when there are no missing values. The average variance for the available case method with α=20% is .965 (standard deviation is .09) for y_1 and 1.243 (sd is .22) for y_2. The hot deck method gives the averages: .989 (.07) for y_1 and 1.237 (.25) for y_2.

The results for the standard deviations indicate that the hot deck method is only performing better than the available case method for y_1 and with a large amount of missing values. The expected values of the variance of both variables is 1.00. The target values are 1.030 (.05) and 1.287 (.18), hence a slight

overestimation when there are no missing values. The average variance for the available case method with α=20% is .965 (standard deviation is .09) for y_1 and 1.243 (sd is .22) for y_2. The hot deck method gives the averages: .989 (.07) for y_1 and 1.237 (.25) for y_2.

The standard deviation of the variances is much larger for y_2 than for y_1. Thus, when a variable cannot be predicted very well the variance estimator will obtain a larger standard deviation than when it can be predicted well. This effect is due to the design weights, which increase the standard deviation of an estimator. Since y_1 correlates more with x_1 (by which the design weights were defined) than y_2, the variance increasing effects of design weights are less. Furthermore, y_2 contains more random error than y_1, which also causes an increase in variance.

4.3 Correlation coefficients

In Table 3 the deviations from the target correlations are presented. To reduce the number of cells Table 3 contains the averages across the five correlation coefficients of interest.

Table 3. Average Differences and Absolute Differences between Target Correlations and Computed Correlations

Method	Differences			Absolute Differences		
	α=.05	α=.10	α=.20	α=.05	α=.10	α=.20
AC	-.030	-.018	.020	.053	.058	.072
HD	.006	.009	.015	.022	.028	.040

There are two criteria: (i) the differences between target and obtained correlations, and (ii) the absolute value of these differences. With (i) one can see whether there is a systematic under- or overestimation of the correlations computed by a particular method. The absolute differences give an indication of the general bias of the method.

The (absolute) differences between target and obtained correlations increase with the amount of missing values in the data. The hot deck method is better than the available case method, which is mainly due to the correlation coefficients between the predictors and y_1. Positive values in the first three columns of the table indicate under-estimation, thus the target value is larger than the obtained value. The hot deck method underestimates the average target correlations slightly. For the available case method the results are mixed.

5. SECOND EXPERIMENT: STUDY WITH DATA OF THE SURVEY OF LIVING CONDITIONS

The Survey of Living Conditions is conducted by Statistics Netherlands in every four years. The aim of the survey is to collect statistical information about the living situation, the living expenses and realized and planned moves in the Netherlands. The survey is used to evaluate the housing policy in the previous years and to support future housing policy decisions. An important application is the estimation of the number of people who are looking for another house or apartment.

In this simulation experiment one variable of the Survey of Living Conditions 1989/1990 is studied: 'value of the house if it would be sold without people living in it'. Within this paper we will abbreviate the name of this variable to 'value of the house'. This variable was chosen for the experiment because of its importance and big amount of missing values that has to be imputed.

It is not easy to analyse the quality of an imputation because the true values are usually not known. Sometimes the results can be compared with other sources. Such external validations are seldom possible because of different definitions and target populations of other sources. Another possibility to get an idea of the quality of an imputation is to run a simulation experiment. A disadvantage of such an experiment is the need for the assumption that relations between filter and target variables are also valid for the records with missing values.

In a simulation experiment some known values are changed into missing values and scores are imputed for these missing values. If the imputed values are similar or equal to the original (true) values we can have confidence in the applied imputation strategy. A problem for hot deck imputation is that the chance to recover an original value is diminishing the more values in that category are changed into missing values. Therefore in the present experiment it is chosen to change all the scores in one category of the variable 'value of the house' into missing values. In this way one can be sure that the original value will not be imputed because it is no longer in the file. The idea is that the filter variables will cause imputations of values near the original values.

In the present experiment 3922 records with a known score for the variable 'value of the house' are used. The scores for the variable 'value of the house' can be divided into 26 categories. Firstly, the scores for this variable on the 367 records corresponding with category 13 (dfl 150,000) are changed into missing values. Category 13 is the modal category in the distribution of the 3922 records over the 26 categories. By changing these 367 records a large number of observations in the centre of the distribution of the variable 'value of the house' is changed into missing values. Secondly, the scores for this variable on the 137 records corresponding with category 22 (dfl 300,000) are changed into missing values, instead of the 367 records corresponding with category 13. Category 22 is chosen to compare the imputation for values which are in reality in the centre of the distribution with the imputation for values which are in reality in the tail of the distribution. Filter variables used for the imputation process include variables like size of the garden, size of the living room, region where the respondent lives, age of the head of the household, year of construction of the house, total mortgage costs per year, number of rooms in the house, net income of the household, type of the house and number of people living in the community of the respondent. All these filter variables have been categorized (if necessary). The 26 categories of the variable 'value of the house' are temporarily combined to 10 categories for finding the most explanatory filter variables.

6. RESULTS OF THE SECOND EXPERIMENT

Table 4 gives an overview of the results of the imputation process for data which are in reality in category 13. Every category in Table 4 is indicated by its number and by the approximate value of the house in Dutch guilders. In every category the number of respondents is increased by the imputation.

Of course also after the imputation the number of respondents in category 13 is still 0. In the categories near category 13 the number of respondents is increasing more than one could expect from the distribution of the variable before imputation. The current distribution is thus better than what we expect from a random imputation in which the 3555 records have equal probabilities to become the donor record for an imputation. This is caused by the use of the filter variables. However, one should realize that not all imputations give values that are in categories near category 13.

Figure 1 illustrates the results of Table 4 and comparable results for imputation for data which are in reality in category 22. In this figure it can be seen that for data which are in reality in category 13 the percentage of imputations in that category is 0. The same is true for category 22. Further one can see that the modal category of the imputations for data which are in reality in category 13 is in the centre of the distribution, while the modal category of the imputations for data which are in reality in category 22 is in the tail of the distribution. Instead of using a hot deck method we also tried a regression imputation which gave very similar results and therefore the regression results are not presented here.

Because the imputation technique used here is the random hot deck method of the statistical software package SURFOX [2] there is a random component in the imputation process. Therefore it is possibly better to replicate the imputation of both situations (imputation for data which are in reality in category 13 and imputation for data which are in reality in category 22) in order to look at the stability. In the case of imputing for category 13 there are 208 (out of 367) random imputations; the 159 exact matches will not change by repeating the process. In the case of imputing for category 22 there are 66 exact matches and 71 random imputations. Because only part of all imputations can change the differences among the replications are not dramatic, as can be seen in Table 5. In this table three replications of imputing for both category 13 and category 22 are presented. The data of replication 1 of both imputation category 13 and imputation category 22 were used in Figure 1.

Table 4. Number of Respondents by 'Value of the House'

Value of the house		Before Imputation		Imputation		After Imputation	
number	x 1,000 Dfl	number	cumulative	number	cumulative	number	cumulative
1	< 50	58	58	1	1	59	59
2	60	63	121	6	7	69	128
3	70	57	178	5	12	62	190
4	75	54	232	3	15	57	247

5	80	87	319	7	22	94	341
6	90	124	443	14	36	138	479
7	100	178	621	26	62	204	683
8	110	136	757	16	78	152	835
9	120	163	920	12	90	175	1010
10	125	141	1061	24	114	165	1175
11	130	229	1290	30	144	259	1434
12	140	260	1550	40	184	300	1734
13	150	-	1550	-	184	-	1734
14	160	233	1783	24	208	257	1991
15	170	221	2004	26	234	247	2238
16	180	209	2213	26	260	235	2473
17	190	84	2297	8	268	92	2565
18	200	252	2549	22	290	274	2839
19	225	161	2710	17	307	178	3017
20	250	186	2896	22	329	208	3225
21	275	116	3012	9	338	125	3350
22	300	137	3149	14	352	151	3501
23	325	50	3199	5	357	55	3556
24	350	84	3283	2	359	86	3642
25	375	43	3326	1	360	44	3686
26	> 400	229	3555	7	367	236	3922

Figure 1. Percentage of Imputations by 'Value of the House'

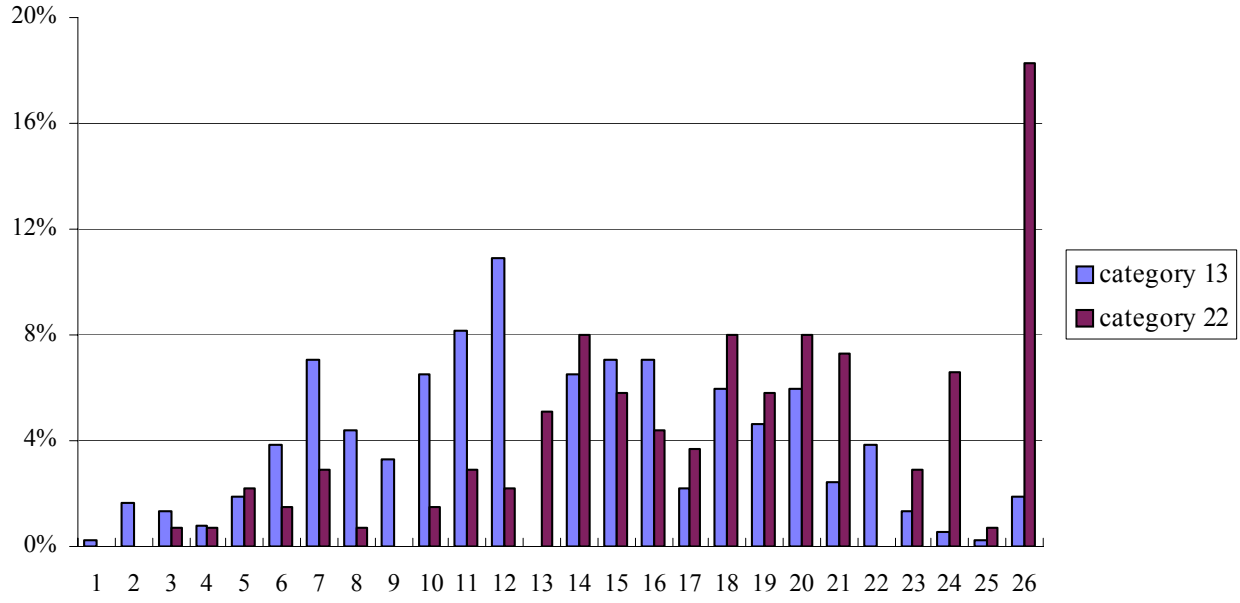


Table 5. Number of Imputations by 'Value of the House' and Imputation Category

Value of the house		Category 13			Category 22		
number	x 1,000 Dfl	r=1	r=2	r=3	r=1	r=2	r=3
1	< 50	1	2	5	0	0	0
2	60	6	3	3	0	2	1
3	70	5	8	5	1	0	1
4	75	3	0	0	1	0	0
5	80	7	6	3	3	1	3
6	90	14	17	15	2	3	2
7	100	26	17	21	4	5	4
8	110	16	13	18	1	1	1
9	120	12	27	20	0	2	0
10	125	24	22	29	2	6	5
11	130	30	27	26	4	3	3
12	140	40	49	43	3	6	4
13	150	-	-	-	7	7	12
14	160	24	20	23	11	9	10
15	170	26	28	30	8	7	5
16	180	26	24	28	6	5	8
17	190	8	14	13	5	6	3
18	200	22	19	15	11	16	12
19	225	17	9	14	8	6	10
20	250	22	21	20	11	10	11
21	275	9	7	8	10	7	8
22	300	14	13	9	-	-	-
23	325	5	3	4	4	4	2
24	350	2	2	2	9	8	7
25	375	1	3	2	1	0	0
26	> 400	7	13	11	25	23	25
number of imputations		367	367	367	137	137	137

r = number of replications

To compare the results of the Survey of Living Conditions experiments two measures of performance are calculated and presented in Table 6. To calculate these measures of performance the number labels of the variable 'value of the house' (1, 2, 3, etc.) are used as category variables. The first measure is the mean absolute deviation of the imputed value from the real value and gives an indication of the quality of the imputation. The second measure is the mean deviation of the imputed value from the real value. This measure is equal to the difference between the mean of the imputed values and the real value and thus gives an indication for the systematic bias.

Table 6. Measures of Performance by Imputation Category

Measures of	Category 13			Category 22		
performance	r=1	r=2	r=3	r=1	r=2	r=3
Mean abs. deviation	4.61	4.57	4.47	5.73	6.20	6.12
Mean deviation	0.44	0.50	0.39	-3.91	-4.57	-4.43

r = number of replications

It is possible to calculate the averages across replications. In Table 7 the resulting mean values before imputation and mean values of the imputation are presented per imputation category. The mean value of the imputation for category 13 is a little bit larger than its true value, while the mean value of the imputation for category 22 is much smaller than its true value. This indicates that it is more difficult to impute for values which are in reality in the tail of the distribution than for values which are in reality in the centre of the distribution (given the chosen measures of performance). In both cases the mean value of the imputation is located between the mean value before imputation and the true value. So though we will not always have imputed values near the true values and we even do not always get nice mean imputed values, the filter variables diminish the selectivity of the item non-response. This shows that it is better to impute values for the missing values by making use of the filter variables than to impute without making use of those variables or simply to drop the records with item non-response.

Table 7. Mean Values by Imputation Category

Imputation Category	Mean value before Imputation	Mean value after Imputation
13	14.3	13.4
22	13.9	17.7

7. DISCUSSION AND CONCLUSIONS

The results of both experiments indicate that the hot deck method generally performs better for estimating the mean. However, for the variance estimation the hot deck method may not be the best choice, as was shown in the Monte Carlo study. Furthermore, the estimation of correlation coefficients between variables with missing values also benefits from the hot deck method, compared to the available case method. It was also shown that the use of design weights does not seem to alter the results drastically.

The hot deck method is particularly useful in situations where we can have filter variables which correlate highly with the variable under study. Apparently, the data in the Survey of Living Conditions study appeared to contain good filter variables for the variable 'value of the house'. It is expected that in many practical situations good predictive variables are available in the data set for many variables with missing values.

Simulation experiments described in this paper are based on random fluctuations. This implies that different results could arise when other experiments with (more) replications are carried out. However, the stability of the results seems sufficient for yielding a reliable indication of the quality of the imputation methods.

REFERENCES

[1] Allen, J. D. An overview of imputation procedures, Staff Report SMB- 90-06, U.S. Department of Agriculture, 1990.

[2] Hooft van Huijsduijnen, J. and A. van Zijl. Surfox, release 1.0, user's manual, 1989 (in Dutch).

[3] Kalton, G. Compensating for missing survey data, Survey Research Centre, Institute for Social Research, The University of Michigan, 1983.

[4] Little, R. J. A. and D. B. Rubin. Statistical analysis with missing data, John Wiley & Sons, New York, 1987.

[5] Verboon, P. and A. Z. Israëls. A simulation study on the treatment nonresponse in continuous data, Research Report, Statistics Netherlands, 1994.

IMPUTING NUMERIC AND QUALITATIVE VARIABLES SIMULTANEOUSLY

By Mike Bankier, Jean-Marc Fillion, Manchi Luc, Christian Nadeau, Statistics Canada

ABSTRACT

For the 1996 Canadian Census, a new minimum change hot deck imputation methodology (NIM) will be implemented. It will perform edit and imputation for the variables age, sex, marital status and relationship to person 1 for everyone in a household simultaneously. Missing, invalid and inconsistent responses will be resolved by the imputation actions. The NIM allows, for the first time, minimum change hot deck imputation of a mixture of numeric and qualitative variables simultaneously. It is also less likely than the old imputation algorithm to create implausible imputed responses. For the 2001 Census, it is planned to generalize the NIM so that it will be able to process all the Census variables.

Keywords: minimum change; hot deck imputation.

1. INTRODUCTION

Many minimum change hot deck imputation systems, both at Statistics Canada and internationally, are based on the imputation methodology proposed by Fellegi and Holt (1976). Examples of such edit and imputation (E&I) systems are CANEDIT [8] and SPIDER [3] used in the Canadian Census to impute qualitative variables and GEIS [4] used in Statistics Canada business surveys to impute numeric variables.

In preparation for the 1996 Canadian Census, the best way to carry out E&I for the demographic variables age, sex, marital status and relationship to person 1 was reassessed. SPIDER was designed to handle small imputation problems and could not be modified to handle E&I of the demographic variables. CANEDIT had been used since the 1976 Census to do E&I for these variables. While CANEDIT successfully identified and imputed the minimum number of variables, many individual imputation actions were implausible and small but important groups in the population had their numbers falsely inflated by the imputation actions. For some households (particularly those with six or more persons), CANEDIT unnecessarily used two or more donors to impute the demographic variables when only one donor was needed. This may have contributed to the implausible combinations of responses. Finally, because CANEDIT could only process qualitative variables, decade of birth had to be used in the edits. Much better edits and imputation actions would have resulted if the discrete numeric variable age could have been used in the edits.

A New minimum change hot deck Imputation Methodology (NIM) has been developed, programmed and applied on a test basis to approximately 80,000 six and eight person households from the 1991 Census. This imputation methodology takes a somewhat different approach to that used by Fellegi and Holt while at the same time capitalizing on some of their insights. The NIM will be used in the 1996 Canadian Census to carry out E&I for the demographic variables. SPIDER will process the other 1996 Census variables sequentially as a series of E&I modules. Each of these SPIDER E&I modules will involve relatively few variables and fairly simple edit rules. For the 2001 Census, the NIM will be generalized and incorporated into SPIDER so that these other E&I modules can be processed, if desired, by the NIM.

The NIM offers some significant advantages as compared to CANEDIT. It allows, given the donors available, minimum change imputation of qualitative and numeric variables simultaneously. It is less likely to falsely inflate the size of small but important groups in the population. The imputation actions for individual households are often more plausible with NIM than with CANEDIT. In addition, it can carry out minimum change imputation for larger groups of variables and larger sets of edits than CANEDIT. Finally, NIM will always perform imputation based on a single donor.

Section 2 explains what the primary objectives for an imputation methodology should be. Section 3 outlines the common features of any single donor hot deck imputation methodology. Section 4 describes the NIM while Section 5 examines the CANEDIT imputation methodology. Some concluding remarks are provided in Section 6.

The NIM was first described in Bankier [3].

2. PRIMARY OBJECTIVES FOR AN IMPUTATION METHODOLOGY

Census edit rules are used to define invalid (including blank) responses for the demographic variables gathered for everyone in Canada. In addition, the edit rules check for responses that are inconsistent within a person and between persons in a household. A household record fails the edits if it contains invalid or inconsistent responses. Otherwise the record passes the edits. An imputation methodology is used to determine which variables to impute for each failed edit household and what values these imputed variables should take on. Often one insists that the imputed values come from a household that passed the edits. This household will be called a donor.

Table 1 displays a household that failed the demographic edits in the 1991 Census along with the CANEDIT imputation action which is underlined. This household failed the edit rule that "The decade of birth for a son or daughter is the same or precedes the decade of birth reported for Person 1". Studying this household, the most reasonable imputation action is to change person 2's relationship to person 1 to spouse. This makes sense because person 1 and person 2 are similar in age, opposite in sex, are married and the ages of the four daughters are reasonable.

Table 1: Failed Edit Household - With 1991 CANEDIT Imputation Action Underlined

Relationship to Person 1	Sex	Marital Status	Age
Person 1	M	Married	34
Son/Daughter	F	Married	32 22
Son/Daughter	F	Single	14
Son/Daughter	F	Single	11
Son/Daughter	F	Single	6
Son/Daughter	F	Single	2

When available donors were investigated in Luc [6], it was found that there were 97 (person /spouse/four child) households for every 3 (person 1/five child) households. Of the existing (person 1/five child) households, few if any would have a married daughter of age 22 present with a person 1 of age 34 and four children of ages 2 to 14. CANEDIT has thus increased the number of a rare type of household (person 1/five child) when creating a (person 1/spouse/four child household) would have been more plausible. CANEDIT, on average, will impute a four child household 1/3 of the time and a 5 child household 2/3 of the time in this situation.

This is because one of three variables (person 2's relationship to person 1 and the decade of birth of person 1 or 2) can be imputed to make the household pass the edits. As described in Section 5, each of these variables has one chance out of three of being selected for imputation by CANEDIT. Thus, in this situation, CANEDIT creates implausible responses while at the same time falsely inflating the number of (person 1/five child) families.

Based on this and other similar examples, it is apparent that the objectives for an automated hot deck imputation methodology should be as follows:

(a) The imputed household should closely resemble the failed edit household. This is achieved, given the donors available, by imputing the minimum number of variables in some sense. The underlying assumption (which is not always true in practice) is that a respondent is more likely to make only one or two errors rather than several. This assumption is made because it is important that a national statistical agency be conservative in the amount of Census data that it modifies.

(b) The imputed data for a household should come from a single donor if possible rather than two or more donors. In addition, the imputed household should closely resemble that single donor. Achieving these two objectives will tend to insure that the combination of imputed and unimputed responses for the imputed household is plausible.

(c) Equally good imputation actions, based on the available donors, should have a similar chance of being selected to avoid falsely inflating the size of small but important groups in the population. The

emphasis is placed on small groups because a relatively low percentage of the demographic data is imputed in the Census. Thus even very poor imputation actions are unlikely to have much impact on large groups in the population.

These objectives are achieved under the NIM by first identifying as potential donors those passed edit households which are as similar as possible to the failed edit household. By this it is meant that the two households should match on as many of the qualitative variables as possible while having small differences between the numeric variables. (Households with these characteristics will be called close to each other or nearest neighbours.) Then, for each nearest neighbour, the smallest subsets of the non-matching variables (both numeric and qualitative) which, if imputed, allow the imputed household to pass the edits are identified. One of these possible imputation actions is randomly selected. As a result, the imputed household will be as similar as possible to the failed edit household while closely resembling the donor.

3. GENERAL SINGLE DONOR HOT DECK IMPUTATION ALGORITHM

It is useful to discuss the general features of any hot deck imputation algorithm whose aim is to impute data for a failed edit household from a single donor. Once this general algorithm is defined, alternative ways of choosing imputation actions within this common structure can be examined.

It will be assumed that the households being edited are split into a number of disjoint imputation groups that will be processed independently. For example, 2000 geographically close six person households might be placed in one imputation group.

Assume that an imputation group has F failed edit households V_f, f = 1 to F, and P passed edit households $\tilde{V}_p$, p = 1 to P. The households are classified into those which fail or pass based on J edit rules which have I variables (either qualitative or numeric) entering at least one of these J edit rules. Each failed edit household V_f will be compared to each passed edit household $\tilde{V}_p$. For a specific V_f and V_p, assume that I_{fp}^* of the I variables do not match. The $2^{I_{fp}^*}-1$ imputation actions possible for that V_f and that V_p can be listed. With $I_{fp}^* = 2$, for example, one can impute the first non-matching variable, the second non-matching variable or both non-matching variables. The possible imputation actions can be identified for each of the P passed edit households. There will then be possible imputation actions V_a for a specific failed edit household V_f. A size measure will be assigned to each of the N_f possible imputation actions and one will be selected with probability proportional to these size measures. Imputation algorithms only differ in what size measure is assigned to each of the N_f possible imputation actions. It will be assumed here, however, that the size measure for imputation actions that do not pass the edits will always be set to zero.

The basic underlying assumption for any hot deck imputation algorithm is that there are donors available which closely resemble the failed edit record. It is also assumed that these donors show the correct distribution of imputed responses for the failed edit record. One of the imputation actions associated with one of these donors will be randomly selected for use with the failed edit record. If there are not enough such donors available, then donors are used which somewhat resemble the failed edit record. In extreme cases, donors are used which do not resemble the failed edit record that closely. In this situation, the required distributional information for the failed edit record is not present in the donors and it is likely that implausible imputed responses will result. Under these circumstances, no imputation algorithm will perform well.

4. DESCRIPTION OF THE NIM

The approach used by the NIM will now be more precisely described. A distance measure $D(A,B)$ is defined which measures the distance between the variables of the two households A and B. With qualitative variables, the distance measure is a count of how many of the qualitative variables of A do not equal (or match) the qualitative variables of B. With a numeric variable such as age, a value in the range 0 to 1 inclusive is added to the distance. If the age of the person in the donor household is similar to the age of the person in the failed edit household, a value close to 0 is added to the distance. Otherwise a value close to 1 is added.

The weighted average (with $0.5 < \alpha \# 1$)

$$D(\underset{\sim}{V_f},\underset{\sim}{V_p},\underset{\sim}{V_a}) \text{ ' } \alpha D(\underset{\sim}{V_f},\underset{\sim}{V_a})\%(1\&\alpha)D(\underset{\sim}{V_a},\underset{\sim}{V_p}) \qquad (2)$$

is calculated for each of the N_f possible imputation actions $\underset{\sim}{V_a}$ which pass the edits. A value of α equal to approximately 0.9 is chosen so that more emphasis is placed on minimizing $D(\underset{\sim}{V_f},\underset{\sim}{V_a})$ rather than minimizing $D(\underset{\sim}{V_a},\underset{\sim}{V_p})$. Those imputation actions which minimize (or nearly minimize) $D(\underset{\sim}{V_f},\underset{\sim}{V_p},\underset{\sim}{V_a})$ are identified, given similar size measures and then one is randomly selected. Other imputation actions are given zero size measures and cannot be selected.

The NIM usually imputes, with a single donor, the minimum number of variables given the donors available. Often the NIM imputes the same number of variables as CANEDIT which is the theoretical minimum. Sometimes, however, CANEDIT used two or more donors to impute the minimum number of variables while the NIM was able to impute the minimum number of variables using a single donor. In a few cases, NIM imputed more than the theoretical minimum number of variables. Usually, however, this was the result of the NIM changing two ages by a little rather than one age by a greater amount so imputation actions of similar quality resulted.

The NIM ensures that the imputation action resembles both the failed edit record and the donor as closely as possible and that equally good imputation actions are selected with similar probabilities. Thus the NIM imputation actions are generally more plausible than those of CANEDIT. Also, small groups are less likely to be adversely affected. More details on the NIM theory is provided in Bankier, Luc, Nadeau and Newcombe (1995) along with computationally efficient algorithms used to implement it.

To illustrate further how this new imputation methodology works, Table 2 lists the nearest neighbour for the failed edit household of Table 1. This nearest neighbour was found in the sample of 350 six person households from the 1991 Census studied by Luc [6]. It can be seen in Tables 1 and 2, for these two households, that all the qualitative variables match except for person 2's relationship to person 1. Four of the age variables do not match, but they are relatively close in value. If all 5 non-matching variables are imputed using the donor's values, the imputed household will pass the edits. Then, however, more than the minimum number of variables will have been imputed. For this reason, the smallest subset of these 5 non-matching variables is determined that, if imputed, will result in the household passing the edits.

Table 2: Nearest Neighbour - Variables That Do Not Match Failed Household of Table 1 are Underlined

Person 1	M	Married	37
Spouse	F	Married	33
Son/Daughter	F	Single	14
Son/Daughter	F	Single	12
Son/Daughter	F	Single	4
Son/Daughter	F	Single	2

For each of the 5 non-matching variables, the value from the failed edit household can be retained or the value from the donor can be imputed. Each of these 5 non-matching variables can only take one of two values and there are therefore $2^5 - 1 = 31$ possible imputation actions. Each of these could be tried to see which ones result in the imputed household passing the edits plus having the minimum number of variables imputed. This, however, could become computationally expensive when the number of non-matching variables increases.

Computational resources can be saved, however, by analysing the edit rules. Only those which the imputed household could possibly fail should be retained. In the six person household stratum in the 1991 Census where all persons in a household are related by blood marriage or adoption, there are 452 edit rules. The edit rule "The decade of birth reported for a father, mother is the same as, or later than that reported for person 1", for example, can be discarded because no one in the failed edit household or the nearest neighbour household is a father or mother of person 1. Other edit rules can also be discarded for similar reasons. This analysis shows that the only edit

rule that the imputation actions for the closest donor can fail is "The decade of birth for a son or daughter is the same or precedes the decade of birth reported for Person 1." This is the same rule that was originally failed.

The failed edit household can be made to pass this edit by changing the decade of birth for person 2, the decade of birth of person 1 or by changing person 2's relationship to person 1. For this particular donor, the only imputation action which imputes a single variable and makes the failed edit household pass the above edit rule is to change person 2's relationship to person 1 to spouse. It should be noted that the 20 closest donors in the sample studied would have all resulted in Spouse alone being imputed. This reflects the fact that there were 97 (person 1/spouse/four children) households for every 3 (person 1/five children) households in the sample studied by Luc [6].

5. CANEDIT IMPUTATION METHODOLOGY DESCRIBED AND COMPARED TO THE NIM

This section describes how the Fellegi and Holt imputation methodology was implemented in CANEDIT. Certain aspects of this implementation, which are not intrinsic to the theory (and could be easily corrected), sometimes resulted in poor quality imputation actions. These are identified below. Other undesirable aspects of CANEDIT, which cannot be so easily corrected, are also discussed. Additional details on the CANEDIT imputation methodology are provided in Appendix A.

To achieve minimum change imputation, CANEDIT first analyses the edit rules to determine the theoretical minimum number of variables to impute in order for the failed edit household to pass the edits. If there is more than one minimum set of variables to achieve this, CANEDIT selects one at random and discards the others. CANEDIT searches for donors which match the failed edit household on certain variables involved in the edits that will not be imputed. It randomly selects one of the donors found in the imputation group which satisfies the matching criteria for the single minimum set of variables retained for imputation. The values from the donor household are substituted for the values in the failed edit household for the variables identified as the minimum number to impute. The matching variables are selected to ensure that the imputed household will pass the edits. This is known as primary imputation. If no donor is found which matches on these variables, CANEDIT attempts to impute the minimum set of variables sequentially using a separate donor for each variable. This is called secondary imputation. If it cannot find a suitable donor for a single variable under secondary imputation, default imputation is used where the left-most allowable response is imputed for a variable (responses are arranged from left to right in alphabetic order).

For each variable under primary imputation that is to be imputed, auxiliary variables can be defined that the failed edit record and the donor have to match exactly. If no donor can be found that satisfies the matching criteria for a minimum change donor plus the auxiliary variables, a donor will be searched for which only satisfies the matching criteria for a minimum change donor.

Under secondary imputation, the minimum set of variables is imputed sequentially. For the first variable in the minimum set, the possible responses allowable for imputation are determined and donors with these responses are retained. Then the first retained donor encountered which matches most closely the auxiliary variables for the first variable in the minimum set is used. This process is then repeated sequentially for the other variables in the minimum set.

In summary, CANEDIT first determines which variables to impute for a failed edit household and then searches for donors. The NIM, in contrast, first searches for donors and then determines the minimum number of variables to impute given the failed edit household and the specific donors. The NIM tries to ensure that the imputed household resembles the donor as closely as possible and that equally good imputation actions are selected with similar probabilities. It should also be noted that the NIM never resorts to secondary or default imputation. The approach used by the NIM is more data driven and is therefore less likely to create implausible imputed responses or falsely inflate the size of small but important groups in the population.

In the subsections which follow, the various components of the CANEDIT imputation methodology are analysed to determine where there are problems.

The difficulties of Sections 5.2 and 5.3 can easily be resolved. The advantages of the NIM compared to CANEDIT, however, based on the above discussion and that in Section 5.1, are clear.

5.1 Determining the Theoretical Minimum Number of Variables to Impute

CANEDIT can determine the theoretical minimum number of variables to impute for qualitative variables. GEIS can determine the theoretical minimum number of variables to impute for numeric variables. CANEDIT can extend its approach to discrete numeric variables by treating them as qualitative variables but it quickly becomes very expensive computationally. CANEDIT, for example, had to use decade of birth rather than age in the demographic edits for this reason. No computationally feasible technique is known that will determine the theoretical minimum number of variables to impute for a mixture of qualitative and numeric variables.

The NIM determines simultaneously the minimum number of qualitative and numeric variables to impute for a particular failed edit record and a particular donor. The problem is much simpler computationally and conceptually because if there are I_{fp}^{*} non-matching variables for a particular V_f and V_p, then there are only $2^{I_{fp}^{*}}-1$ imputation actions that have to be considered.

It should also be noted that determining the theoretical minimum number of variables to impute without looking first at the donors means that preference will always be given to imputing one numeric variable while in some situations imputing two numeric variables by smaller amounts may be an equally valid or better imputation action. Thus GEIS (and CANEDIT if it could handle numeric variables) will sometimes discard legitimate imputation actions.

Finally, if many variables are being imputed and there are relatively few donors, there may in fact exist no single donor which will allow the theoretical minimum number of variables to be imputed. CANEDIT will then go to secondary or default imputation. NIM will impute more than the minimum number of variables in this case but it will be from a single donor and is more likely to be a plausible imputation action.

5.2 Selecting One Minimum Set of Variables to Impute at Random Before Considering the Distribution of Responses

Both CANEDIT and GEIS randomly choose a single set of variables to impute whenever more than one minimum set is found. This is done to save computational resources but is not an integral part of the theory of Fellegi and Holt. The example in Table 1 of Section 2 shows that doing this can artificially increase the size of certain small groups plus create implausible imputed responses. This, if possible, should be avoided. This can be done by considering all minimum sets of variables to impute when searching for donors. SPIDER, in fact, does this.

5.3 Searching for Donors

CANEDIT determines a subset of variables (known as matching variables) which enter the edits but will not be imputed. CANEDIT then searches for donors which match the failed edit household on all the matching variables.

This method of searching for donors is not very satisfactory. Often only a few matching variables are used. In the example of Table 1, CANEDIT only required that the donor match the failed edit household on Decade of Person 1, Relationship of Person 2 to Person 1 and Marital Status of Person 2. This is because the matching variables are chosen to ensure that the imputed household passes the edits. It does not guarantee, however, that the donors which qualify closely resemble the failed edit household. Thus CANEDIT will not necessarily select a nearest neighbour. Some of the possible damage can be mitigated by the use of auxiliary constraints but this requires the user to be aware of the problem and use the auxiliary constraints wisely.

It has also been found that CANEDIT often resorts to secondary imputation actions even when the NIM is able to impute the minimum number of variables using a single donor. This happens because CANEDIT requires that the donor match the failed edit household on all the matching variables under primary imputation and this is not always possible. With secondary imputation, however, a donor will always be found if one exists which has an acceptable value for the variable being imputed. This can result, however, in the donor matching the failed edit record on few if any

variables. Also, if two or more variables are being imputed for a household, two or more donors will be used. CANEDIT used secondary or default imputation actions for 42% of the eight person failed edit households on the East regional data base while the NIM was able to impute the minimum number of variables from a single donor for 95% of the eight person failed edit households on the Ontario regional data base.

In the 1991 Census, the number of persons born in the decade 1870 increased from 60 to 300 because of imputation. This is probably the result of default imputation being applied because CANEDIT was not able to find a donor under secondary imputation. Thus, the size of a small group has been artificially increased by the CANEDIT imputation actions.

The above problems related to searching for donors could be resolved by having CANEDIT search for donors in an improved fashion (e.g. doing something similar to what the NIM does).

6. CONCLUDING REMARKS

The NIM performs minimum change hot deck imputation of qualitative and numeric data simultaneously, given the donors available, in a computationally feasible fashion. It has the potential for application to a wide range of surveys and censuses. The 1996 Census production version of the NIM software is now being programmed. Further study will be done to optimize parameters of the NIM in preparation for its use on the demographic variables in the 1996 Canadian Census.

Appendix A: The CANEDIT E&I Methodology

CANEDIT was first used for E&I of 1976 Census data. In the 1991 Census, its most important application (which will be described below) was to carry out E&I for the demographic variables age, sex, marital status and relationship to person 1. CANEDIT uses the minimum change hot deck imputation methodology proposed by Fellegi and Holt (1976). This methodology is based on the following three principles.

1. The data in each record should be made to satisfy all edits by changing the fewest variables;

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

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

It is shown in Section 5 that CANEDIT respects principle 1 and 3 but does not always respect principle 2.

CANEDIT defines the edits as a series of conflict rules. A household fails the edits if it matches one or more of these conflict rules. Otherwise, it passes. To achieve minimum change imputation, CANEDIT first analyses the edit conflict rules to determine the theoretical minimum number of variables to impute for the failed edit household to pass the edits. If there is more than one minimum set of variables to achieve this, CANEDIT selects one at random and discards the others.

Primary Imputation:

Households being edited are subdivided into a number of strata based on the number of persons in the household and whether everyone in the household is related by blood, marriage or adoption. Each stratum is further subdivided into disjoint imputation groups of approximately 2048 geographically close households. CANEDIT searches within the imputation group for passed edit households (donors) which match the failed edit household on certain variables involved in the edits that will not be imputed. It randomly selects one of the donors found in the imputation group which satisfies the matching criteria for the single minimum set of variables retained for imputation. The values from the donor household are substituted for the values in the failed edit household for the variables identified as the minimum number to impute. The matching variables are selected to ensure that the imputed household will pass the edits. Exact matches on these variables is required between the donor and the failed edit households. This is known as primary imputation.

For each variable to be imputed, auxiliary constraint variables may be defined for which an exact match is required between the failed edit household and the donor. If no donor can be found that satisfies

the matching criteria for a minimum change donor plus the auxiliary constraints, a donor will be selected under primary imputation which only satisfies the matching criteria for a minimum change donor.

If no donor is found which satisfies the matching criteria for a minimum change donor (this happened, for example, for 15.6% of the imputed 6 person households on the EAST data base in 1991), CANEDIT attempts to impute the minimum set of variables sequentially using a separate donor for each variable. This is called secondary imputation. If it cannot find a suitable donor for a single variable, default imputation is used where the left most allowable response is imputed for a variable (responses are arranged from left to right in alphabetic order). More detail on secondary imputation is given below.

Secondary Imputation:

Variables (from the minimum set to impute) are imputed sequentially in the order in which they were listed when the auxiliary constraints were defined. Variables with no auxiliary constraints are imputed in alphabetic order after these other variables.

For the first variable to be imputed for the failed edit household, acceptable responses that are consistent with the responses for other variables not being imputed will be determined. Then passed edit households with these acceptable responses among the 500 households (both passed and failed) that follow the failed edit household in the imputation group will be identified. If none are found, default imputation will be done. Then each of the other variables from the minimum set to impute are processed sequentially in the same fashion. It should be noted that the imputation of earlier variables may restrict the number of acceptable values that can be imputed for the later variables. The donor household selected must either have passed the edits or have been imputed in primary imputation.

When searching for a donor to impute for a particular variable, the households with acceptable responses are evaluated according to how well the auxiliary constraints are satisfied. A weight is assigned to each auxiliary constraint variable defined for the variable to impute. The weights decrease geometrically (e.g. 16, 8, 4, 2, 1) from the first to the last auxiliary variable listed for the variable being imputed. When the failed edit household and the donor household are compared, the weights associated with the auxiliary constraint variables for which there are matches are added. CANEDIT uses the donor with an acceptable response that has the highest sum of weights. If there is more than one donor with the highest sum of weights, CANEDIT uses the first encountered among the 500 households.

Comments on When There is Only One Variable to Impute

When there is only one variable to impute, it is possible that no donor will be found under primary imputation because the donor must satisfy the matching criteria exactly for the minimum change donor. With secondary imputation, however, a donor will always be found if one exists which has an acceptable value. This can result, however, in the donor matching the failed edit record on few if any variables. If the auxiliary constraint variables are carefully selected, however, secondary imputation may result in reasonable imputation actions being selected that were not considered under primary imputation. Having the auxiliary constraint weights decrease geometrically, however, may not be optimal in terms of selecting imputation actions.

REFERENCES

[1] Bankier, Mike. Imputing Numeric and Qualitative Census Variables Simultaneously, Social Survey Methods Division Report, Statistics Canada, April 14, 1993.

[2] Bankier, M. , Luc, M. , Nadeau, C. , Newcombe, P. Imputing Numeric and Qualitative Census Variables Simultaneously, Social Survey Methods Division Report, Statistics Canada, March 22, 1995.

[3] Ciok, Rick. Spider - Census Edit and Imputation System, Social Survey Methods Division Report, Statistics Canada, September 1992.

[4] Cotton, Cathy. Functional Description of the Generalized Edit and Imputation System, Business Survey Methods Division Report, Statistics Canada, July 25, 1991.

[5] Fellegi, I. P. and Holt, D. A Systematic Approach to Automatic Edit and Imputation, *Journal of the American Statistical Association*, March 1976, Volume 71, No. 353, 1993, pp. 17-35.

[6] Luc, Manchi. Preliminary Results on Analysing Age, Sex, Marital Status, Common-law Partner Status and Relationship to Person 1 for Some 1991 Six person Household Data, Social Survey Methods Division Report, Statistics Canada, February 9, 1993.

[7] Nadeau, Christian. Statistics on the Utilisation of Different Imputation Techniques in CANEDIT, Social Survey Methods Division Memo, Statistics Canada, October 26, 1992.

[8] Pageau, François. Features of the CANEDIT Software, Social Survey Methods Division Report, Statistics Canada, September 1992.

Chapter 2

DESIGNING SETS OF EDITS

FOREWORD

By Giulio Barcaroli, National Statistical Institute, Italy

In data editing a fundamental role is played by the *design* of sets of edits. This step is not, or at least should not be, a mere declaration of the rules applied to data in order to detect and eliminate as many errors as possible, discarding the impact of this application on multivariate distribution. On the contrary, many aspects should be considered in this task, such as the fulfilment of the following requirements:

- the *completeness* of the set of edits, not only in the sense stated by Fellegi-Holt methodology (explicitness of knowledge contained in initial edits), but in the sense of full exploitation of available knowledge. That is, all possible rules to identify errors must be defined, with regard both to the structure of the questionnaire (permissible paths of compilation) and to the relationships among objects of the real world that are under investigation;

- the *correctness* of the set of edits, i.e. every edit must correspond to the knowledge we have of the domain, and this knowledge must be correct with respect to the real world. As a consequence, the set must be free of contradictory edits;

- *subject-matter people* should be responsible for the task, and not only informaticians, as the former can do a better job with regard to the previous points;

- edits should be declared in a clear and legible way (preferably defined in *natural language*), mainly because this would allow subject-matter people, who are not computer experts, to define and to maintain them;

- the *effects* of any individual edit, and of the whole set, should be analysed before the real application, in order to be sure that no additional errors are introduced and that multivariate distribution is not distorted.

Clearly, all this could be achieved much more effectively and with less effort if specific software, designed and developed with these targets in mind, is available. Generalized software systems have great advantages over *ad hoc applications*: they obviously reduce application costs, but, much more importantly, they allow the correct application of given methodologies to each suitable situation.

The aim of this chapter is to give a general overview of the current situation in National Statistical Institutes and to show some advanced experiences based on the use of generalized systems, with particular regard to the way edits are defined, tested and applied to data.

The first contribution reports a description of the Italian software DAISY, implemented to apply Fellegi-Holt methodology to edit and imputation of qualitative data. The emphasis is not only on the structure and the functions of the system, but also on the particular methodology for the design and tuning of the edits. The applications related to the Labour Force Survey and the Multipurpose Households Survey are briefly described.

The next two papers describe two edit systems developed at the U.S. Bureau of the Census. The system SPEER (Structured Programs for Economic Editing and Referrals) is adopted in U.S. Bureau of the Census for edit and imputation of quantitative variables. It is characterized by the particular form of definable edits: ratios between couples of variables whose values must be internal to given intervals. The underlying methodology is very similar to Fellegi-Holt schemes for qualitative variables: check of edits to detect inconsistencies and redundancies, generation of implicit edits, localization of errors in data and imputation of variables (the latter two steps following the well-known principles of final correctness and minimization of change of data). A general methodology for the definition of edits is given: in particular, the user is assisted in this task by a specific software, D-MASO, which suggests lower and upper limits for the intervals of the edits on the basis of available data. In the paper the particular interactive application of SPEER to the Annual Survey of Manufactures and to the Census of Manufactures is reported.

The DISCRETE system, developed as a prototype in the U.S. Bureau of the Census, is also an implementation of Fellegi-Holt methodology, with new algorithms for implicit edit generation and error localization.

The fourth paper concerns the National Agricultural Statistics Service of the U.S.A. (NASS) experience of computer generation of instruments for data collection and interactive editing. In NASS a given survey can consist of up to 46 different questionnaires, tailored for the different States in which the survey is held: this is the case, for example, of the Quarterly Agricultural Survey. Moreover, every NASS State Office has to collect and edit data from different surveys, generally using CAPI and CATI techniques with BLAISE software. So, the following idea has arisen: instead of creating a complete BLAISE code for any different questionnaire in a given survey, the standard and common part of all questionnaires is separated, and maintained in a common library by subject-matter specialists, who also define specifications related to different States and repetitions of the survey. A relevant system has already been developed: the next step is the development of a system with one more dimension, not only for one given survey, but for different versions of the survey.

The fifth and final paper reports the results of a survey that was carried out in cooperation with some National Statistical Offices in 1994-95. The following information about software systems and procedures currently used are presented: field of application, if generalized or ***ad hoc,*** profile of users, characteristics of languages for data and edits definition. With regard to this last point in particular, reported examples are quite interesting and they underline the great variety of specialization required of the user of these systems.

DAISY (DESIGN, ANALYSIS AND IMPUTATION SYSTEM): STRUCTURE, METHODOLOGY AND FIRST APPLICATIONS

By Giulio Barcaroli, Marina Venturi, National Statistical Institute, Italy

ABSTRACT

In 1989 IBM-SEMEA and ISTAT began cooperation in the field of generalized software for edit and imputation of statistical data, which led in 1992 to the development of a first version of a system named DAISY (acronym standing for Design Analysis and Imputation system). The final target of this system, which can be regarded as a specialized CASE tool, is very ambitious: the complete coverage of design and execution of data editing procedures for any statistical survey involving qualitative variables (mainly household surveys). These procedures can be very complex, as they are commonly composed of sequences of both probabilistic and deterministic steps, and the system supports the user (the statistician responsible for the survey) during the definition and the analysis of each of them. Up to now, DAISY has been applied to two of the most important household surveys led by the Italian Institute, the Labour Force Survey and the Multipurpose Households Survey, with very satisfactory results.

Keywords: Fellegi-Holt; data editing and imputation; data quality; statistical software.

1. INTRODUCTION

A statistical survey can be viewed as a production process composed of different and strictly interrelated phases. In particular, the phases of data collection and data input usually produce a certain amount of *errors* (we can define an error as the difference between real world and its representation in data), which can be classified as *systematic* (due to identifiable defects in the organizational and conceptual aspects of these two phases) and *stochastic* (randomly appearing among data, not depending on a given cause). A subset of both types of errors can be identified and eliminated, as they produce missing values, out-of-range values and logical inconsistencies among permissible values of different variables. Identification and elimination of errors can be made by men who analyse data in an interactive way, or by programs that scan and modify data in an automatic way. The former is often the best solution from a qualitative point of view, but is much more expensive in comparison with the latter. Programs can be classified as *deterministic* or *probabilistic*. A deterministic program is based on a set of *if-then rules*:

IF (error condition(s)) THEN (imputation action(s))

In this case, corresponding to any given error situation (i.e. a logical inconsistency among values of two or more variables), the attribution of the new

values to the given variables is defined. For example:

IF ((age < 15) and (marital status = married)) THEN (marital status <-- single)

It can easily be shown that the deterministic approach has substantial drawbacks, as it does not fulfill the three fundamental requirements:

a) the final correctness of data with respect to the defined error conditions;

b) the minimality of data modifications;

c) preserving the initial distribution of variables.

On the contrary, the probabilistic approach, based on the Fellegi-Holt methodology [5] requires only the definition of a set of error conditions (*edits in normal form*). A particular algorithm permits the choice of the variables to be modified and the values to be assigned to them in order to eliminate all the inconsistencies, minimize data modification, and not distort the multivariate distributions.

For these reasons, the probabilistic approach is undoubtedly better than the deterministic, and should always be followed. But there is a conceptual limit: *it is applicable only to eliminate stochastic errors*. In the case of systematic errors, deterministic programs must be used.

Another problem that can arise under the probabilistic approach is related to *feasibility*. According to Fellegi-Holt methodology, once the initial set of edits is defined in normal form, it has to be analysed in order to eliminate redundant and contradictory edits, and all *implicit edits* contained in this set must be generated to ensure the optimal imputation. The complexity of implicit edit generation grows exponentially with respect to the number of initial (or explicit) edits. Beyond a certain number of explicit edits the implicit edits can not be generated, and the optimal imputation is not guaranteed. In this case different actions can be undertaken:

- the initial set can be split into two or more subsets, each of them feasible in terms of implicit edits generation;

- the implicit edits are not generated, and a different imputation algorithm is used, attempting to identify the minimal set of variables to be changed for each case.

In the case of surveys whose data contain both stochastic and systematic errors, and where the number of different types of errors is not high, the edit and imputation strategy can be very complex. The resulting procedure will be composed by different deterministic and probabilistic steps. Each of them has to be defined and tuned in a way that ensures the maximization of the probabilistic approach and the limitation of the deterministic one only to systematic errors. Bearing these targets in mind, we developed DAISY and defined a methodology for its use.

The next sections proceed as follows:

- the *general architecture* of the system is described;

- an analysis is made of how the *probabilistic approach* is carried out by the current version of DAISY, and how the next version should allow the optimal application of *the deterministic approach*;

- a *general methodology for the design* of a complex edit and imputation procedure is introduced;

- *two experiences of application* of DAISY are reported: the Labour Force Survey and the Multipurpose Households Survey.

2. SYSTEM ARCHITECTURE

DAISY has three most important characteristics:

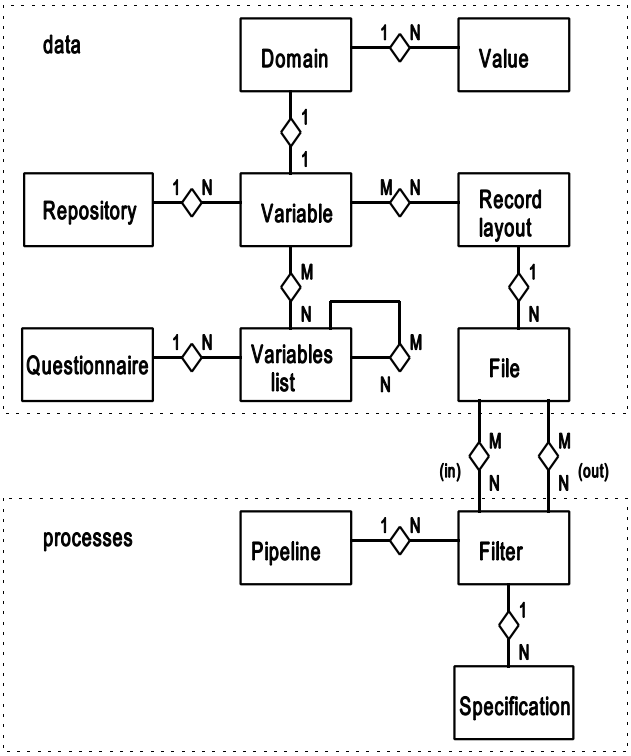
- *open structure*: the first version (available from June 1992) allows the entire probabilistic approach, while next versions should contain modules for handling the deterministic approach, plus more complex tasks, such as the inter-records / inter-files editing; this will be possible thanks to the modularity of the system which allows the incremental implementation of new functions;

- the direct *co-ordination of the phase of data design with the phase of data editing*: the internal structure of the system is based on a repository which contains all the information about data (variables, domains, layouts);

- the relatively high *independence of the statistician* in using the system: the need for programmers and analysts is reduced to a minimum with the current version, and should be almost zero with the next one.

The structure of the system is shown in the Entity-Relationships scheme in figure 1.

Figure 1. Entity-Relationships scheme of the internal structure of DAISY



This scheme can be subdivided into an upper part, containing information about *data*, and a lower part, containing information about *processes*.

For each survey, the statistician must define a repository, containing the following elements:

- *variable*, including information about data considered in the imputation process. It is linked to *domain* and *value*, containing the set of values that a given variable can assume. Variables can be grouped in one or more *lists of variables*, in order to follow the structure of the *questionnaire*: these elements represent the *data conceptual level*;

- variables must be inserted in one or more *record layouts*, which characterizes one or more data *files*: this is the *data logical-physical level*.

Once data is defined at the different levels, the statistician has to design processes that must be applied to them in order to detect and eliminate errors. Using a metaphor, the edit and imputation procedure can be viewed as a flow of raw data going through a *pipeline* and purified by a suitable sequence of *filters* until they become clean. Each filter is characterized by *specification*s which show how it purifies data. The concepts of pipeline and filter require some additional explanation.

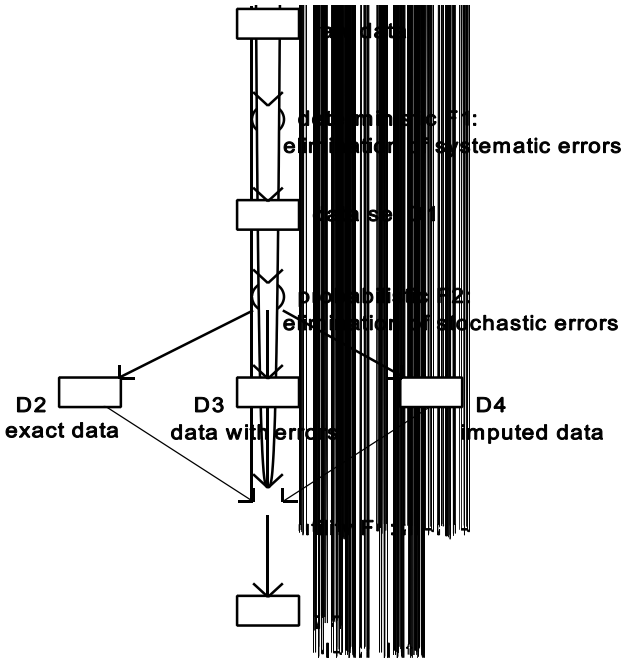
A pipeline is a directed and acyclic graph, whose nodes and edges represent filters (processes) and files (data), characterized, respectively, by specifications and record layouts. The construction of a pipeline is made by indicating all the filters contained in it, taking care of observing the following constraints:

- at least one filter, whose input file(s) does not constitute the output of any other filter, must exist (*initial filter existence constraint*);

- at least one filter, whose output file(s) does not constitute the input of any other filter, must exist (*final filter existence constraint*).

Each filter has to be defined together with the associated data sets and record layouts. The system permits the building of the correct sequence of filters by considering the function (input/output) of data sets linked to each filter: if a data set D is output file of filter F_i and input file of filter F_j, then F_i is directly connected to F_j and is executed just before it. An example of pipeline is shown in figure 2.

Figure 2. Example of pipeline



Only probabilistic filters can be defined in the current version of DAISY; in the final version two kinds of filters will be available:

- *proper filters*, whose execution produces a real transformation of data;

- *utility filters*, producing reorganization, composition or enrichment of initial raw data.

As examples of the latter, we can mention filters

of SORT, MERGE, JOIN (reorganization and composition filters), and filters of DERIVATION (enrichment filters: the computation of new variables based on the existing ones).

Among the former, the most important are edit and imputation filters, probabilistic or deterministic, intra-record or inter-record.

Filter specifications permit the indication of the characteristics of the processing: their structure and complexity depend on the type of filter. The simplest case is the SORT filter: only sorting variables and sorting order (descending or ascending) should be indicated. Edit and imputation filters are more complex: specifications consist of edits in normal form and if-then rules.

While utility filters are immediately executable after defining their specifications (only syntax check is required), edit and imputation filters need a pre-processing of specifications to ensure their correctness, non redundancy and completeness.

2.1 Probabilistic imputation

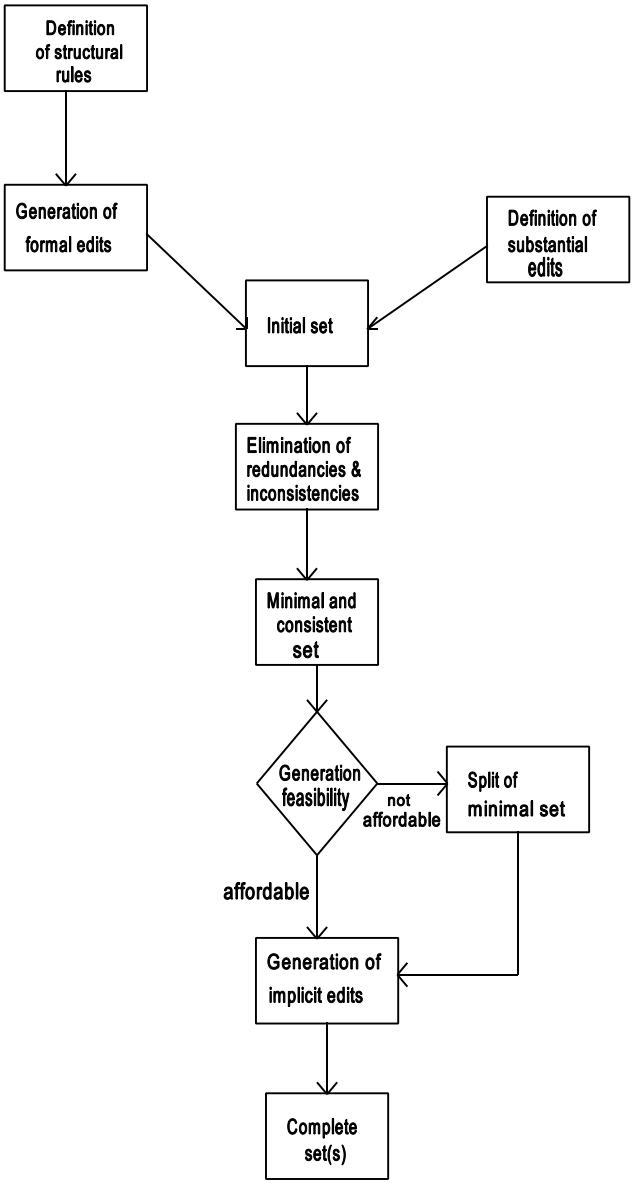
DAISY implements an extension of Fellegi-Holt algorithm for probabilistic imputation. Following this approach, the job of the statistician is limited to declaring the error situations. After this the algorithm provides the best solution for imputation considering actual errors in any current record.

Definition of edits in real surveys is not an easy task. It is composed of the following steps:

- definition of the *initial set of edits*;
- elimination of redundant edits (*minimal set*);
- elimination of contradictory edits (*consistent set*);
- generation of implicit edits (*complete set*);
- if generation can not be accomplished, the initial set will be split into two or more subsets, and implicit edits will be generated for each of them, or alternatively, the minimal and consistent set of edits will be used for data imputation.

The general flow of steps is shown in figure 3.

Figure 3. Flow of probabilistic filter definition



The initial set of edits in normal form can be divided into *formal edits* and *substantial edit*s. The formal edits are derived from the structure of the questionnaire (particularly from the rules for the compilation of its sections and subsections, the so- called *structural rules*). The substantial edits depend on the knowledge of relationships among objects in the real world.

Formal edits can be derived directly from structural rules, which have the following syntax:

BLANK | NONBLANK variable_name | list_name IF(F) condition

In other words, a structural rule is a statement that defines the obligation to complete a variable (NONBLANK) or a list of variables (generally corresponding to a section of the questionnaire), or to

leave them empty (BLANK), depending on conditions concerning other variables. The double F in IFF stands for *if and only if* and indicates that the reverse must also be true. For example:

BLANK professional_condition IFF age < 15

means that the variable professional_condition *must not* contain any value if the age of the person is less than 15 years, and also that professional_condition *must* contain value if age is greater or equal to 15. This single structural rule produces the following two edits in normal form:

professional_condition (not blank) AND age (0-14)
professional_condition (blank) AND age (15-max) [1]

The transformation is performed automatically by DAISY. This is useful in real cases, where the structural rules allow the accurate and rapid definition for lists of variables of tens and even hundreds of formal edits.

Once the initial set is defined, DAISY analyses it in order to discover and eliminate redundant and contradictory edits. Redundancies are eliminated in a single step together with indicating contradictions found by comparing couples of edits. On the contrary, discovering contradictions by the analysis of triples, ... , n-tuples of edits is a more complex task. This can be performed in another step, or can be a sub-result of the step of implicit edits generation. In any case, the contradictions are not removed automatically. Any single inconsistency must be solved by the user.

The obtained minimal set of edits is also a consistent set with respect to contradictions resulting from couples of edits, I. e. it can be considered *partly consistent* and not *totally consistent*. The next step, the generation of implicit edits, may require a great amount of time and memory which could go beyond the available resources (as mentioned before, the complexity is exponential in comparison with the number of initial edits). To assess the feasibility of the task, DAISY computes the number of operations needed for the first cycle of the generation. This can give a good idea of the overall complexity.

If the generation can be performed, the complete set of edits is obtained, and the filter can be applied to data with no additional operations. Otherwise, DAISY permits the subdivision of the initial set in two or more subsets, each of them feasible from the generation point of view. This can be done in two different ways:

- manually, by flagging edits belonging to one or the other subset;

- automatically by the system, using one of two available algorithms (*topological* and *KNN*) which cluster edits by minimising the number of common variables.

Alternatively, we can decide not to generate the complete set of edits, and use instead the minimal and consistent set to impute data. This can be done, but may lead to the following inconveniences:

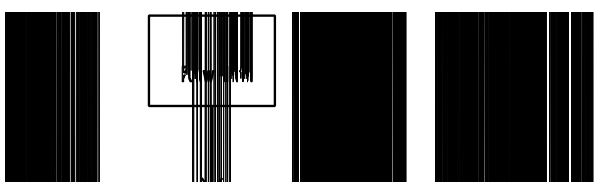
- more time needed for data imputation, as the system has to make attempts to find a solution for any current record, and the number of attempts may be very high;

- impossibility to correct some records (this is a limitation of the current version of DAISY and is not due to theoretic impossibility).

The higher the ratio between the number of implicit edits and the number of initial edits, the more often these inconveniences can occur. Therefore, before choosing between the first or second strategy, this ratio has to be evaluated.

The statistician must also define *how* the filter should be applied to data, by choosing the values for a set of parameters. To explain the meaning of these parameters we have to recall the basic concepts of Fellegi-Holt methodology concerning the imputation phase. When a record activates ("fails") one or more edits, the first step is the determination of the *minimal set of variables* to modify in order to de-activate the failed edits. The minimal set is given by the minimal number of variables which "covers" all the failed edits. By using the parameter FIXED, the user can assign to any variable a fixity degree, from a minimum 1 to a maximum 9 (in this case the variable can never be chosen in the minimal set). The values 1 to 9 depend on the judgement of the statistician about the error probability related to the variables. The value 9 must be assigned if the related variable has already been edited and imputed in a previous step, and therefore cannot be modified again.

Figure 4. Flow of imputation step



The second step is the determination of the potential values to assign to the variables in the minimal set. This is done by computing for each variable the complement to the intersection of ranges that appear in all the edits in which the variable is present. At this point the real attribution of values can occur in one of the following ways (see figure 4):

1. *joint imputation by donor*: one record is chosen in a set of "good" records, and its values of the variables in the minimal set are given to the current record. The search for the donor is undertaken by considering the so-called *matching variables*, i.e. the variables that appear in the failed edits but not in the minimal set of variables that have to be modified. There are two possibilities:

   1.1. a record with the *same* values in the matching variables is searched (*restricted joint imputation*); this is the best way to impute, ensuring the maximum adherence to the multivariate distribution;

   1.2. a record with values in the matching variables that fall in given ranges is searched (*enlarged joint imputation*);

2. *sequential imputation by donor*: for any variable in the minimal set, a donor must be found whose value falls in the computed interval for that record;

3. *imputation by marginal distributions*: the value to assign to a variable in the minimal set is chosen by generating a random number with probability given by the frequencies distribution of the variable, computed on real data or given by the statistician.

The system, if not set otherwise, makes attempts to execute the restricted joint imputation. If the appropriate donor records can not be found, first the enlarged joint imputation is tried, and in case of failure the sequential imputation. If this too is unsuccessful, the imputed values are randomly generated by the marginal distributions. The user can alter this sequence. For example, since the restricted joint donor imputation can be very time-consuming, it can be avoided by specifying a certain parameter (NOIMPR). Or, if the statistician has more confidence in the marginal distributions method for a given set of variables, and one of these variables is in the minimal set for a given record, the donor imputation is ignored (parameter MARGINAL). Finally, in case the statistician wants to furnish a distribution that does not depend on the real distribution computed on raw data, this can be obtained by using the parameter WEIGHT.

The search for the donor is made in a "tank" containing a number of possible donors. How the tank is built has a great influence on the quality of the imputation. By default, a tank of maximum 2000 records is randomly created at the beginning of the imputation phase; it is never renewed and any record can be used as donor an unlimited number of times.

The user can specify the number of records in the tank (which can include all "good" records in the data set), the period of "refreshment" of the tank, and the maximum number of times that the same record can be used as a donor. Very important for the user is the possibility to specify "key variables" for the creation and the dynamic refreshment of the tank. This ensures that the donor and the receiving record will have the same values (or at least not "distant" values) in those variables: in real applications, regional variables were chosen as key variables.

2.2 Deterministic imputation planned for the future

The probabilistic approach is suitable in the case of stochastic errors, but when systematic errors have been detected, the deterministic approach has to be used. Only if-then rules are suitable for these kinds of errors that can affect one or more variables, caused by structural problems in the questionnaire, in the organization of data collection and data storing.

The current version of DAISY does not yet have functions to handle deterministic filters, but specific requirements for the implementation of these functions have already been defined. The system will allow the insertion of *deterministic imputation rules* (DIRs), and the following functions will be executed:

1. *elimination of redundancies* [8];

2. *analysis of inconsistency*, in two distinct ways:

   - by considering only the condition part of DIRs, which is conceptually identical, or can be reduced to edit in normal form, and by performing an inconsistency check as in the probabilistic approach;

   - by considering also the imputation part of DIRs. In this case the analysis is much more complex, though feasible. The real problem is that the elimination of inconsistencies must be

accomplished manually by the statistician, and the final consistency is difficult to reach.

3. starting from DIRs, *automatic generation of executable code* in a general purpose language (COBOL, FORTRAN or C).

These three steps can be successfully performed only if the number and the complexity of DIRs are reasonably low. Actually, this can be obtained if the deterministic approach is followed only for systematic errors, which should be a small subset of all possible errors. The methodology to ensure this result is introduced in the next section.

3. A METHODOLOGY FOR THE DESIGN OF COMPLEX EDIT AND IMPUTATION PROCEDURES

The preparation of an automatic procedure for edit and imputation of data can be a very complex operation, as has been the case for Italian Labour Force and Multipurpose Households surveys. Hundreds of edits and if-then rules have to be defined, their correctness, consistency and completeness must be checked, and their effect on data has to be estimated. The final target is a procedure which "captures" and eliminates as many errors as possible, but does not introduce distortions in data.

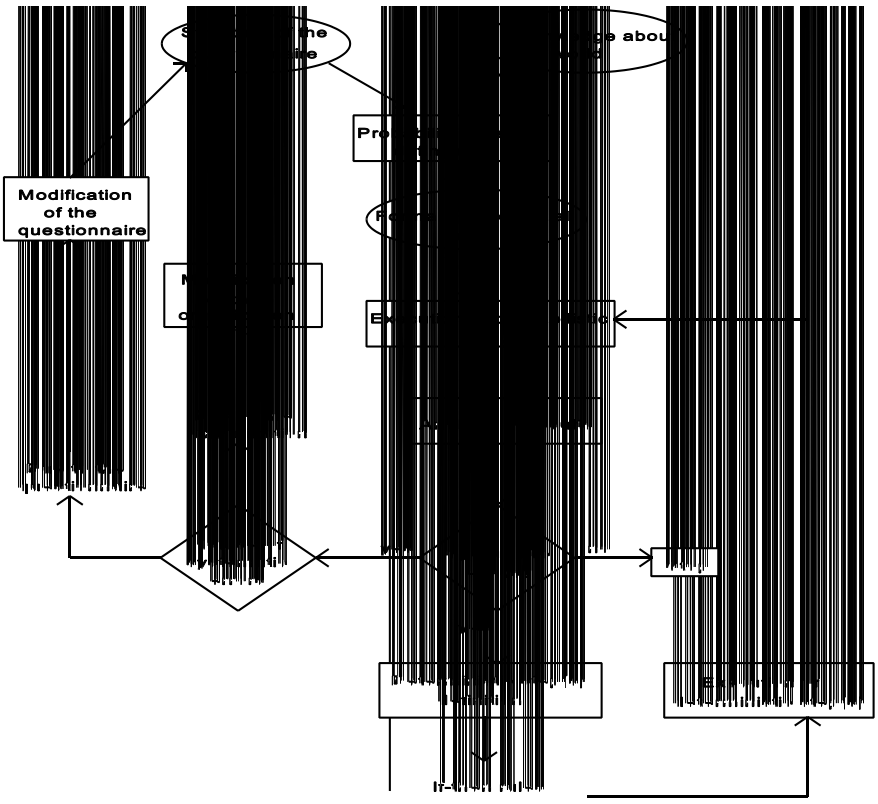
The probabilistic approach would ensure this result, but it can be applied only to stochastic errors. As the deterministic approach is likely to introduce distortions, *it must be minimized in the procedure*, i.e. we have to be sure that its application is limited to systematic errors, and their presence must be removed by using proper strategies in the long term.

In figure 5 the scheme of the proposed methodology is shown.

The first step is the *definition of edits in normal form*. Formal edits are defined considering the structure of the questionnaire, while substantial edits are derived through knowledge about relationships and constraints among objects in the portion of real world which is the target of the survey.

These edits (after being processed as introduced in par. 2.1) are applied to data, by executing the filter(s). At this point, an *analysis of imputation* is performed, in order to verify the correctness of edits and detect the presence of systematic errors.

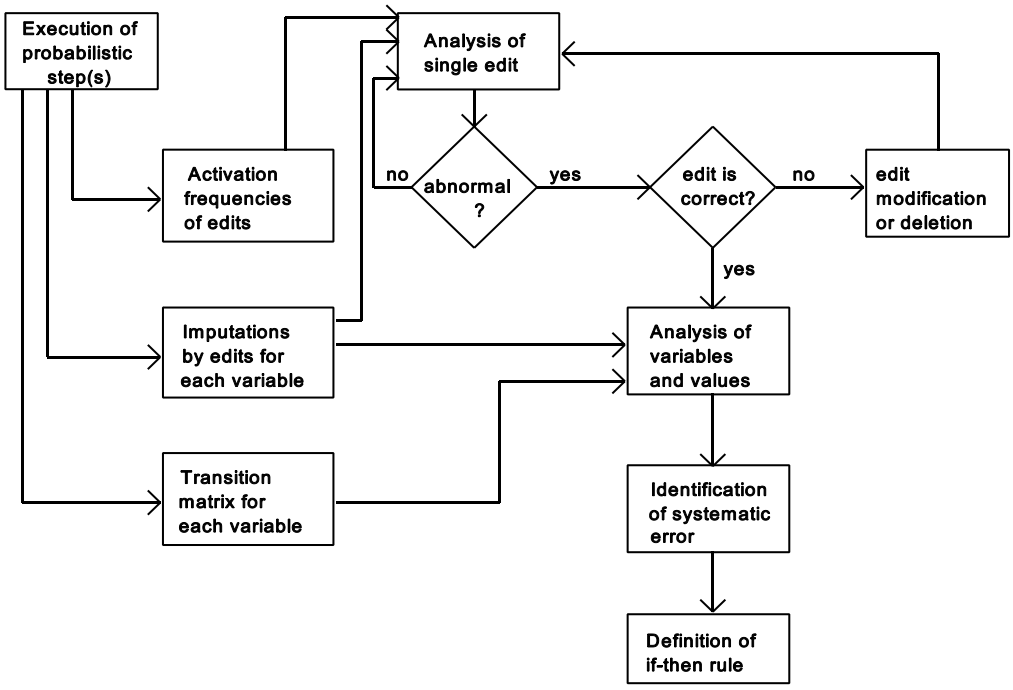
Figure 5. Methodological Scheme of the preparation of the edit and imputation procedure



The analysis (see figure 6) is performed by considering mainly:

a) the frequencies of edits' activation;
b) the imputations made on each variable, classified by the edits which caused the imputations;
c) the "transition matrices" linking values in raw data and values in clean data for each variable.

Figure 6. Analysis of results of imputation



Experience showed that there are two possible causes of the abnormal values of activation frequencies:

- the related edit is not correct (i.e. it does not represent a logical incoherence, but something acceptable) and therefore has to be removed or modified;

- the error identified by the edit is not stochastic, but systematic, and in this case has to be eliminated by using an if-then rule.

In the second case, the systematic error must be identified by considering the transition matrices of the imputed variables. When systematic errors affect a variable, its transition matrix generally presents a concentration of frequencies outside the main diagonal. The analysis of these concentrations can help to identify the cause of the systematic error in order to define the if-then rule that can eliminate it.

Once the analysis has been completed, if systematic errors were detected, a set of deterministic rules is defined. They must be applied to data *before* the application of probabilistic steps. Then analysis of results is performed again to see if other anomalies can be detected.

Consequently, in a short period the procedure is a mixture of both probabilistic and deterministic steps. In the long run the deterministic steps should disappear, as the structural causes determining systematic errors have to be eliminated. Analysis of these causes reveals shortcomings in the questionnaire design and/or in the organization of the data collection and storage. After these problems are solved, systematic errors should decrease. This can be monitored by considering the frequencies of activation of if-then rules.

The same strategy can be further developed in surveys that are particularly oriented to the production of "strategic aggregates", for example, the employment and unemployment rates computed on Labour Force data. In this case the aggregates can lead the process of analysis, as we are particularly interested in the variables that are used to calculate the most important aggregates. The basic idea is to compute all aggregates on raw data and clean data, and to measure the distance

between the two values. When the distance is too big, the analysis is oriented particularly to the variables which determine the given aggregate. Imputation is performed until the distance falls in an acceptable interval. This strategy has been followed in the Labour Force Survey.

4. APPLICATIONS

Until now DAISY has been applied to the two most important household surveys: the Labour Force survey and the Multipurpose Households survey. In both of them very complex edit and imputation procedures were needed. The results were satisfying. The quality analysis was performed for the old and new procedure of Labour Force survey [3], [4].

4.1 The Labour Force survey

The new Italian Labour Force survey (since October 1992) is a sample survey on a quarterly basis involving nearly 73,000 households and 200,000 individuals. The questionnaire is composed of the following sections:

- *section 1:* variables concerning the composition of households and the structural characteristics of individuals (related to head of household, age, sex, marital status, education);

- *section 2:* questions concerning the target of the survey: professional conditions, working activity, job search, training courses, etc. These questions are to be answered exclusively by individuals older than 14 years. The structure of this part of the questionnaire is rather complex: the completion of the numerous variables of this section (82) is conditioned by several compilation rules.

DAISY has been very useful in the phase of definition of formal edits. Based on only 59 structural rules, the system generated 402 formal edits in normal form. Together with 206 substantial edits defined by the statisticians, the total initial set comprised 608 edits.

Feasibility analysis showed the practical impossibility of generating implicit edits based on such a large initial set. The choice was made to divide the initial set into five subsets following the internal structure of the questionnaire. The subsets contain edits concerning:

1. the structural variables of section 1 (filter PROBAB1);

2. age and the most important variables of section 2: professional condition, working activity, job experiences, job search, professional training (filter PROBAB2);

3. variables related to working activity and job experiences (filter PROBAB3);

4. economic activity and profession (filter PROBAB3A);

5. variables related to job search (filter PROBAB4);

6. variables related to professional training (filter PROBAB5).

Table 1 presents the number of structural rules, formal and substantial edits and the dimension of complete sets related to the different filters.

The methodology introduced in par. 3 permitted the detection of a set of systematic errors which were mainly due to misunderstandings of the interviewers. Deterministic programs were developed to be applied before the execution of probabilistic steps.

Table 2 presents the results of a comparison of the old and the new edit and imputation procedure.

The new procedure generally gives better results, which is particularly evident in the case of the second aggregate. The quality of the procedure will improve as soon as the restricted joint donor imputation (which is now available in DAISY) is used for data imputation.

4.2 The Multipurpose Households survey

The last Multipurpose Households survey was held in December 1993, and involved nearly 20,000 households and 56,000 individuals. The questionnaire is very long. It contains 321 variables in the two main sections, the first compiled by the interviewers and the second by the respondents.

Table 1. Filters and edits in Labour Force survey edit and imputation procedure

Filters	Structural rules	Formal edits	Substantial edits	Complete set
PROBAB1	3	6	23	113
PROBAB2	3	167	0	567
PROBAB3	35	138	54	1020
PROBAB3A	1	1	96	97
PROBAB4	11	80	18	393
PROBAB5	6	10	15	35
Total	59	402	206	2225

Table 2. Comparison of the effect on main aggregates of old and new procedure

Aggregates	Old procedure (July 1992)		New procedure (April 1993)	
	Raw data (%)	Clean data (%)	Raw data (%)	Clean data (%)
Employed	37.37	37.70	35.37	35.88
Looking for a new job	1.37	0.80	1.62	1.51
Looking for the first job	2.43	2.34	1.88	1.69
Other persons looking for a job	1.32	1.51	0.85	0.75

The definition of 70 structural rules led to 354 formal edits, which, together with 79 substantial edits, determined an initial set of 433 edits. It was not possible to generate the implicit edits from such a large initial set, so it was divided into two: one set the comprised 239 edits concerning the first section (compiled by interviewer) and the other set comprised the 194 edits concerning the second section (self-compiled).

The first subset was still too big, and needed to be split again. The final composition of filters is shown in table 3.

Table 3. Filters and edits in Multipurpose Households survey edit and imputation procedure

Filters	Initial set of edits	Complete set of edits
PERSPRB11	116	564
PERSPRB12	123	147
PERSPRB2	194	194
Total	433	905

The analysis of the quality of the imputation is still in progress.

5. CONCLUSIONS

The first version of DAISY, currently used in the Italian National Statistical Institute, follows the probabilistic approach of the Fellegi-Holt procedure. From many aspects this is highly preferable to the deterministic approach, but it should be used only for the correction of stochastic errors. The presented methodology used in design and tuning of the edit and imputation procedure permits the maximisation of the probabilistic approach, and limits the deterministic one only to handling systematic errors. The characteristics of the first application of DAISY and of the methodology (Labour Force and Multipurpose Households surveys) have been introduced. New implementations of DAISY will concern deterministic steps and inter-record edit and imputation of data.

REFERENCE

[1] Barcaroli, G. An integrated system for edit and imputation of data in the Italian Statistical Institute, *Survey and Statistical Computing*, 1992, pp.167-177.

[2] Barcaroli, G., Di Pace, L. The automatic generation of statistical incompatibility rules from Entity-Relationship schemes, *Proc. of Seminar on New Techniques and Technologies for Statistics*, Bonn, February 1992.

[3] Barcaroli, G., Di Pietro, E., Venturi, M. La nuova indagine trimestrale sulle forze di lavoro: aspetti metodologici e analisi dell'impatto delle innovazioni introdotte sulla stima degli aggregati, *Politiche del Lavoro* n.22-23, Franco Angeli Milano, 1993.

[4] Barcaroli, G., Venturi, M. An integrated system for edit and imputation of data: an application to the Italian Labour Force survey, *Proceedings 49th session of the International Statistical Institute*, Florence, 1993.

[5] Fellegi, I. P., Holt, D. A systematic approach to edit and imputation, *Journal of the American Statistical Association*, 1976, vol.71, pp.17-35.

[6] Ford, B. L. An overview of Hot-deck procedures in *Incomplete data in sample survey*, Academic Press, New York, 1983, vol.2, p. 191.

[7] Friedman, J. H., Bentley, J. L., Finkel, R. A. An algorithm for finding best matches in logarithmic expected time, *ACM Transaction on Mathematical Software*, 1977, vol. 3, pp.209-226.

[8] Garcia Rubio, E., Villan Criado, I., Sistema DIA, Sistema de deteccion e imputacion automatica de errores para datos cualitativos, Instituto Nacional de Estadistica, Madrid, 1988.

[9] Kalton, G., Kasprzyk, D., The treatment of missing survey data in *Survey methodology*, 12, 1, Statistics Canada, 1986.

[10] Platek, R., Gray, G. B. Imputation methodology in *Incomplete data in sample survey*, Academic Press, New York, 1983, vol.2, p. 283.

THE 'SPEER' EDIT SYSTEM

By William E. Winkler and Lisa R. Draper, Bureau of the Census, USA

ABSTRACT

This document provides background on the workings and application of the system SPEER (Structured Programs for Economic Editing and Referrals) designed for ratio edits of continuous economic data. The first three sections consist of a description of how the system operates, and how it should be developed and maintained. Its strong points and limitations are also analysed. The fifth section presents the applications of SPEER system on two of the largest U.S. surveys of manufacturing and industries. Finally, an example is given, which shows the specific details of the input and output files used by the software.

Keywords: Fellegi-Holt, localization, imputation

1. INTRODUCTION

The SPEER edit system is the generalized software developed for the design of edits via creating ratio between couples of quantitative variables. The system utilizes the Fellegi-Holt model of editing. The first version of SPEER was written by Brian Greenberg [2], [3] and the current version was written by William Winkler [5]. The computational algorithms, much of the imputation methodology, and the source code in the current version is new.

2. DEVELOPING AN EDIT SYSTEM USING SPEER

There are three facets to the development, (1) analysis of the data using statistical and other packages, (2) development of a pre-edit system, and (3) development of a SPEER system. If data from a prior time period are not available, then data obtained during the collection can be used.

2.1 Stage 1: analysis of the data

Stage 1 proceeds with a variety of steps. The analyst begins by running various tabulations on the data to determine means, variances, ranges, and other values. Next, a regression package is run 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.

2.2 Stage 2: development of a pre-edit system

The pre-edit system 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.

2.3 Stage 3: development of a SPEER system

Stage 3 begins with determining the edit bounds for ratios. With appropriate test data, an auxiliary program D-MASO 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 applied SPEER software consists of two programmes. The first program, gb3.for, uses as the most important input the file of explicit edits defined by an analyst. Based on that, the program generates the logically implied edit bounds and checks the consistency of the entire edit system. It does not require test 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 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 consist of the failed edits, the minimum fields to impute, and the imputation methodology that was utilized for each field.

3. MAINTENANCE OF SPEER CODE AND SPEER SYSTEM

The code may not require any maintenance. If larger data structures are needed, then 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 SPEER system includes the documentation on how the "beta" coefficients from the regression are obtained. Based on that, a special program can quickly produce the set of "beta" coefficients.

4. LIMITATIONS AND STRENGTHS

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 find the variables that are linearly related and to determine the associated regression ("beta") coefficients 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 system 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.

The main advantage of the software is that it is very easy to apply. Only the format statement describing the locations and sizes of the quantitative data being edited needs to be changed [5]. In situations where storage does not exceed the default storage of the program, the 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 2000 variables.

The software is also 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.

5. APPLICATION

SPEER is currently being used in two large interactive applications: 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 centre in Jeffersonville, Indiana has shut down. Previously, late data were entered but generally

left unedited. Analysts can also perform additional 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 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.

LRPIES 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.

6. THE HARDWARE/SOFTWARE ENVIRONMENT AND DOCUMENTATION

The SPEER 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.

Three documents describe the overall SPEER methodology and capabilities. They are Greenberg and Surdi [2], Greenberg and Petkunas [3], and Greenberg, Draper, and Petkunas [1]. Winkler [4] describes how to develop and run a SPEER system. The details of the software and how to run it are covered in Winkler [5]. 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.

7. 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.

7.1 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,

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
```

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.

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.

7.2 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
     QPR3       11.429       4.000       6.682      21.092
     FBR4        6.000       6.000        .812      10.870
```

REFERENCES

[1] Greenberg, B. G., Draper, Lisa, Petkunas, Thomas. On-Line Capabilities of SPEER, presented at the Statistics Canada Symposium, October 1990.

[2] Greenberg, B. G., Surdi, Rita. 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, 1984.

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

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

[5] Winkler, W. E. SPEER Edit System, unpublished documentation, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C., USA, 1995.

THE 'DISCRETE' EDIT SYSTEM

By William E. Winkler and Thomas F. Petkunas, Bureau of the Census, USA

ABSTRACT

This document provides background on the workings and an application of the DISCRETE edit system. The system is based on the Fellegi-Holt model and applies to general editing of discrete data. 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.

Keywords: Fellegi-Holt; set covering; error localization; integer programming; imputation.

1. INTRODUCTION

The DISCRETE edit system is designed for general edits of discrete data. The system utilizes the Fellegi-Holt model of editing. Source code for DISCRETE was written by Winkler [3] and is based on theory and computational algorithms from Fellegi-Holt [1] and Winkler [4].

2. DEVELOPING AN EDIT SYSTEM USING DISCRETE

The data editing in the DISCRETE system is performed by two programs which operate in the following way:

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 with 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. As proved by Garfinkel, Kunnathur and Liepins ([2], hereafter referred to as GKL), the class of originally defined explicit edits plus the class of maximal implicit edits is sufficient for solving the error localization problem 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 [4] 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 [4] is significantly faster than an error localization due to GKL. While both error-localization algorithms use a series of cutting plane arguments to reduce the original integer programming problem, the algorithm of Winkler contains two modifications. The first modification is to use a greedy heuristic to first find a solution. A characterization allows determination of whether the greedy solution is the optimal one. If it is not, then the solution method goes to a general cutting-plane type of argument. Such arguments are generally known to be the most effective for solving integer programming problems. The main difference between Winkler [3] and GKL is that according to the first methodology the number of edits passed to the error localization stage grows at a much slower exponential rate. The bounding is due to a more precise characterization of the implicit edits needed for error localization [4]. 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 consists of the failed edits, the minimum fields to impute, and other information related to specific data records.

3. LIMITATIONS AND STRENGTHS

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, it is not known how large SETS of problems 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 it is presently written, the software is an early prototype version. Therefore the code is not sufficiently well organized and documented so that it can be maintained. Also 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.

The main advantage is that 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.

4. FURTHER DEVELOPMENT

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 table, 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.

5. 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 used to determine the rent control regulations for New York City. With previous edits, these fields were edited sequentially based on if-then-else rules. The sequential edits are often easily implemented but are not easily checked for logical consistency. For the AHS survey, another disadvantage is that there has to be an initial field (e.g. TENURE) from which the remaining fields will be edited. The initial field is never edited in the sequential edit application but can be edited 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. TENURE did hold a higher weight because of its response reliability. It would be a correct assumption that most respondents are aware of their living arrangement, therefore, TENURE is a very reliably reported field. However, there are other circumstances that would cause it to be incorrect. It still needed to be edited.

The explicit edits needed for DISCRETE were developed by combining the edits from 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 were 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 because of the complicated nature of the edit situation. 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 42 edits. Again, because of the use of the data for this survey was 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.

6. THE HARDWARE/SOFTWARE ENVIRONMENT AND DOCUMENTATION

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 only documentation associated with the DISCRETE edit system is Winkler [3]. The documentation is minimal and only describes how to compile and run the software on the example included with it.

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
```

7. 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.

7.1 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 k th variable takes values 1, ..., n_k, where the number of values n_k 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.

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.

The example of this section is a modified version of the example of GKL. The modification consists 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 generates 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 edits 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.

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	

7.2 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.

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
```

REFERENCES

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

[2] Garfinkel, R. S., Kunnathur, A. S., Liepins, G. E. Optimal Imputation of Erroneous Data: Categorical Data, General Edits, *Operations Research*, 34, 1986, pp. 744-751.

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

[4] Winkler, W. E. Editing Discrete Data, unpublished document, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C., USA, 1995.

GENERATION OF INSTRUMENTS FOR DATA COLLECTION AND INTERACTIVE EDITING

By Mark Pierczhala, National Agricultural Statistics Service, USA

ABSTRACT

For some agricultural surveys, the National Agricultural Statistics Service (NASS) has many versions of a questionnaire, sometimes one version for each of 46 state offices. An instrument-generating procedure has been developed for the case of one survey with multiple versions of questionnaires. This paper discusses an extension of the instrument-generating techniques to generate instruments that can handle two or more ongoing surveys, where each of the surveys may have many versions. Instruments are produced for both data collection and interactive editing. This programming research was conducted using the BLAISE system from the Statistics of Netherlands.

Keywords: generated questionnaires; multi-survey tool

1. INTRODUCTION

Agricultural surveys in the United States present a formidable challenge in the programming of electronic interviewing and editing instruments because agriculture changes greatly from state to state.

The version challenge is greater in forms-based systems than in question-based systems. In the former, the screen is used to portray multiple questions to some extent like they are printed on the paper form. In the latter, one question at a time is displayed on the screen. The forms-based approach is much better liked by the interviewers for a variety of reasons; one large reason being the navigational ability they have. In such a system you do not want to waste valuable screen real estate with large blocks of questions that are inappropriate to the state, nor do you want to be jumping around in seemingly random order over several pages. Thus the forms-based approach means that appropriate screen displays must be somehow obtained for each state separately. In a question-based system, where the question is the only thing being displayed, there is no corresponding consideration. The programmer is free to jump around from question to question without having to ever worry about a multi-question display. However, even in a question-based system, programming for 45 different versions of one survey is a tedious time-consuming task that is prone to error.

In order to implement a new forms-based system it was necessary to devise a new way of handling multiple versions. The discussion is particular to Blaise. This does not mean, however, that it could not be done in other ways in other systems. There are two major aspects of the Blaise system that make the generation of instruments relatively easy: 1) the block-based language with its dot notation for multi-level names, and 2) the fact that screens are automatically generated by Blaise during compilation (or preparation) based on the source code (i.e., you don*t personally draw each form that appears in the instrument). Among other aspects of Blaise III (the latest version of Blaise) that are very helpful are parallel blocks and parameters. Parallel blocks allow you to put part of a questionnaire on hold and maintain routing integrity while doing other parts. Parameters allow blocks of code to be programmed without having to know beforehand the details that will drive them in an actual interview.

2. GENERATING VERSIONS OF QUESTIONNAIRES

Within each QAS questionnaire are several standard sections such as crops, grain stocks, hogs, and chickens. The crops sections of the December QAS

offer the most extreme example of the version challenge, as demonstrated by the following list:

STATE A	STATE B	STATE C
Corn	Irrigated Corn	Corn
Soybeans	Non-irrigated Corn	Soybeans
Potatoes	Irrigated Sorghum	Alfalfa Hay
Sorghum	Non-irrigated Sorghum	Sunflowers
	Peanuts	Tobacco
	Potatoes	
	Rice	
	Irrigated Soybeans	
	Non-irrigated Soybeans	
	Cotton	

There are different crops in different states, some states distinguish between irrigated and non-irrigated crops while others do not, and the order of the crops differs from state to state. In addition, the different states gather different items of information about each kind of crop.

Other aspects of the version challenge that are not as apparent include difference in question text, units of production (e.g., pounds, tons, or bushels), SAS variable names, and unique item codes that have to be associated with each question.

2.1 Principles for Automatic Generation of Versions

Generating version of questionnaires is carried out according to the following principles:

- **Building upon similar structures**

Each structure has a number associated with it and the code for each structure is held in a file with the number for its name.

- **Breaking down the version challenge into manageable parts**

The necessary strategy is to separate aspects of programming that change from crop to crop from those that do not. In Blaise this separation is made neatly. The things that do change from crop to crop appear mostly in the FIELDS paragraph, while the things that stay the same appear in the RULES paragraph. Thus it is necessary to generate a separate FIELDS file for each crop while it is possible to have only 17 different RULES files on hand that are applied appropriately to each of the separate FIELDS files.

- **Customization of Blocks**

It remains to customize each block as regards edit limits and other things as well. In Blaise III this is done in the RULES paragraph at the block level (lowest level of organization possible)

In practice the variable edit limits are calculated from an external file as they change from state to state for the same crop. Automatically generating code eliminates tediousness, is time effective, and less subject to error.

2.2 Requirements for Automatic Generation of Versions

Three major requirements for automatically generating versions of instruments are a library of programmed code, a parameter file of specifications, and a program that generates the versions of instruments. The library serves as a resource of coded segments that the generator program assembles according to the parameter file of specifications.

The concept of a library is very important. Breaking up the Blaise coded segments as described creates potentially hundreds of small files where there were just several larger ones before. They cannot all be placed in one directory. They must be placed in a simple subdirectory tree whose structure is readable and understandable to those who must work with it.

2.3 Parameter File of Specifications

The parameter files of specifications must contain pertinent information for each state. Ideally there would be one parameter file where each state*s specifications would be contained on one line. NASS*s specification file is a Blaise III data set. Specifications are entered into this subsidiary Blaise instrument especially designed for easy entry of specifications. It is designed so that a clerk can fill in the proper information. This external file embodies a very powerful idea, that the state-specific code is held in a place where it is easily manipulated and maintained by non-programmers.

2.4 Generator Program

The program that generates the versions from the library of code and the parameter file needs to be able to read information from the parameter file and draw upon the code from the library. NASS has used Manipula (the tool of Blaise aiming to manipulate data sets) as it can read Blaise data sets directly and produce ASCII files as a result of what it reads. The sequence of steps follows: 1) the generator program reads a line from the parameter file, 2) lines are written into a text

file by the generator program, (these lines refer to code from the library), 3) the instrument is built up from a standard front part, the text file of library references, and a standard ending part, 4) the Blaise program is prepared by each state, and 5) the prepared program is copied into the proper production directory.

Each state generates its own instruments. Headquarters provides all necessary specifications files and all the Manipula programs and library. Since each state now generates its own instrument, it is possible to generate only the FIELDS paragraphs that the particular state needs as the instrument is being built. If problems are found in the instrument during production, then the instrument is not changed. Rather, the generator program or the library is fixed and the instrument is regenerated.

Surveys where the instrument generation approach is or will be implemented include the QAS and the state-based Acreage and Production (A&P) Surveys which are used to obtain county-level estimates. In the QAS the specifications are set in headquarters. In the A & P Surveys, the specifications are set in the state office.

3. ONE GLOBAL INSTRUMENT WORKING 45 DIFFERENT WAYS

There are surveys in NASS where it is better to produce one global instrument and make it perform 45 different ways depending upon an external specification. Among these are the June Area Frame Survey (JAF) and the Chemical Use Survey. Most work so far has been done with the JAF.

3.1 June Area Frame Survey

Work is almost completed for the JAF. There is one electronic instrument that can be driven 45 different ways based on an external file of specifications. This involves much more than just setting flags for whether a question should be asked or not. The external file contains a large amount of information including question text, item codes, SAS VAR names, and other meta information. The external file itself is based on national-level specifications especially for crops and grain stocks.

From 1,200 questions specified on national and state levels, 8,000 questions are handled by the global instrument.

TOGGLES is the central source of specification for one state where a crop (or stock) can be chosen. Meta data from the CROPS (or STOCKS) national data base is then transferred into the TOGGLES data base. This prevents a lot of retyping of the same information. TOGGLES includes question-by-question toggles, item codes and SAS variable names for appropriate questions, some question wording, edit limits, and for the crops, a state-specific screen design for interviewers in CAPI. In all, there are about 1,200 questions and work cells in TOGGLES.

Under this method of instrument production, a question (or cell or code) in the crops or stocks part of the questionnaire can have one meaning in one state and another meaning in a second state. Thus the instrument itself does not hold all meta data. When reading data into or out of the instrument, or when flashing questions on the screen for the interviewer, meta data from the external TOGGLES must also be used. Again, this worked well in Blaise III. The external data models CROPS, STOCKS, and TOGGLES are all Blaise III data models. They are set up so that non-programmers can state the specifications without having to get into the source code. One advantage of this was seen in Pennsylvania and Wyoming in 1995 when it turned out that specifications were wrongly stated. The state personnel were able to invoke the CROPS and TOGGLES data models and correct specifications which allowed them to proceed without re-preparing the production instrument.

In 1995 the JAF instrument was set up for all states and was tested in 3 of them. Due to laptop memory constraints, one instrument was used in Indiana for Computer Assisted Personal Interviewing (CAPI) and for Interactive Editing, while another was used in Pennsylvania and Wyoming for only Interactive Editing after data were collected on paper questionnaires. The original goal was to use the same instrument for all three states (or for all states). This could have been done if the laptop computers in Indiana had more than 4 MB of memory. However, in principal, we proved that this could be done and in 1996 expect to have one instrument for all states for the JAF regardless of the use of the instrument.

3.2 Generated Instruments Versus One Global Instrument

Whether it is better to generate an individual instrument for each state or provide a global instrument that is driven different ways by an external file must be judged on a survey-by-survey basis. Technically it is possible to use either technique on all of NASS*s

agricultural surveys. The following table gives a list of pros and cons of each technique.

There are various trade-offs in the selection of the method of handling versions. The QAS could use the one global questionnaire approach but would probably need a tabular presentation for crops and stocks to do so. This is because of the way Blaise presents tables as opposed to other displays. By using tables, much of the concern about navigation is reduced.

4. MEGA-VERSION INSTRUMENTS

4.1 Background

One aspect of NASS*s survey program, that is perhaps shared by other agencies that do economic surveys, is that there are some farms that are sampled often throughout the year. In fact, in it*s quarterly and monthly survey programs, it is possible that a farm can be sampled for 2 or 3 surveys at once. Usually it is one person that is interviewed for all surveys. In such a case on paper the interviewer takes all appropriate forms to the farmer at one time and just plows through all of the appropriate questions.

TECHNIQUE OF MANY GENERATED INSTRUMENTS	TECHNIQUE OF ONE GLOBAL INSTRUMENT USED IN DIFFERENT WAYS
Crop blocks (or grain blocks) are all uniquely generated. As a result: - All meta data are contained within each instrument. - Unnecessary questions do not appear in the crop*s (stock*s) block. Thus there is less 'air* in the data set. - Since all blocks are uniquely defined, the prepared instrument is larger than if one block was programmed and used many times in different ways as a macro or subroutine.	There is one crop (stocks) block that is used in different ways. As a result: - Meta data for some sections are contained in external files. This includes question text, SAS VAR names, item codes, etc., which must then be read into the instrument. - Unnecessary questions are in the block (for any crop) and must be routed around. Thus there is more 'air* in the data set. - The instrument is smaller than if many uniquely defined blocks were included because the formal macro capabilities of Blaise are used.
Is used more often in questionnaires where large tables are not employed.	Is used primarily in personal interview surveys which feature large instruments with huge tabular arrays.
Allows use of a structure appropriate to each crop*s questions	Uses one (or a few) overall crop structure(s) regardless of which questions will be asked.
Distribution of files from headquarters may be a major challenge if instruments are generated in headquarters, however if state generate their own instruments, this is less of a problem.	Only one instrument and external file are distributed from headquarters to each state.
Each state needs its own customized program to read data in and out of Blaise. This too is automatically generated.	Each state uses the same read-in and read-out program which is driven by the same external file used to drive the instrument itself.
Mistakes in the specification usually require regeneration and re-preparation of the instrument. An exception is edit limits.	Mistakes in the specifications can usually be repaired directly in the specifications file without re-preparing the instrument.

In phone surveys, where one farmer is interviewed for more than one survey, the first one is done in CATI then all remaining ones on paper. Since the farmer is questioned over the phone he doesn*t detect any switch in modes. The drawback is that now some data must be captured and edited after the interview instead of during the interview. In CAPI, the situation is worse in the sense that we take this expensive laptop out to the field, collect data on computer in full view of the farmer for one survey, then put it away so the other interviews can then be done on paper. A mega-version instrument would be one where all appropriate survey questions can be put into one instrument and where the interview would flow smoothly from one survey to the next.

The mega-version approach was first proposed in 1993 at the Annual Research Conference of the Bureau of the Census. This paper presented some thinking about what such an approach would entail.

An important prerequisite for this method is to have a top-level commitment for its usage and thorough planning well in advance. Even in a powerful system, such as the new Blaise III which provides the technical capability and capacity, mega-versions cannot be expected to happen automatically.

4.2 Description of Mega-Versions

Mega-version instruments would handle two or more surveys at one time. Consider the following situation. Surveys A, B, and C are held within a time frame, for example Month 1 of the current year. One or more of Surveys A, B, and C may have multiple versions across states. In the mega-version scenario presented in 1993, instead of Instruments A, B, and C, there would be a Month1 Instrument that would handle Surveys A, B, and C. In effect there would be multiple versions of a time-frame instrument, one for each state. A respondent would only be required to answer those questions for which he was sampled. Thus if he is in survey A only, then he would just get survey A questions, not the questions for surveys B or C. If he is in all surveys, then he would get all appropriate questions, without redundancy. The mega-version approach would be a respondent-centric approach as opposed to a survey-centric approach now used by almost all agencies. It would require some changes in the organization to manage the survey preparation, survey collection, and data movement.

4.3 Benefits

The respondent, the interviewer, and the survey organization all stand to benefit from the mega-version approach. For the respondent, questions that are common across surveys are asked only once. The interview proceeds smoothly between all surveys. Overall time of interviewing may decrease but should not ever increase.

The enumerator benefits in 3 ways. First, the interviewer does not have to jump out of one CATI or CAPI instrument onto paper or into another CATI or CAPI instrument. Not only does this save time, it also reduces the need to rely on experienced enumerators for these kinds of interviews to know where to jump to on the paper form. As well, edits and computations can now be defined between different surveys. If there is a conflict between answers in Surveys A and C then the enumerator can jump to either part of the one instrument to resolve the problem (and get back to the interviewing point very easily).

Second, CATI management would be eased. In CATI, all appropriate surveys are handled in one call scheduler. Parameters are manipulated in the call management file in order to control when the interviews for each survey appear. In some states respondents are notified of a survey (or surveys) by mail and are given a toll free number to call the office to respond to the survey. If all appropriate surveys are held in one call scheduler the interviewer could just ask the respondent for his phone number, call up the form, and immediately be guided through the appropriate questions.

Third, CAPI interviewers avoid some embarrassment. For example, the interviewer would be looked upon as inefficient if she carried a laptop onto the farm, performed one interview on the laptop, then jumped to paper for the second and third surveys.

The organization would benefit from increased productivity in interviewing and data processing, from explicitly managing respondents as well as surveys, by encouraging standardization between surveys, and by ensuring consistent data between surveys.

4.4 Extension of Generating Tools to the Multi-Survey Realm

It will not be enough to just paste code from different surveys into one instrument; there may be additional code that holds edits and computations that relate one survey to another. There will also be some additional routing code that avoids asking duplicate questions between surveys. To produce the

instruments, the electronic specification data base discussed above would be enlarged to encompass all eligible surveys. The library would be extended with additional directories holding additional files of tested code. Part of the specification program will coordinate how questions from two or more surveys should agree with each other through edits. These are called concordance questions or blocks of questions. Some of this additional concordance code can be anticipated and stored in the library. Some cannot and must be generated dynamically according to pre-specified rules.

Representatives from across the organization will have to come together to decide how all these questions should appear in the instrument, and which questions are redundant of each other. Some harmonizing of questions or blocks of questions may be necessary.

4.5 Challenges

The nature of the survey organization, the inherent technical nature of the various surveys, and the limitations of software and hardware all give rise to challenges in producing and using mega-version instruments.

From an organizational perspective, there would have to be a highly visible mandate from senior managers and then cooperation on many levels between disparate units, starting with survey design and procedures. Every one will get what they need and almost all of what they want but may have to compromise on details. For example, placement of sections, designation of item codes, priorities for the call scheduler, all need to be sorted out formally before the programming begins.

NASS uses item codes for many purposes, including high-speed data entry. Each question has an item code assigned to it, and once that assignment is made, it is set in concrete. This item code assignment is done for each survey independently and the same item code can be used for different items in different surveys. NASS refers to item codes in Blaise when reading in survey data from the standard data entry package (Key Entry III). Where these item codes are not unique within the mega-version instrument as a whole they can easily be made unique for each major section within the instrument.

Another challenge is that surveys must be managed according to their own schedules. Therefore data from one survey must be read out into edit, analysis, and summary before data from another survey. Here again, data readout and conversion can be handled by major section of the questionnaire. In the call scheduler, it is possible to influence the delivery of forms to the enumerators, however some decisions will need to be made about what to concentrate on at different times during the surveys. It will likely be each state individually that makes the decisions, thus processing flows must allow some flexibility.

Screening questions for each survey differ. For the software there is no technical problem, the planners of the instrument will just have to make sure that the needed questions are asked of the appropriate respondents.

A larger challenge is that even within one survey, a farmer may have to respond on behalf of 2 or 3 different farming operations or tracts of land. For example, in the JAF where tracts of land are the unit of interview, a farmer may operate land in 2 or 3 of these tracts. This was handled by rostering some sections (i.e., rosters of rosters) twice in the instrument, including a large table of field usage (about 2,000 questions and a like number of edits in each manifestation of the table). This is actually a variation on the mega-version theme and might be called a mega-instrument. This worked well for the interview, however, it did complicate data management and survey accounting.

All of these complications can be handled by planning and clever approaches to rostering and generation. They are stated here for the benefit of those who will embark on this kind of project in the future. The generation techniques described above are very powerful and not very hard to implement. They allow you to think of and accomplish agency goals such as mega-versions or customizing an interview per respondent. For a more detailed description of mega-version instruments and how they would be put together see Pierzchala [2].

REFERENCE

[1] Pierzchala, M. The 1995 June Area Frame Project, *Proceedings of the Third International Blaise Users* Conference*, Helsinki, Finland, Statistics Finland, 1995.

[2] Pierzchala, M., Computer Generation of Mega-Version Instruments for Data Collection and Interactive Editing of Survey Data, *Proceedings of the Annual Research Conference*, Washington, DC: U.S. Bureau of the Census, 1993, pp. 637-

645.

[3] Pierzchala, M. Generating Multiple Versions of Questionnaires, *Proceedings of the First International Blaise Users Meeting*, Voorburg, Netherlands: Statistics Netherlands,1992, pp. 131-145.

SURVEY OF GENERALIZED SYSTEMS USED FOR EDIT SPECIFICATIONS

By Giulio Barcaroli, National Statistical Istitute, Italy

ABSTRACT

Every member of the UN/ECE - Data Editing Joint Group has been required to answer some questions for any software system currently used for edit and imputation of data in their National Statistical Organization: identification (name) of the system, field of application, if generalized, together with indications about the user profiles and the characteristics of embedded languages. Information has been given by the following countries: Austria, Canada, Czech Republic, Germany, Greece, Hungary, Netherlands, Poland, Slovenia, Sweden, United Kingdom and United States (Bureau of the Census and NASS). Information was already available on Italy and Spain. In this short note we try to give a general overview, particularly of the characteristics of languages, reporting the examples given by respondents.

Keywords: generalized software; statistical application; user language

1. SYSTEMS IDENTIFICATION AND FIELD OF APPLICATION

The following generalized systems have been indicated:

- GRAN78 (Sweden): also known as EDIT/78, is a mainframe system for data editing, batch and on-line as well;
- RODE/PC (Sweden): used for data entry and interactive editing;
- DATAMAN (Czech Republic): used for data entry and interactive editing;
- GEIS (Canada): an edit and imputation integrated system applicable to numeric, non-negative and continuous data;
- DC2 (Canada): a large-scale data collection and capture system, applicable both to qualitative and quantitative data, particularly for interactive editing;
- NIM (Canada): an edit and imputation system based on minimum change donor technique for both qualitative and quantitative variables;
- QUANTUM V5.6 (Greece): used for editing and tabulation of data in a WINDOWS environment;
- G/VS (Slovenia): a subsystem of GODAR SYSTEM, used for the interactive editing of both qualitative and quantitative variables;
- LINCE (Netherlands): system for automatic editing of qualitative data based on Fellegi-Holt methodology (from Statistical Bureau of the City of Madrid, Spain);
- BLAISE (Netherlands, Hungary, NASS): the well known system for data capture and processing;
- SPEER (U.S. Bureau of the Census): generalized system for automatic or interactive editing of quantitative variables;
- DISCRETE (U.S. Bureau of the Census): generalized system for automatic editing of discrete data, developed on the basis of Fellegi-Holt methodology but with different algorithms for the generation of implicit edits and for imputation of data.

All these systems are generalized.

Other countries, especially Austria, Ireland, Poland, described methods and procedures currently used at a fairly detailed level. Information was already available about DIA (Spain), DAISY and SCIA (Italy).

2. USER PROFILES

With regard to user profiles, the following situations occur:

1. Statisticians are not directly involved in implementation of editing procedures, i.e. they

give specifications and informaticians develop the corresponding application;

2. Statisticians co-operate with informaticians in the definition of editing and imputation rules with direct access to generalized systems;
3. Statisticians use generalized systems directly with no need of informaticians.

The first situation is typical of traditional applications, in which no generalized system is available: editing specifications are transmitted to programmers in natural language, and the latter must translate them in a language processable by the computer.

The last two situations occur when generalized and advanced systems are available; the second is much more frequent than the third.

As examples of the first case, we can mention Poland, Austria, Ireland and Sweden. In Poland editing applications are specified by statisticians, who prepare tables whose rows represent editing rules with a given syntax: informaticians take these specifications and prepare prototypes that are tuned together with the statisticians. In Sweden, RODE/PC applications are created by EDP professionals.

Sweden (GRAN78 is used both by data processing personnel and subject-matter specialists), Canada (in DC2 statistical methodologists consult subject-matter statisticians in order to give specifications to informaticians) and Czech Republic (users of DATAMAN are mainly informaticians, but many statisticians are able to use it) are examples falling into the second case. U.S. Bureau of the Census software, i.e. DISCRETE and SPEER, until now required computer expert intervention as data definition is contained inside the source of FORTRAN programs.

Finally, edits in GEIS are directly specified by statistical methodologists and subject-matter statisticians, so we can consider this as a pure example of the third case. Also Italian software DAISY and SCIA should allow statisticians to define and apply editing rules without the need of computer programmers. Experience has showed that in the initial applications the support of informaticians is required.

3. APPLICATIONS

The following information concerns the use and applications of some of the cited systems:

DATAMAN

There are three versions of DataMan, numbered from 1 to 3. As far as the first version is concerned, in 1994 80% of surveys carried out internally by the Czech Statistical Office used DataMan1. The second version, DataMan2, was developed in 1994 and used in 1995 and was applied to 85-90% of the surveys. Finally, the third version, DataMan3, whose use is planned for 1996, will be applied to 95% of all surveys.

DC2

Over one hundred Statistics Canada applications have been designed in DC2, mostly for Data Capture. The following surveys use, or are being redesigned to use, DC2:

the Survey of Employment, Payrolls and Hours;
the Census of Agriculture;
the Annual Survey of Manufactures;
many others in areas such as Education, Culture, Services and Health.

GEIS

More than twenty surveys have used GEIS, including the Census of Agriculture, the Survey of Employment, Payrolls and Hours, the Annual Survey of Manufactures, the Annual Wholesale and Retail Trade Survey, the Survey of Labour and Income Dynamics, and various agricultural and transportation surveys.

GODAR G/VS

GV/S has been applied to about 50 surveys at the Statistical Office of Slovenia. The most important are: Monthly Report on Industry, Annual Report on Industry, Natural Increase and Decrease of Population, Migration of Population, Survey on Agricultural Enterprises, Survey on Number of Livestock, Report on Hotel Trade, Monthly Report on Constructions, Survey on Primary, Secondary and High School Teachers.

NIM

The NIM is being used to carry out editing and imputation in the 1996 Canadian Census for the demographic variables

QUANTUM

In the National Statistical Service of Greece, the software QUANTUM for data editing and automatic correction has been applied to the Labour Forces Survey and to the Vital Statistics (Births, Marriages, Deaths).

RODE/PC

The software RODE/PC is largely used inside Statistics Sweden, as it is currently applied to about

100 surveys.

SCIA and DAISY

In the Italian National Institute of Statistics, the two generalized systems have been applied essentially to surveys concerning households and persons. The most important are the Census of Population and Dwellings, Labour Force Survey, Multipurpose Households Surveys, Survey on Households Expenditures, Survey on Marriages, Survey on Births, Survey on Graduates Occupation.

4. LANGUAGES FOR EDITS DEFINITION

4.1 Major characteristics

Edits in **GEIS** are specified by simply connecting variables with mathematical operators: a little knowledge of commands and syntax is needed, as the system is menu-driven. For example:

```
PASS: FOOD + BEVERAGES = TOTAL_SALES
PASS: MILK_LITRES >  15 * MILK_COWS
FAIL: X1 + X2 + X3 < 10
```

(the meaning of these edits is evident).

In **DC2**, on the contrary, specification of rules is made by using a subset of PROLOG, ESL, that is not so user-friendly:

```
cons_item1(fields(Start,Buy,Lose,Sell;End),pass,_,_,_):
-Start+BuyLoseSell=:=    End
cons_item1(_,nopass,cons_item,msg_values(item1),3)
```

(this consistency edit checks that the sum of the variables Start and Buy, minus Lose and Sell, equal End).

In **NIM**, edit rules are specified using Decision Logic Tables (DLTs), which follow the syntax of SPIDER, i.e. the front end editing system, requiring a RAPID data base organization. The following is an example of conflict rule (age check between the head of household and his/her father/mother-in-law): if a household matches its conditions, it fails the edit:

```
H    NAME(DE6PAC) TYPE© 6,2,2
C    R2P12BU(02) = CLASS(PERSON1S_SP)      ;Y;      ;
C    R2P12BU(#1) = FATH_MOTH_INLAW         ;Y;      ;
C    R2P12BU(#2) = FATH_MOTH_INLAW         ;Y;      ;
C    AGEU(#1) - AGE(02) < 15               ;Y;      ;
C    AGEU(#2) - AGE(02) < 15               ;Y;      ;
C    R2P12BU(#1) = FATHER_MOTHER           ;Y;      ;
C    R2P12BU(#2) = FATHER_MOTHER           ;Y;      ;
C    AGEU(#1) - AGE(01) < 15               ;Y;      ;
C    AGEU(#2) - AGE(01) < 15               ;Y;      ;
```

The table DE6PAC represents a generic table which is replicated by SPIDER to create 10 tables: each table represents one of the 10 possible positions in which a Father/Mother-in-law pair could appear in a six person household.

Edits are specified in **BLAISE** by using a Pascal-like language. This task requires training of users:

```
Datamodel Person
...
Rules
      Name
      BirthYear
      Age := Year(SysDate) - BirthYear
      Age.Show
      MaritalStatus            if Age < 15 then
       MaritalStatus = Single "^Name is too young to be
married"
      endif
EndModel
```

GRAN78 makes use of a keyword-based language:

```
 VALID   SVAR=(1:9);
         MONTH=(1:12);
         DAY=(1:31);
TRANSGEN NEW = VAR+VARX;
GRUPP  TEST1=0(NEW>15)
         TEST2(MONTH,DAY)=1(4|6|9|11,>30),1(2,>29);
TEST         TEST2(SVAR,VARX);
```

RODE/PC: The Language is the *Advanced Validation Language* (AVL), that can be considered, like BLAISE, as another PASCAL-like language:

```
FIELD_END var2;
BEGIN
      IF OVERRIDE=2 THEN
      ACCEPT;
          IF (var2=2) AND (var1=1) THEN
          REJECT (var2, '#You cannot enter a 2 when var1 is 1');
END.
```

DATAMAN from Czech Republic belongs to the family of PASCAL-like languages.

The **GV/S** system from Slovenia, on the contrary, has peculiar characteristics. For example:

```
ERRN(001) = (KONSK(FILE1,A) & B>75) ! (C<10) ;
```

where KONSK is a function in GV/S language, FILE1 a consultant file, A, B, C are variables and ERRN(x) a boolean representing a given error. The meaning of this statement is the following: "If A is in FILE1, and

is true that B is greater than 75 and C is less than 10, then error #1 occurs".

In **SPEER** an edit is defined as a *ratio* between a couple of quantitative variables, bounded by a lower and an upper limit. An example of the syntax is the following:

```
110  1  1  2  .0212400  .0711125  .0369900  EMP1/APR2
```

whose meaning is the following: in data group characterized by a '110' code, the #1 edit is a ratio between variables #1 and #2 (respectively, EMP1 and APR2), whose lower limit is '.0212400' and upper limit is '.0711125', with a central value of '.0369900'.

The family of systems for edit and imputation of qualitative data according to **Fellegi-Holt methodology (DISCRETE, LINCE, DIA, DAISY, SCIA)** has very similar languages to define the so-called *edits in normal form*.

In **DAISY**, for example, the statement:

```
AGE(MIN TO 14) MAR_STATUS(NOT 1)
```

represents an edit in normal form whose meaning is "if the age is less than 15 and marital status is different from 1 (single), than an error occurs". This is equivalent to SCIA notation:

```
AGE(0-14) MAR_STATUS <1)
```

or to DIA expression:

```
AGE(L15) MAR_STATUS(\1)
```

The system **DIA** allows also to define *rules for deterministic imputation*, like the following:

```
VAR1(1-3) VAR2(3) VAR3(b) = VAR3 ( = VAR1 + 2*VAR2 )
```

whose meaning is: " if the conditions in the left-hand member are verified, then compute VAR3 as specified in the right-hand member".

As an exception, **DISCRETE** has a rather different syntax. See, for example, an edit with the same meaning of the previous ones:

```
Explicit edit #1: 2 entering field(s)
  AGE            15   response(s) : 0 1 2 3 4 5 6 7 8 9 10
                                     11 12 13 14
  MAR_STATUS     4    response(s) : 2 3 4 5
```

4.2 Possible classification of languages

In this short note we have tried to give an overview of the different languages for edits definition, used in the National Institutes that required to our survey. It seems that they can be classified in different groups:

1. Keyword-based languages: for example, GRAN78 and DATAMAN;
2. Languages with characteristics of already existing commercial languages, as in the case of the languages adopted by BLAISE and RODE/PC (PASCAL-like), and DC2 (PROLOG-like);
3. Declarative languages: this is the case of any system following the Fellegi-Holt methodology, that requires the definition of the only validity or error condition, with no indication of how to solve imputation problems.

Languages of the first class are perhaps the most difficult for statisticians to use directly and can be handled more properly by programmers.

Declarative languages are the most appropriate for subject-matter experts: actually, they are embedded in systems developed for direct use by these experts (with some exceptions, for example, SPEER) .

Second group languages seem to be intermediate, as their syntax requires specific training, but they are not directed specifically at computer programmers for use.

In both second and third group languages, a great variety of user-friendliness (degree) can be found, and the general trend is clearly in this direction, as the characteristics of most recent systems prove.

5. SUMMARY TABLE

Country	System	Purpose	User profile
Austria	ad hoc applications		survey staff give specifications to informaticians

Canada	DC2	large scale data collection and capture system	methodologists consult survey staff to give specifications to informaticians
Canada	GEIS	edit and imputation system applicable to numeric data	methodologists and survey staff
Canada	NIM	edit and imputation system applicable to qualitative and quantitative variables	survey staff with some help from methodologists and informaticians
Czech Republic	DATAMAN	used for data entry and interactive editing	mainly informaticians but also survey staff
Greece	QUANTUM V5.6	editing and tabulation of data	informaticians
Ireland	ad hoc applications		statisticians give specifications to informaticians
Italy	DAISY	edit and imputation system applicable to qualitative variables	survey staff and methodologists
Italy	SCIA	edit and imputation system applicable to qualitative variables	survey staff and methodologists
Netherlands, Hungary, NASS	BLAISE	data capture and processing system	survey staff and informaticians
Netherlands (from Statistical Bureau of Madrid)	LINCE	edit and imputation system applicable to qualitative variables	survey staff and methodologists
Poland	ad hoc applications		survey staff give specifications to informaticians
Slovenia	G/VS (GODAR)	interactive editing of both numeric and discrete variables	informaticians
Spain	DIA	edit and imputation system applicable to qualitative variables	survey staff and methodologists give specifications to informaticians
Sweden	GRAN78 (EDIT/78)	batch and on-line data editing	survey staff and informaticians
Sweden	RODE/PC	data entry and interactive editing	survey staff give specifications to informaticians
U.S. Bureau of the Census	DISCRETE	edit and imputation system applicable to qualitative variables	survey staff give specifications to informaticians
U.S. Bureau of the Census	SPEER	edit and imputation system applicable to numeric variables	survey staff give specifications to informaticians

DOCUMENTATION

BLAISE

BLAISE III Reference Manual, Statistical Informatics Unit, Statistics Netherlands, 1994.

BLAISE III Developer's Guide, Statistical Informatics Unit, Statistics Netherlands, 1994.

DATAMAN

DataMan, Referencni prirucka (Reference Manual, in Czech only).

DataMan, Uzivatelska prirucka (User's Guide, in Czech only).

DataMan2, Referencni prirucka (Reference Manual, in Czech only).

DataMan2, Uzivatelska prirucka (User's Guide, in Czech only).

DAISY

G. Barcaroli, C. Ceccarelli, O. Luzi. Una metodologia di editing e imputazione per variabili qualitative, 1994, (in Italian only).

DC2

DC2 Concept and Facilities Release 1.0, Generalised Survey Function Development (GSFD), July 1992.

DC2 R1.5 Sample Verification Facilities Addendum, GSFD, Oct. 1992.

DC2 R1.5 Export Facilities Addendum, GSFD, Oct. 1992.

DC2 R1.5 Edit Specification Language Reference Manual, GSFD, Oct. 1992.

DC2 R2.0 Navigation and Inter-Field Communications Facilities Addendum, GSFD, March 1993.

DC2 R2.0 System Messages Manual, GSFD, April 1993.

W. Mudrik, J. Croal, B. Bougie. DC2 R2.0 Interviewing and Appointment Management Facilities Manual, July 1994.

W. Mudrik, B. Bougie. DC2 R2.0 Sample Verification Quality Control Procedures Manual for Supervisors and Verifiers, July 1994.

DIA

E. Garcia Rubio, Avalon Criado. SISTEMA DIA - Sistema de detecciòn e imputaciòn automatica de errores para datos cualitativos, 1988, (in Spanish only).

GEIS

J. G. Kovar, J. H. MacMillan, P. Whitridge. Overview and Strategy for GEIS, February 1991.

C. Cotton. Functional Description of GEIS, July 1991, revised August 1993.

S. Auger, J. M. Fillion. Tutorial Introduction to GEIS, March 16, 1992.

I. Schiopu-Kratina, J. G. Kovar. Use of Chernikova's Algorithm in GEIS, January 1989.

S. Legault, D. Roumelis. The use of GEIS for the 1991 Census of Agriculture, November 1992.

GODAR GV/S

GODAR Vega/Stat System. Data Editing Working Session, Stockholm, 1993.

GV/S Reference Manual, in Czech only.

GV/S User's Guide, in Czech only.

GRAN78 (EDIT/78)

Description and Features of Selected Data Editing Software Systems, SCP/DA/WP.76, 1984, pp.51-54.

LINCE

LINCE Reference Manual, in Spanish only.

LINCE User's Manual, in Spanish only.

NIM

M. Bankier, M. Luc, C. Nadeau, P. Newcombe. Imputing numeric and qualitative variables simultaneously, March 1996.

QUANTUM

Analytical User Manual, QUANTIME LTD, London.

Reference Guide, QUANTIME LTD, London.

RODE/PC

RODE/PC User Guide Release 3.0. VM-Data Dataservice, Malmoe, Sweden.

SCIA

G. Barcaroli, C. Ceccarelli, O. Luzi, A. Manzari, E. Riccini, F. Silvestri. The methodology of editing and imputation of qualitative variables implemented in SCIA (VM/CMS version), ISTAT, Rome, 1995.

Chapter 3

GRAPHICAL EDITING

FOREWORD

By Ron James, United Kingdom

For the purposes of this chapter, the graphical editing is the exploitation of human visual perception ability to identify an anomaly, a pattern or an implied relationship in data that might take much more time and effort to find by analytic non-graphical means. Graphical data editing methods proved to be in many National Statistical Offices an efficient tool for the reduction of statistical survey costs.

Furthermore, the experiences show that traditional editing on a case-by-case basis does not often allow to see clearly the impact of an individual data point on the aggregate estimate. The obvious examples are the identification of an outlier for specific data and a simple relationship between different data in a set of cases. The outlier may represent an error or an exceptional movement. The simple relationship, perhaps represented as a straight line with suitably transformed data, may provide an easier visual method for finding an error or an important exception. Sometimes it is the nature of the relationship that it is valuable and easier to see as a result of graphing.

In those cases where the statistician has little prior knowledge of the response to a survey, preliminary exploration of the early data using graphics can be of significant help in setting up the bounds of the editing process. Also they can be used in imputation and editing itself, monitoring and tuning the editing process and in evaluating imputation procedures by exposing the aggregate properties of imputed data as compared with the overall data set.

The paper by Bienias et al. develops the data exploration theme using graphics and transformations to help expose trends and other properties of the data that may be difficult and more time consuming to find by purely analytical means. Knowledge gained in this way may then be used to define, improve or enhance the editing process.

Graphics can be used in a more specific way as an integral part of the editing process. These processes are essentially identifying exceptional data. They are exceptional either because they have a large effect on the aggregate results of the survey, or because they are erroneous, or both. In either case the graphical process has led the editor to the key data directly and efficiently. The technique of identifying those data that have most impact on the aggregate results is known as 'macro' or 'top down' editing. Such a technique which enables the editors to concentrate their resources on confirming or correcting the most influential data presumes that the remaining large number of small errors are non systematic and have little effect on the result. This has been confirmed by several studies. It does not necessarily mean that editing stops after the largest errors or outliers have been identified. Where the data is to be used for subsequent analyses, especially for purposes other than the initial aggregates it may be essential to complete the process of editing. However, such techniques are an efficient way of reducing the cost and time required to achieve specific results from a survey. The paper by Per Engström of the Swedish Statistical Bureau is a classic example of macro or top down editing.

It is typical of graphical methods that they provide insight, revealing things that were sometimes not dreamt of before the process began. This is a point made by Esposito et al. in describing the ARIES system used in the US Department of Labour. It is thought important that these graphical systems are open ended with flexible facilities to extend and develop new views responding to what the graphical information is revealing. A further example is provided in the paper on Graphical Editing and Analysis Query System by Paula Weir et al. of the US Department of Energy. The PC platform with its Graphical User Interface and an Object Oriented approach would seem to be an ideal environment for these developments.

IMPROVING OUTLIER DETECTION IN TWO ESTABLISHMENT SURVEYS

Chapter 3

GRAPHICAL EDITING

FOREWORD

By Ron James, United Kingdom

For the purposes of this chapter, the graphical editing is the exploitation of human visual perception ability to identify an anomaly, a pattern or an implied relationship in data that might take much more time and effort to find by analytic non-graphical means. Graphical data editing methods proved to be in many National Statistical Offices an efficient tool for the reduction of statistical survey costs.

Furthermore, the experiences show that traditional editing on a case-by-case basis does not often allow to see clearly the impact of an individual data point on the aggregate estimate. The obvious examples are the identification of an outlier for specific data and a simple relationship between different data in a set of cases. The outlier may represent an error or an exceptional movement. The simple relationship, perhaps represented as a straight line with suitably transformed data, may provide an easier visual method for finding an error or an important exception. Sometimes it is the nature of the relationship that it is valuable and easier to see as a result of graphing.

In those cases where the statistician has little prior knowledge of the response to a survey, preliminary exploration of the early data using graphics can be of significant help in setting up the bounds of the editing process. Also they can be used in imputation and editing itself, monitoring and tuning the editing process and in evaluating imputation procedures by exposing the aggregate properties of imputed data as compared with the overall data set.

The paper by Bienias et al. develops the data exploration theme using graphics and transformations to help expose trends and other properties of the data that may be difficult and more time consuming to find by purely analytical means. Knowledge gained in this way may then be used to define, improve or enhance the editing process.

Graphics can be used in a more specific way as an integral part of the editing process. These processes are essentially identifying exceptional data. They are exceptional either because they have a large effect on the aggregate results of the survey, or because they are erroneous, or both. In either case the graphical process has led the editor to the key data directly and efficiently. The technique of identifying those data that have most impact on the aggregate results is known as 'macro' or 'top down' editing. Such a technique which enables the editors to concentrate their resources on confirming or correcting the most influential data presumes that the remaining large number of small errors are non systematic and have little effect on the result. This has been confirmed by several studies. It does not necessarily mean that editing stops after the largest errors or outliers have been identified. Where the data is to be used for subsequent analyses, especially for purposes other than the initial aggregates it may be essential to complete the process of editing. However, such techniques are an efficient way of reducing the cost and time required to achieve specific results from a survey. The paper by Per Engström of the Swedish Statistical Bureau is a classic example of macro or top down editing.

It is typical of graphical methods that they provide insight, revealing things that were sometimes not dreamt of before the process began. This is a point made by Esposito et al. in describing the ARIES system used in the US Department of Labour. It is thought important that these graphical systems are open ended with flexible facilities to extend and develop new views responding to what the graphical information is revealing. A further example is provided in the paper on Graphical Editing and Analysis Query System by Paula Weir et al. of the US Department of Energy. The PC platform with its Graphical User Interface and an Object Oriented approach would seem to be an ideal environment for these developments.

IMPROVING OUTLIER DETECTION IN TWO ESTABLISHMENT SURVEYS

By Julia L. Bienias, David M. Lassman, Scott A. Scheleur, and Howard Hogan, Bureau of the Census, USA

ABSTRACT

Previous researchers have successfully used various graphical methods to improve both the efficiency and accuracy of the editing process (e.g. [2], [3], [5], [6]). This paper describes an application of graphical methods from exploratory data analysis to the task of identifying potentially incorrect data points. The use of various plots and transformations in exposing potentially incorrect data points is demonstrated. The graphical methods are illustrated with data primarily from the Annual Survey of Communication Services and the Monthly Wholesale Trade Survey.

Keywords: exploratory data analysis; boxplots; scatter plots; linear regression; resistant regression; fitting methods.

1. INTRODUCTION

An important step in producing estimates from survey data is editing. In many settings, trained analysts examine the data to find unusual or unexpected values, which may be the result of errors made by the respondent or in the data-capture processes. Having found a questionable case, the analyst then tries to verify its accuracy by checking the original form, obtaining related data from other sources, and/or contacting the respondent. One would like to correct as many errors as possible within the time limitations for a given survey. Thus, accurately identifying the cases whose values are most likely to be the result of errors is an essential part of efficient editing.

Previous researchers have successfully used various graphical methods to improve both the efficiency and accuracy of the editing process (e.g. [2], [3], [5], [6]). We describe the application of graphical methods from exploratory data analysis to the task of identifying potentially incorrect data points. Our report is the result of a working group of analysts, research statisticians, and programmers devoted to this effort. We illustrate the methods with data primarily from the Annual Survey of Communication Services and the Monthly Wholesale Trade Survey. We first describe the two surveys and the current methods used for editing.

2. DESCRIPTIONS OF THE TWO SURVEYS

2.1 The Annual Survey of Communication Services

The Annual Survey of Communication Services (ASCS) is a mail survey covering all employer firms that are primarily engaged in providing point-to-point communication services (e.g., telephone, television, radio), as defined in Major Group 48 of the 1987 edition of the *Standard Industrial Classification Manual*. The ASCS provides detailed revenue and expense statistics from a sample of approximately 2,000. The Census Bureau introduced the survey in 1991 to track the explosive growth and change in the industry. The Bureau of Economic Analysis is the primary federal user of the data collected; other users are the Bureau of Labor Statistics and private industry [12].

2.2 The Monthly Wholesale Trade Survey

The scope of the Monthly Wholesale Trade Survey (MWTS) is all employer firms engaged in wholesale trade, as defined by Major Groups 50 and 51 of the 1987 edition of the *Standard Industrial Classification Manual*. Particularly, the survey covers merchant wholesalers who take title to the goods they buy and sell, collecting sales and inventory information. The MWTS, conducted since the 1940's, is a mail survey of approximately 7,000 firms, of which 3,500 receive forms in a given month. As with the ASCS, the Bureau of Economic Analysis is the primary federal user of the data [13].

3. ISSUES INVOLVED IN THE CURRENT EDITING PROCEDURES

After the data from the questionnaires are keyed, a computer program flags cases failing completeness, internal consistency, and/or tolerance edits. Editing review is divided among several analysts for a given survey. Each analyst finds which edits have failed for a case through an interactive correction system or a paper listing, on a case-by-case basis. They can also use a database query system to try to find problem cases that have not already been identified.

There are several disadvantages to this approach. Examining one case at a time does not permit the analyst to obtain a broad view of the behavior of the

industry as a whole, and such a view can be of great benefit in determining the impact of an individual unit on the aggregate estimate. In addition, it undoubtedly leads analysts to examine more cases than necessary. Finally, for a few of the ASCS tolerance edits, constant parameter levels derived from previous surveys have been hard-coded into the programs. This implicitly assumes the relationships among the variables are static over time, which may not be the case.

4. APPLICATION OF EXPLORATORY DATA ANALYSIS METHODS

4.1 Background

Exploratory data analysis (EDA) can be described as "a set of tools for finding what we might have otherwise missed" in a set of data (see [11]). These tools, combined with the analysts' subject-matter expertise, are particularly well-suited to the task of data editing. In this setting, we are not interested in ascertaining the truth of a postulated economic model or a similar estimation or hypothesis testing problem. Rather, our goal is to determine which cases are unusual with respect to the bulk of the cases and to follow up those cases. In addition to providing methods for displaying data in a variety of ways, EDA emphasizes fitting data using methods that are relatively insensitive to the presence of outliers in the data ("resistant" methods). Such fitting is a way to define and then account for (remove) certain aspects of the data so the analyst can concentrate on other aspects. (See [4], [14])

EDA fits well with the survey processing environment. Because in the editing stage we expect to find wild observations that might be off by orders of magnitude from the bulk of the data, transformations and resistant techniques are particularly useful in helping us find order amid the chaos. In addition, these techniques allow for efficient examination of large amounts of information at once, an aspect that is particularly valuable in the time- and resource-constrained survey production environment.

From the arsenal of tools collectively called "exploratory data analysis," we considered both univariate boxplots and the more general bivariate fitting. We describe boxplots first, followed by scatter plots and some methods for fitting. In addition, although transformations are applicable to all tools, we describe them in the context of scatter plots, because that is where we used them most.

4.2 Boxplots

Boxplots allow quick visual analysis of the location, spread, and shape of a distribution. Our boxplot has its box spanning the lower and upper quartiles, with whiskers extending from the box to the furthest data point within a distance of one-and-one-half times the interquartile range from the box. We considered data values beyond the whiskers as potential outliers. If the data are reasonably symmetric, then these cutoffs provide a good working definition of cases which may need review. See [11] for a discussion of boxplots in general, and [4] for a discussion of the expected number of outliers for different sample sizes. Note that the whisker definition could be modified to suit the needs of a particular survey operation (e.g., one could use 2 times the interquartile range instead of 1.5).

Figure 1

Figure 1 demonstrates the use of the boxplot for operating ratio (expenses/revenue) data from the ASCS. The boxplot shows that the median operating ratio is .7978 and fifty percent of the points lie between .7269 and .9811. The left and right whisker values are .3760 and 1.3401. The cases flagged by

the use of the boxplot are different (and fewer in number) than the cases that would have been flagged by the current hard-coded edit parameters, .9 and 1.1.Those parameters fail to help us isolate the "true" outlier cases, as they result in too many cases being flagged.

Alternatively, we could flag cases that would appear beyond the whiskers as in our boxplot, an approach that is "dynamic" in that it relies on incoming data to set parameters. At minimum, we could use values from Figure 1 as new hard-coded edit bounds, noting that these revised bounds would no longer be symmetric around one (consistent with the findings of Granquist [3]).

4.3 Scatter Plots

A scatter plot of two variables is a simple and particularly useful technique. When the data are appropriately transformed, one can use a variety of methods to remove linearity in the scatter and then examine the residuals from the linear fit. This allows us to see patterns that we might otherwise miss when looking at the original data; looking at the residuals from a fit allows us to examine the data on a finer scale (see Section 4.5).

Figure 2

As a vivid illustration of the kinds of problems

encountered in editing data, we used another survey for which we had raw responses to a particularly problematic question. One item in the Motor Freight Transportation and Warehousing Survey is the percent of revenue derived from local trucking, a question believed to be confusing to respondents as we may define "local" in different ways. Figure 2, a scatter plot of these unedited data for the current versus prior period, shows a weak linear relationship. Cases along the 45 degree line are companies whose year-to-year reports are consistent. The reports become more inconsistent the further they are from the 45 degree line. Some of the cases along the vertical axis are "births" to the survey (cases selected during the current period to reflect new firms). Births should be analyzed separately, because they have only current-year data.

4.4 Transformations

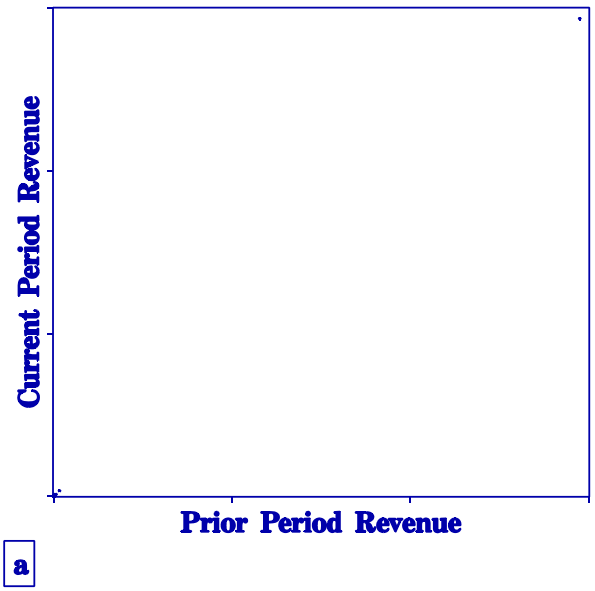
Transforming the data so patterns can be more easily discerned is a technique that is important to all graphical and data-fitting methods. It is used to obtain symmetry in the data, to promote linearity, and to equalize spreads between data sets. These properties are assumed, implicitly or explicitly, by many of the techniques we use to analyze data. For example, when we look for outliers by examining a boxplot, we are implicitly assuming the data are supposed to be symmetric. If the data are naturally skewed, many of the points in the tail that appear to be outliers are actually values that are consistent with the underlying distribution. Thus, "discovering" such outliers in the long tail would not be very meaningful. With skewed data, we want to spend our time investigating those data points that are particularly unusual, given that we expect many points far from the bulk of the data. If we transform skewed data to be generally symmetric, we can then find those points.

Because economic data are typically positively-skewed, transformations that lead to the expansion of lower data values and to shrinking the spread of larger data values are particularly useful. (See [4] for more details on types of transformations.)

Figure 3 is an example of the use of transformations for the ASCS. The scatter plot of untransformed revenue data (Figure 3a) reveals little, as one case is many times larger than the other cases. Hiding the large case was unsuccessful, as the next largest case was still many times larger than the remaining cases. Instead, taking logs of the data showed a useful scatter plot (Figure 3b). We see a strong linear relationship, which is what we expect for a plot of current and prior data. Cases that do not appear to be following this linear relationship would thus be considered unusual. We also see that the case that appeared to be an outlier in Figure 3a is, in fact, very much in line with the rest of the data.

Figure 3

For the MWTS, a scatter plot of the current inventory data against the current sales data shows that most of the data are bunched in the lower left corner (see Figure 4a). Because both variables are skewed, we first tried a natural log transformation (log(x+1)). (We added one because a value of 0 for inventory data does not indicate the case is a birth, and thus it may be of value to include such cases.) This overtransformed the data, skewing them in the opposite direction (Figure 4b). This is because there was a big gap in values between 0 and the next largest value. Such an effect would also occur if there were many establishments with very small reported data and a few with very large values. We then tried taking the square root (Figure 4c) and fourth root (Figure 4d). The latter resulted in the most useful transformation, as most of the data can be seen clearly.



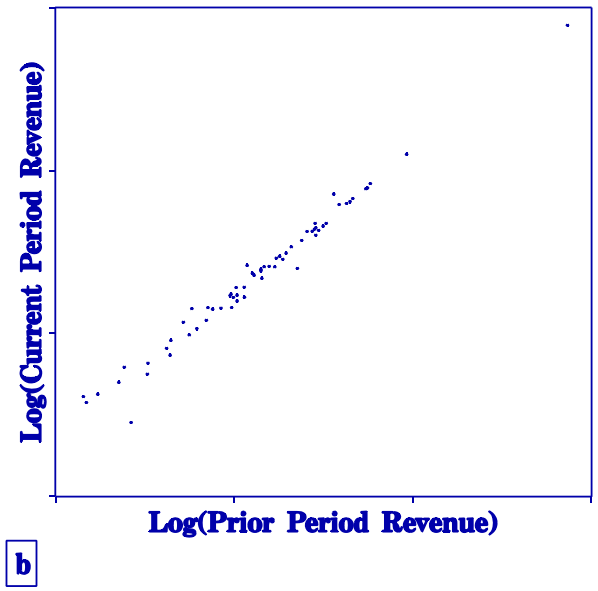


Figure 4

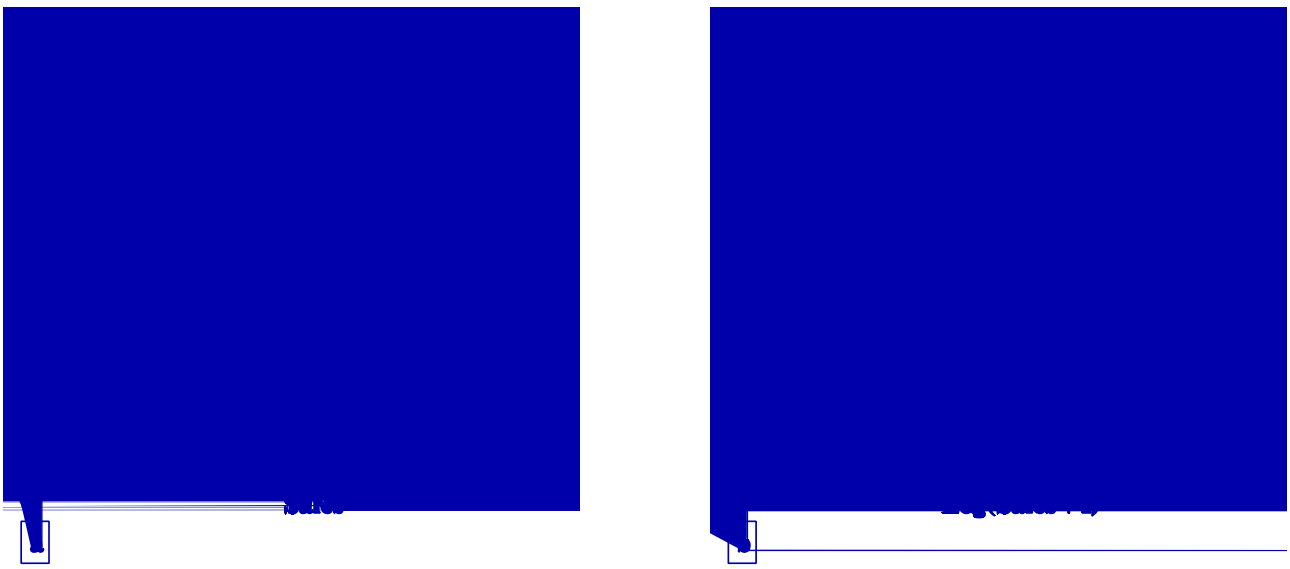
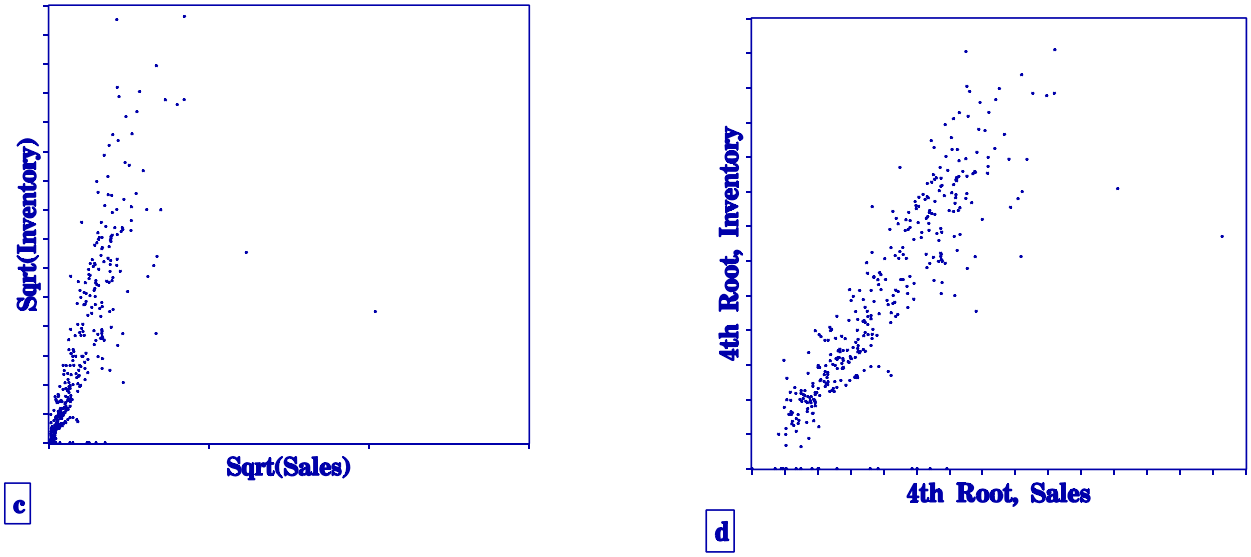


Figure 4



4.5 Fitting

In this section we describe two approaches to fitting, ordinary linear regression and resistant regression. Both were useful, in different ways.

In analyzing ASCS data, we considered the relationship between revenue and payroll for current year data. Figure 5a shows the ordinary least squares regression of revenue on payroll; there are many points clustered near the origin and two cases in the upper right corner. First, we tried removing the two large cases. Again the distribution showed points clustered in the left corner. Such an approach, of iteratively hiding points and refitting, has the disadvantage of being subjective and of essentially requiring analysts to identify outliers first.

One alternative is to use ordinary least squares on transformed data. In this example, logs were useful. Figure 5b shows the fit to the logged data, depicting a strong linear relationship. The point labeled A is an obvious outlier. Examination of the residuals revealed a pattern, which allowed us to discover that tax-

Figure 5

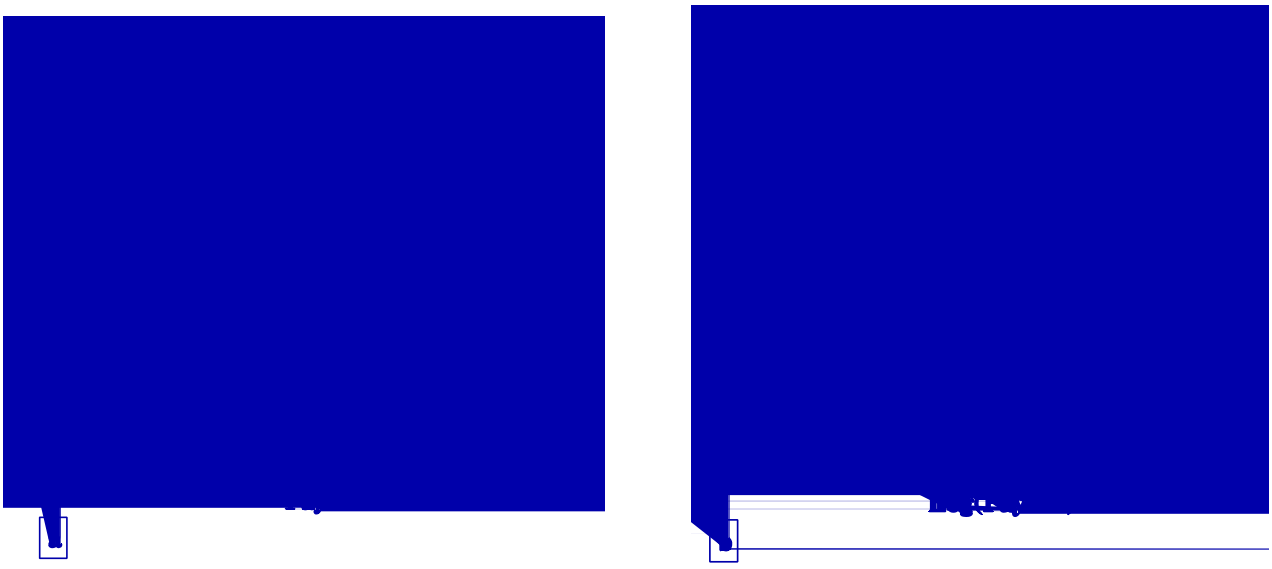
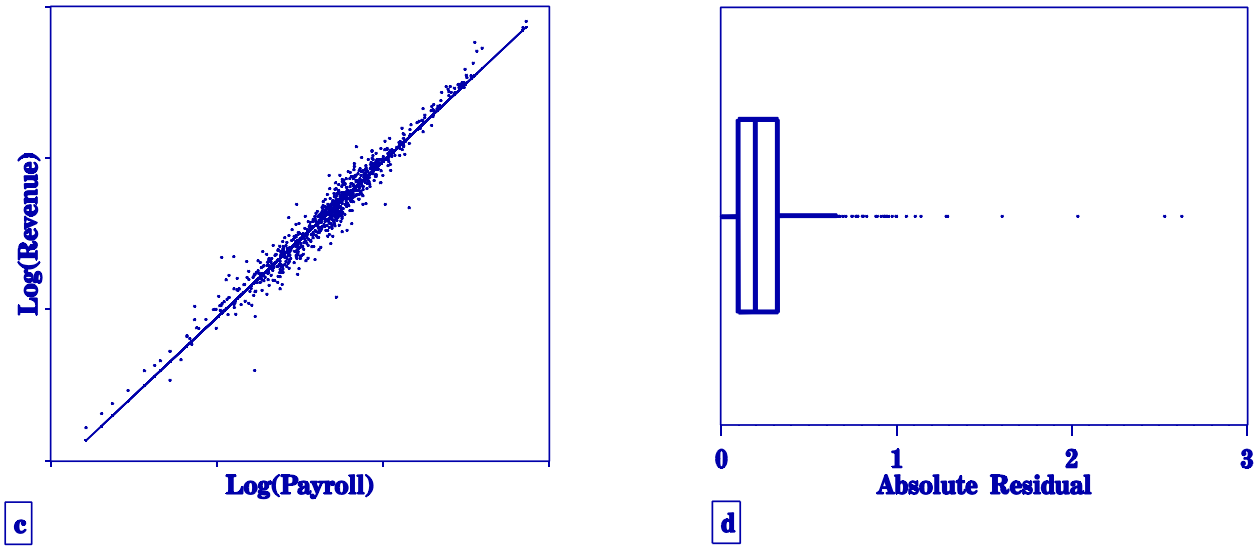


Figure 5



exempt cases were inadvertently being included in the analysis. Tax-exempt cases should be examined separately from taxable cases, because our revenue item only includes taxable receipts. Removing both those cases and point A and refitting the data (Figure 5c) led to the distribution of absolute residuals shown in Figure 5d. This plot can be used to detect outliers, as with a cutoff level *C*:

$$C = K * \text{(median absolute residual)}.$$

We found *K*=4 (corresponding to *C* = .7868) to be the best. All cases above .7868 were examined and most were "true" outliers. For our example, this method was judged by the survey analysts to be excellent for finding outliers.

Unfortunately, ordinary least squares (OLS) can give great weight to fitting a few wild values. It may work well, as in our example, when there are only a few wild cases and the demarcation between usual and unusual is clear. As an alternative, we investigated resistant fitting using the biweight function developed by Tukey [8], [9]. This widely-tested iterative weighted-least-squares fitting procedure uses a weighting function defined as:

$$w_i = \begin{cases} (1-u_i^2)^2 , & u_I < 1 \\ 0 \text{ otherwise} \end{cases}$$

where $u_i = (r_i / (c*s))$

r_i / residual from previous fit for point *i*

s / mean absolute residual from previous fit

c / scale factor.

Setting c = 4 is quite resistant, c = 8 is moderately resistant. We stopped iterating when the proportionate change in *s* was less than 0.01. This required few iterations; resistant regression is a very efficient and fast procedure.

We applied resistant regression to the MWTS, predicting logged current inventory data by logged inventory data from the prior year. We expect a linear relationship. Figure 6a shows the data and the line from the OLS fit, and Figure 6b shows the residuals from that fit. It is easy to see the OLS fit misses the central tendency of the point cloud. Figure 7a shows the fit resulting from resistant regression (c=4). This fit more effectively removed the linearity from the data. The residuals now cluster around 0, as we would want (Figure 7b).

5. A NOTE ON USING RATIOS

In many instances, data review has relied on calculating ratios (e.g., sales/payroll) and looking for unusually large or small ratios. There is nothing wrong with this approach per se, but it would be wrong to rely too strongly on it.

The use of ratios assumes a rather simple model of the true relation between the two variables, specifically a straight line through the origin. The true relation may differ markedly, there may be data clouds following different straight lines. For example, the relationship might be different for a small company than for a large company. It is essential that the data reviewer plot the data and look at the shape. Further,

Figure 6

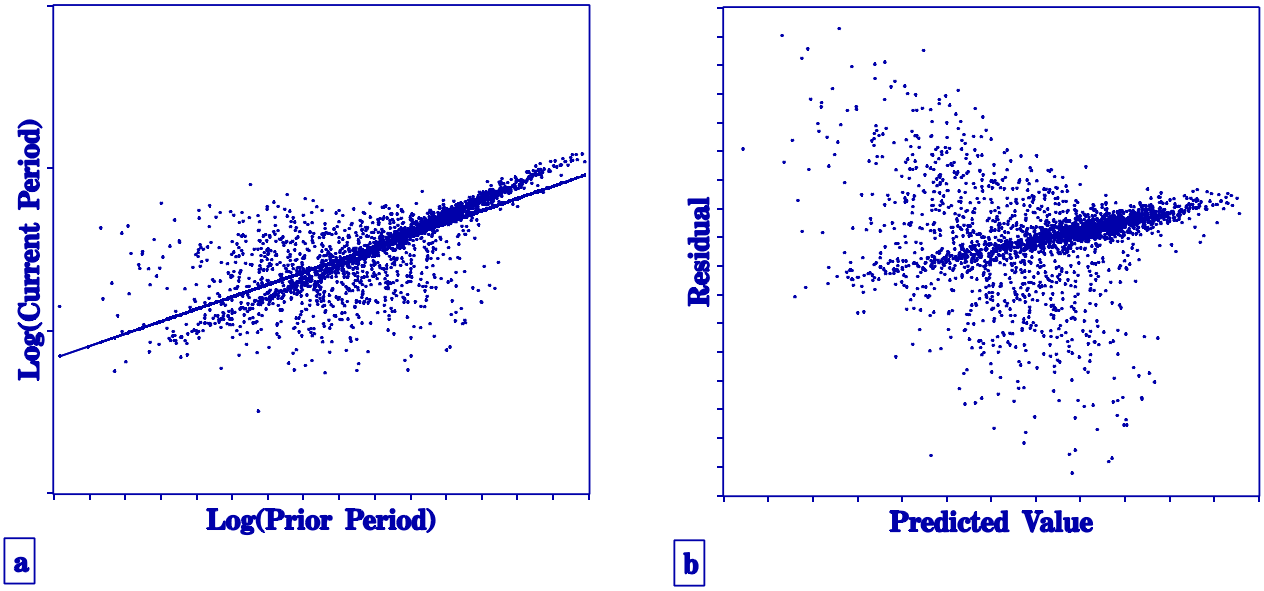
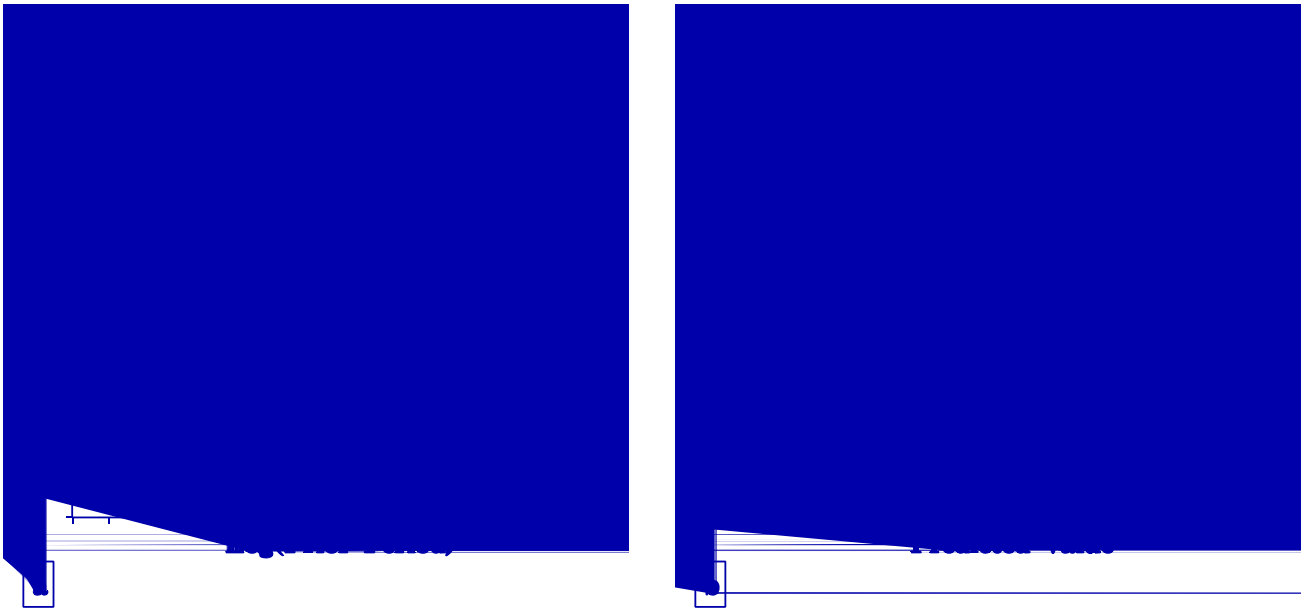


Figure 7



the "acceptable ratios" are often set from historic data, last year's or last census'. The relationships can change systematically throughout the business cycle. One could iterate, calculate the average ratio from the current survey, calculate its standard deviation, identify and remove outliers, and start again. However, given the existence of rather fine iterative resistant fitting tools, it is hard to see the advantage of this approach.

6. SUMMARY AND EXTENSIONS

We have described how principles and methods from EDA can be used to improve the efficiency and

accuracy of editing, by helping analysts see patterns in the data and use that information to prioritize cases for follow-up. Building a successful editing system using this approach is more than just selecting the correct statistical tools. The system must be acceptable to the people who will use it. Creating such acceptance requires training the analysts in the methods described here, as well as incorporating the tools into the current production environment and existing computer systems. To date, we have been successful in getting many people to try the methods on several surveys. In addition to the surveys described previously, these methods are currently being applied to the Motor Freight Transportation and Warehousing Survey, the Service Annual Survey, and the Commodity Flow Survey.

Analysts for these surveys reported that being able to ascertain the effect of a given case on the estimate was quite useful. Other specialized programs written for data editing provide this feature (e.g., [2], [5]). Incorporating sampling weights in the procedures described here provides a similar utility.

The EDA approach can be combined with batch-type edits (e.g., [1], [7]). One could examine the data flagged from a batch program along with the unflagged data using the tools described here. Or, the graphical-based methods could be the basis for batch-type dynamic edits. For example, a program could transform the data to be more symmetric and then flag all cases that would be beyond the whiskers of a boxplot. Finally, in settings in which hard-coded edit parameters must be used, these methods can be used on a subset of data to help find or evaluate such cutoffs.

REFERENCES

[1] Draper, L ., Greenberg, B., Petkunas, T. On-line capabilities in SPEER (Structured Programs for Economic Editing and Referrals), *Proceedings of Statistics Canada Symposium 90: Measurement and Improvement of Data Quality*, Statistics Canada, Ottawa, 1990, pp. 235-44.

[2] Esposito, R., Fox, J. K., Lin, D., Tidemann, K. (in press). ARIES: A visual path in the investigation of statistical data, *Computational and Graphical Statistics*.

[3] Granquist, L. A review of some macro-editing methods for rationalizing the editing process. *Proceedings of Statistics Canada Symposium 90, Measurement and Improvement of Data Quality*, Statistics Canada, Ottawa, 1990, pp. 225-34.

[4] Hoaglin, D.C., Mosteller, F., Tukey, J. W. (Eds.) Understanding Robust and Exploratory Data Analysis,. Wiley NY, 1983.

[5] Houston, G., Bruce, A. G. Graphical editing for business and economic surveys, Technical report, New Zealand Department of Statistics, Mathematical Statistical Branch, February 1992.

[6] Hughes, P.J. , McDermid, I. , Linacre, S. J. The use of graphical methods in editing (with discussion), *Proceedings of the 1990 Bureau of the Census Annual Research Conference*, U.S. Department of Commerce, Washington, DC, 1990, pp. 538-54.

[7] Lee, H. Outliers in survey sampling. In B. Cox et al. (Eds.), *Survey Methods for Business, Farms, and Institutions*, Wiley, NY, (in press).

[8] Mosteller, F., Tukey, J. *Data Analysis and Regression*, Addison Wesley, Reading, MA, 1977.

[9] McNeil, D. R. Interactive Data Analysis, Wiley, NY, 1977.

[10] Office of Management and Budget, *Standard Industrial Classification Manual*. Available from National Technical Information Service, Springfield, VA (Order no. PB 87-100012). 1987.

[11] Tukey, J. W. Exploratory Data Analysis,. Addison-Wesley, Reading, MA, 1977..

[12] U.S. Bureau of the Census, *Annual Survey of Communication Services: 1992,* U.S. Government Printing Office, (Current Business Reports, Item BC/92), Washington, DC, 1992.

[13] U.S. Bureau of the Census, Combined Annual and Revised Monthly Wholesale Trade, January 1987-December 1993, U.S. Government Printing Office, (Current Business Reports, Item BW/93-RV),. Washington, DC, April 1994.

[14] Velleman, P. F., Hoaglin, D. *Applications, Basics, and Computing of Exploratory Data Analysis*. Boston: Duxbury Press,1981.

THE GRAPHICAL AND QUERY SYSTEM 'ARIES'

By Richard Esposito, Dong-yow Lin, and Kevin Tidemann, Bureau of Labor Statistics,USA

ABSTRACT

The paper gives an overview of a graphical and query system ARIES that has been successively used in the Current Employment Statistics (CES) program of the Bureau of Labor Statistics (BLS) of the U.S. Department of Labor since November, 1991. The ARIES (Automated Review of Industry Employment Statistics) system has improved the data review capabilities of CES analysts, and has allowed to rethink fundamentally the possibilities for improving data review in the BLS. ARIES uses a top-down approach for outlier detection and treatment, the search for suspect sample data is driven by preliminary identification of suspect estimates. Estimates which deviate from historical trends are presented graphically which enables viewers to limit their search for sample outliers to a small subset of the numerous estimates computed. Query and graphical search techniques are then used to pinpoint individual suspect sample members which may have caused those estimates to be suspect. Finally, outliers are given appropriate weights and new estimates calculated using tabular screens on the PC.

Keywords: Graphics; editing; statistical survey

1. BACKGROUND

The Current Employment Statistics (CES) program of the BLS estimates total employment, employment of women, non-supervisory employment, average weekly hours, and average hourly earnings for virtually all non-agricultural industries in the United States, and average overtime hours in manufacturing industries. These estimates are computed from a month-to-month matched sample survey of over 350,000 governmental and private establishments each month, with estimates computed for over 1600 sample-based and 1000 aggregate level industry cells, for each of the data elements mentioned. About a dozen industry analysts are responsible for the quality of sample data and estimates. They guide the data through a series of quality control steps, from the receipt of the sample data from state offices which collect the data from establishments, through the process of computing final estimates.

Formerly, data review in the CES program has been an inefficient mix of main frame computers, and a lot of human work. Many tasks which would have made the analysis less arduous were never assigned to the machine, and humans and useful information which would have made both sides' tasks more efficient was not exchanged. This is probably the point on which ARIES makes its most significant contribution. The information provided by the ARIES computer system is essentially a better understanding of the data on a current basis. The old system of reviewing many thousands of lines of computer paper owed its success to the ability of the human industry analysts to construct over time their own internal pictures of the industries, gaining, through arduous labor and from bits and pieces, some idea of the whole.

2. EDITING USING ARIES

Industry employment and earnings estimates are computed from establishment sample data 3 times each month; 1st preliminary, 2nd preliminary, and final estimates. Each estimate is computed using more cumulated sample as establishments report their data for the given month. Typically, for 1st preliminary estimates, 200,000 sample data reports are first processed on an IBM mainframe. For all industry estimating cells which fail tolerance checks, all sample data and associated historical data are downloaded from the mainframe to 12 individual 486-based PC's, according to the industries for which each of 12 industry analysts is responsible. This transfer of data from the mainframe to PC's is done automatically overnight, using a combination of programs written in C and ACS Excellink/Host-V, although there are faster ways of performing this task currently. An investigator program automatically notifies the mainframe computer to begin alternative fall-back procedures if any individual PC fails to receive its data.

By the time industry analysts arrive at work in the morning, all relevant data has been automatically downloaded to their individual PC's and the review procedure can begin. A short description of the ARIES review process is as follows: each analyst uses an estimate level graphical representation of all of their industries to identify those industries with suspect estimates, based on normal historical month-to-month changes. For each suspect industry estimate identified, the analyst will use either of two graphical methods, or a query method, to identify suspect sample reports which may have contributed to the suspect estimates.

Suspicious sample thus identified can then be automatically corrected or given an appropriate weight for estimation, and new estimates are automatically and immediately calculated based upon any weightings performed.

All changes made are automatically entered into an audit trail, and also copied to a LAN server, so that a permanent record is kept for supervisory review and later analysis. At the end of the review process, corrected and weighted sample data are uploaded back to the mainframe for storage, with the entire process to be converted to the Client/Server environment in the future.

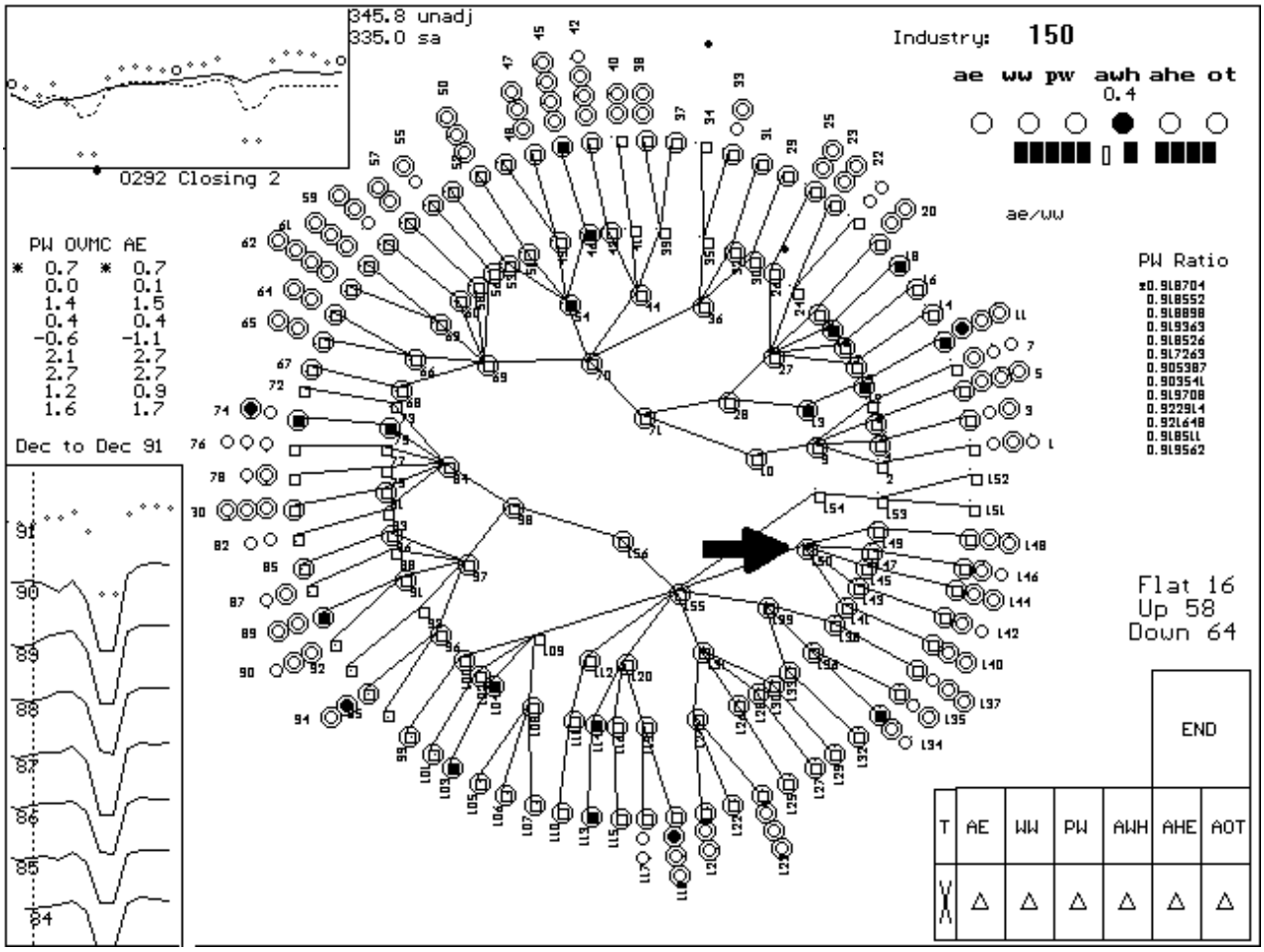
3. DEVELOPMENT CONSIDERATIONS IN ARIES

Each step in ARIES has not only fulfilled a specific need, but has also generated ideas for future improvements.

Industry analysts begin their review with the "Anomaly Map" for their own specific industries. A typical anomaly map is shown in Figure 1, with all industry identifying information eliminated because of confidentiality requirements.

The anomaly map is a tree of the industry structure for that industry analyst, with more finely differentiated estimating cells on the perimeter, and more aggregate levels closer to the center of the map.

Figure 1. Anomaly Map



Essentially, each node of the tree represents a specific industry. Estimation from sample data is done for perimeter cell industries, and estimates for more aggregate level industries towards the center are obtained by addition from the perimeter cells. On the PC, colors are used to mark industry nodes whose

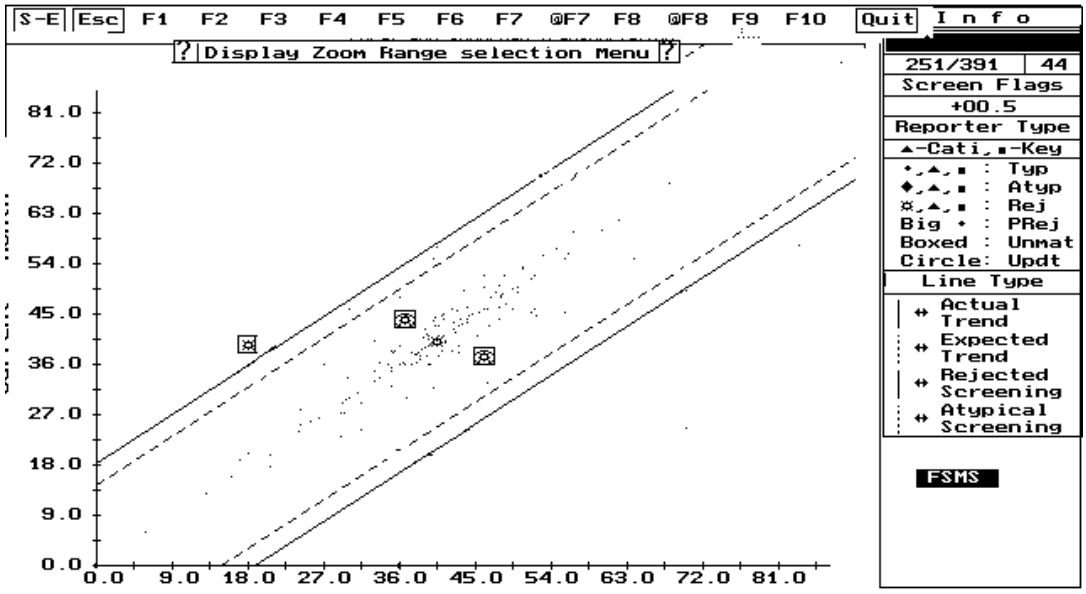
estimates are outside historically determined tolerances, and the specific color used indicates how much out of tolerance that industry's estimate is. The connecting lines of the tree are also colored to connect parent and child nodes which have similar tolerance failures. When any given industry node is clicked on with the mouse, historical monthly estimates are charted stacked-by-year in the lower left corner, and for the most recent two years in the upper left corner. In Figure 1, historical time-series are shown for the industry marked with the larger arrow (for this article, a large size circle has been substituted for the color indicator on the PC). The smaller arrow in the upper left corner points to the current estimate, which can clearly be seen to be far below the historical trend for this industry. The anomaly map shown here shows tolerance anomalies for Average Hourly Earnings estimates; by using the mouse, the colors will change to show tolerance exceptions for Employment, Woman Worker, Production Worker, Average Weekly Hours, or Average Overtime Hours.

The anomaly map was created for two reasons. The first motivation was to give industry analysts a better picture of what is happening to the estimates in their industries, and the second was to enable a quick top-down search for industries whose sample may be causing an estimate to be suspect. Suspicious estimates at the published level can quickly be traced back to the specific basic estimating node which caused the deviation. The design impulse to create the anomaly map was to try to fit as much information on all of the industries on a single screen. The anomaly maps can be seen as a first step in obtaining an overall picture of what is happening in a large subgroup of industries. At present, only one category of estimates is shown at a time, for example, in Figure 1, only the colors for average hourly earnings would be plotted. Ideally, the relationships between all categories for all industries could be pictured at once. We supply a partial solution to that problem by plotting the color representation, in the six circles in the upper right corner, for all the associated data categories for a single industry, once that industry's node has been pointed to by the mouse. That eliminates a lot of paging back and forth between screens.

The anomaly map is used to concentrate the search for suspect sample data on only those industries which have problems with their estimates. Once an industry analyst has identified those industries on the anomaly map, he or she will begin the search for suspect *sample* data, confining the search to only those suspect industries.

To isolate suspect sample data in a given industry, two graphical methods and one query method are available. The first graphical method is to identify sample outliers from a scatter gram such as Figure 2.

Figure 2. Scatter Gram for a Single Industry



Each industry's scatter gram takes a few seconds to plot on the PC screen. Each point on the scatter gram represents the current month's reported data (on the vertical axis) and the previous month's reported data (on the horizontal axis) for an individual establishment. In the CES survey, month-to-month matched sample data is used to calculate estimates. A natural expectation is that data points will group along the 45 degree line, with seasonality and trend factored in. Employment estimates tend to group along this line

better than average hourly earnings, average weekly hours and average weekly overtime hours worked. Automatically computed screening ranges are constructed and plotted on the screen at a set number of standard deviations from the (approximately) 45 degree expected value line. The establishment points outside the solid lines (extreme failure region) and those outside the dashed lines (less extreme failure region) would normally be the establishments whose data the industry analyst would select, using the mouse, to look at more carefully. The information boxes to the right in Figure 2 contain a legend for the various actions performed on each point, and these actions and various categories of sample data are marked with special symbols on the scatter gram. By selecting one or more scatter gram points with the mouse, the analyst can quickly bring up on the screen detailed tabular and graphical historical sample data corresponding to the establishment represented by that point. At that point, the analyst can optionally call up a tabular or graphical data report of the most recent 16 months of sample values for that individual reporter. An example of such a graphical view for the sample member of Figure 2, and for six data types, is shown as Figure 3. (Overtime hours are not collected in this industry).

The second graphical method used to identify sample outliers is through the plotting of the sample distribution of the month-to-month change in all sample establishments in an industry. An example of such a distribution graph is seen in Figure 4.

Selection of sample outliers to further investigate is done from the distribution graph by moving the tall double bars to isolate sample reports on either end of the distribution. The sample distribution itself is shown as the darker bars; a comparison normal distribution is also plotted as the less dark bars.

The third method used to isolate sample outliers is through the use of pre-set query questions; various parameters can be set for any of 108 different queries, and the PC will search the particular industries sample data to select those sample members which meet the conditions set by the queries. The query mode of selecting outliers is especially suitable for finding sample with specific characteristics, or with particular relationships among their different sample data items.

Figure 3. Example of a Graphical 16-Month Data Report for a Selected Sample Report

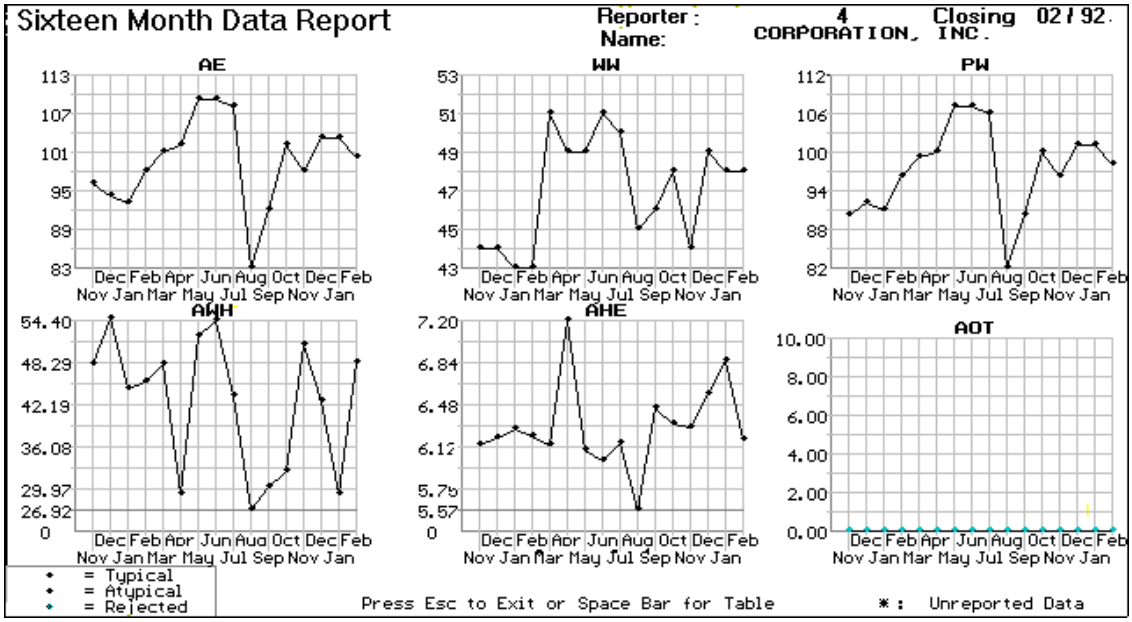
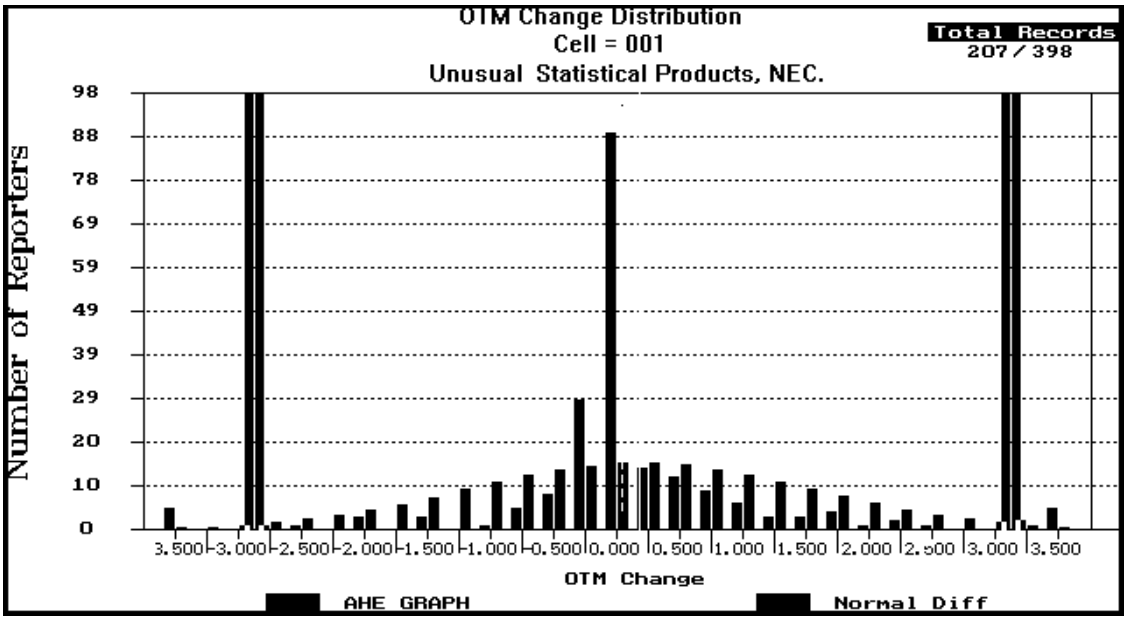


Figure 4. Sample Distribution for a Single Industry



Once outliers are isolated using any of these three methods, the industry analysts will give suspect sample data appropriate weights, and new estimates are immediately calculated on the PC and added to more aggregate industry levels.

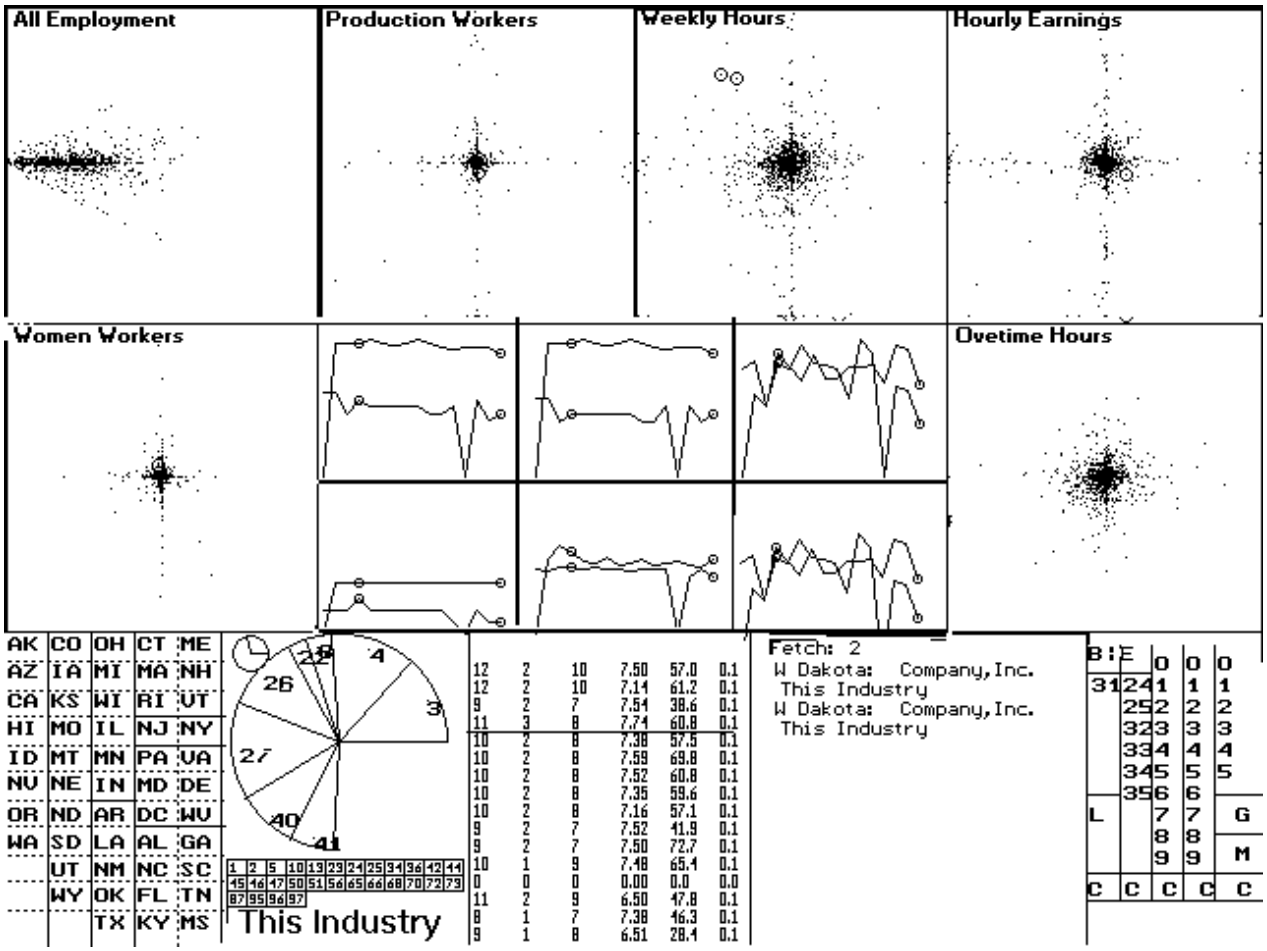
4. RECENT PROTOTYPE DEVELOPMENTS

In order to provide a better and more comprehensive visual representation of data, we are attempting to provide solutions to several specific goals which the ARIES system has brought into focus. For this purpose we have constructed a prototype which we anticipate will be integrated into the ARIES system. This "Phase 2" prototype is seen in Figures 5 and 6.

In Figure 5, the scatter grams for all six sample data types for all sample reports for an industry are shown on one screen, so that each sample establishment is represented by six points, one in each scatter gram. When a single point in any of the six scatter grams is clicked using the mouse, the associated points in the other five scatter grams for that establishment are circled, as well. When more than one establishment point is captured by the mouse, associated circles are color coded to keep the establishment information separate. At the same time, the establishment name and address and other identifying information appear in the bottom row, fourth box from the left (replaced here by company names from West Dakota), and a graphical time-series for the most recent 16 months of sample data are displayed for all six data types in the smaller boxes in the center of the screen (also color coded on the PC screen). In the bottom row, 3rd box from the left in Figure 4, are shown the numerical values for those 16 months for any of the establishments shown, which appear by mouse-clicking on the company name. In Figure 5, we have mouse-clicked the points for two sample member establishments.

These scatter grams are an evolution from the scatter gram of Figure 2. In Figure 2, only one sample data type is pictured on the screen at one time; in Figure 5 all data types are shown, and this makes it much easier to see relationships among data types for the same establishments. For our review, that is very important. In Figure 5, the horizontal axis represents the change from the previous month to the current month in the current year, and the vertical axis represents the change from the previous month to the current month's value lagged one year (An exception is the All Employees scatter gram). Thus, points along the 45 degree line represent establishments whose data activity this year matched their own data activity last year. Zero month-to-month change this year and zero month-to-month change last year is represented in the very center of each scatter gram. As an option, the horizontal axis here for the All Employment scatter gram represents the actual magnitude of the establishment, rather than the change from last month, in short, the size of the firm. In this case, the vertical axis has zero point at the midpoint of the vertical axis,

Figure 5. Interactive Scatter Grams for All Data Types and Information by State, Comment Code and Firm



and represents the difference in month-to-month movement between this year and last year. Thus, when any point is selected by the mouse, the size of that establishment's employment is instantly obvious.

How would an industry analyst use Figure 5 on his or her PC? Selection of the industries for display of scatter grams is controlled by the selection box in the lower right hand corner. The number code of an industry is selected by the mouse, and the analyst can choose to cycle through a large group of industries one at a time, or view combined industries. The scatter grams can be limited to show only establishments of a certain size, which would be important in finding only the most significant sample outliers. Once the selection of an industry to view has been made, the scatter grams for that industry appear. At the same time, each state in the state map changes color to indicate an index of sample activity for that state. One can then see if an estimate level problem is limited to a single state or subgroup of states. The state map can also be used to plot scatter grams for only a particular state.

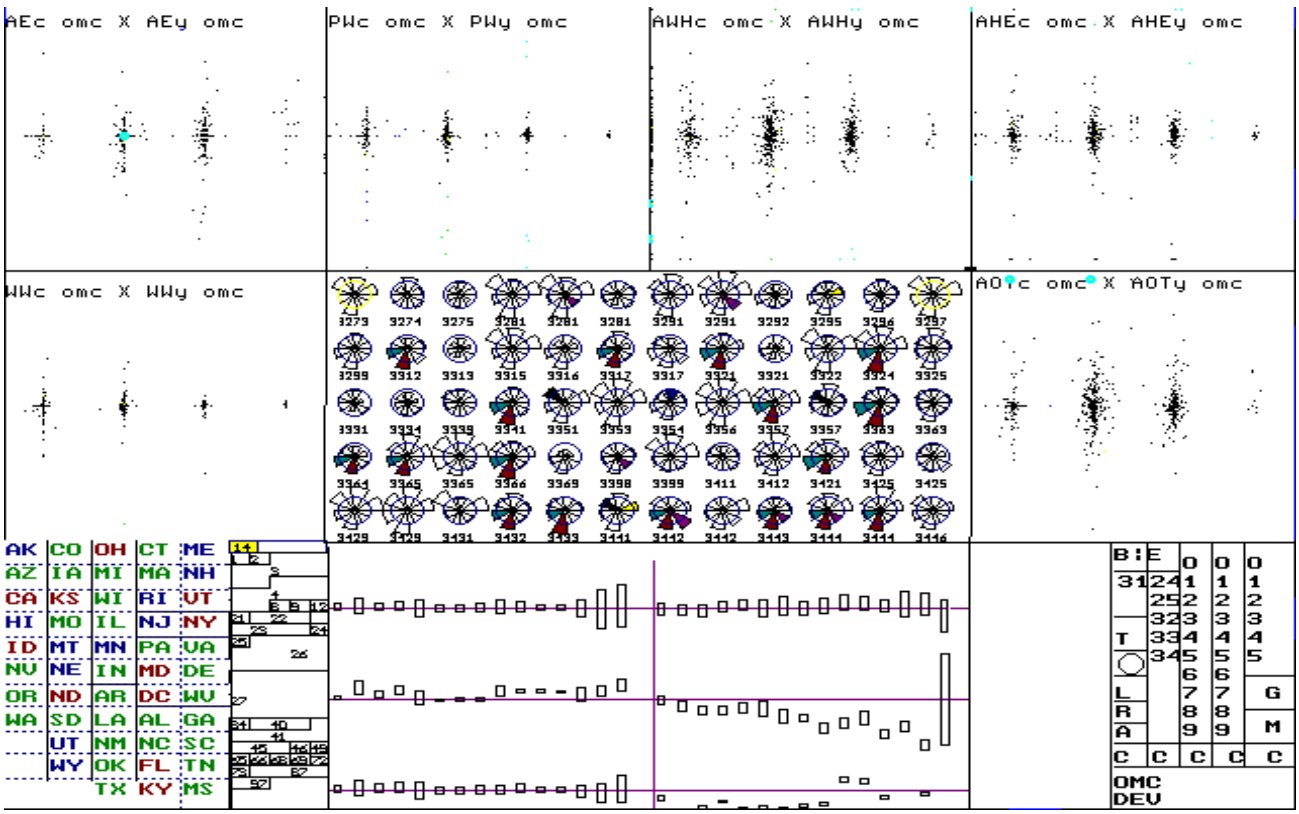
The pie chart to the right of the state map is a pie chart of the percentage of specific explanation codes used when reporting sample data. For example, a state agency will affix a "strike" code if sample values for an establishment have been affected by a strike. The pie chart thus shows at a glance some of the most important factors influencing the data in the current month. Codes which individually comprise less than two percent of all comment codes received are arranged as tiny boxes below the pie chart.

In Figure 6, Colors & Segmenting for Size Class, Rough Tukey, & Daisy Variations, we have segregated the sample into 4 size classes, and used colors and the horizontal axis to indicate the size class of the sample members, as well as the sample *movement* for 1 year ago, and on the vertical axis we have plotted the sample *movement*, rather than the raw values, for the current month in the current year. In this way, we can see the relationships of how sample behaved this year

in comparison to last year. The current month's movement will be the one attracting the most attention, so the vertical axis representing that is unreduced, while the one-year lagged movements are at one-quarter scale, and segmented by size class. At the center of this graph in the bottom section, Tukey box plots for the most recent 15 months of sample data for each of the six variables are plotted. Using these plots, the user can quickly see whether the standard deviation for the current month is substantially larger than previous months, and perhaps consider the presence of a sample outlier as the cause. The "daisy variations" comprise the very center of the graphic, and represent an index to the box plots. Each one of the 60 individual daisies is a 6-petaled representation of the current sample standard deviations for a single industry, and these daisies are used in a manner similar to the anomaly map. Each petal represents one of the six variables measured, and the "flowerpot" or circle within each daisy represents the average standard deviation (not including the current month) for the most recent 14 months for a single variable. Under the presumption that data for previous months are already more or less "clean", the user can then immediately see how large the current standard deviation is relative to the previous months' average by seeing how far over the average or edge of the "flowerpot" each petal is. If a daisy indicates a particularly large current standard deviation, we can mouse-click on the daisy and immediately see the associated scatter grams and box plots for that particular industry. The box plots shown here are for just one of the 60 industries represented by the central daisies.

In designing the prototypes, we were motivated to include a simultaneous picture of all 6 variables, and to include more immediately accessible 1-year lagged data for the sample members, so that we could take advantage of seasonal expectations and long-term trends. The daisies were invented so that we could see patterns in historical standard deviations, and to take advantage of the deviations as indicators of current sample reliability.

Figure 6. Colors and Segmenting for Size Classes, Rough Tukey, & Daisy Variations



5. CONCLUSION

We view the successful introduction of the ARIES system as just a beginning. The principles that we have followed and have been led to can be said to be the following:

1. Make the system applicable to the task at hand.

2. Try to make information digestible by using graphics.

3. Make the system interactive, to increase flexibility of use.

4. Make a system for which it is more natural to make improvements rather than a system for which it is more natural to just use and ignore. This means, make a system which gives information which is useful to guiding future improvements.

The graphics features of ARIES are designed to efficiently view micro-level sample data within the larger picture of industry level measures of movement. This principle is a continuing motivation for the prototype, which seeks to expand the capabilities of ARIES in several directions. The daisy variations and associated Tukey plots move the system towards more formalized and useful statistical representations for the data. While originally designed for outlier detection, these representations will be used to investigate in greater depth the issues of sample adequacy and estimate quality. The use of computer graphics has given us a powerful method to simultaneously obtain a better view of data at multiple levels of analysis, and our future research will be on how to further benefit from the comprehensive vision made possible by this approach.

REFERENCES

[1] Esposito, Fox, Lin, and Tidemann. ARIES: A Visual Path in the Investigation of Statistical Data, *Journal of Computational and Graphical Statistics*, Volume 3, Number 2, June 2, 1994, pp.113-125.

[2] Esposito, Lin, and Tidemann. The ARIES Review System in the BLS Current Empoyment Statistics Program, *ICES Proceedings of the International Conference on Establishment Surveys*, Buffalo, New York, June 27-30, 1993.

[3] Esposito, Fox, Lin, and Tidemann. ARIES: Visual Techniques for Statistical Data Investigation at the Bureau of Labor Statistics, *American Statistical Association 1994 Proceedings of the Section on Statistical Graphics,* 1994.

[4] Galitz, Wilbert 0. *Handbook of Screen Format Design*, QED Information Sciences, Inc.,Wellesley, Massachusetts, 1989.

[5] Granquist,L. On the Role of Editing, *Statistisk Tidskrift,* 2, 1984, pp.105-118.

[6] Hughes, Phillip J. , McDermid, I. , Linacre, Susan J. The Use of Graphical Methods in Editing, *U.S. Bureau of the Census: 1990 Annual Research Conference Proceedings*, 1990, pp.538-550.

[7] U.S. Department of Labor. Employment, Hours, and Earnings from the Establishment Survey, *BLS Handbook of Methods*, 1988, pp.13-40.

[8] Madow, Lillian H. and Madow, William G. On Link Relative Estimators, *ASA Proceedings of the Section on Survey Research Methods*, 1978, pp.534-539.

A GRAPHICAL MACRO-EDITING APPLICATION

by Per Engström and Christer Ängsved, Statistics Sweden

ABSTRACT

This paper presents a graphical macro-editing PC application made for the Short Periodic Employment Survey at Statistics Sweden. The authors conclude that developing graphical macro-editing tools is neither difficult nor time-consuming using modern developing software.

Keywords: macro-editing; interactive graphics; Visual Basic; scattergram

1. INTRODUCTION

The main purpose of this paper is to describe a newly developed graphical macro-editing system made with Visual Basic as programming tool. The application is made for the establishment survey Short Periodic Employment Survey, Private Sector (SEP).

The main advantages of the PC-system are:

I) Data editing can easily be integrated with data entry (in this survey the subject matter personnel do the data entry and editing, which includes contacts with the respondent);

ii) Hardly any further editing is needed after the input editing which evidently is effective and sufficient (although the bounds for the statistical checks are wide);

iii) Survey officers have better access to the data (compared to the main frame) and can interactively query the data base to get information.

After a system revision of the SEP survey in 1993 we decided for a Client/Server solution with Microsoft SQL Server as the back-end and a Visual Basic application as the front-end.

Since a macro-editing method [2] was chosen, it was decided that the input-editing should concentrate on consistency checks, and the statistical checks should be focused on identifying gross errors with big influence on estimates. Inspired by the ARIES system [1] we applied a graphical solution to review data causing suspicious estimates. A working prototype with all the desired features was developed within two weeks.

2 THE GRAPHICAL MACRO-EDITING APPLICATION

2.1 Description of the SEP Survey

The survey is a quarterly establishment survey in the private sector. The sample consists of about 12000 establishments. The main purpose is to measure employment (permanent and temporary employed) at detailed industrial level. Other variables are absence, newly recruited and leaving persons.

2.2 Input-Editing

By input-editing we mean editing at the data entry stage. No editing is carried out before data entry. The reason is to keep the original data automatically for later studies.

The checks are focused on identifying formal errors and extreme values. For instance vacancies should not be more than the total of employees for a specific establishment (consistency check). Big changes compared to the previous period have to be identified, because that could indicate that a respondent for an establishment has responded wrongly for the whole enterprise.

If errors are rare and random it is not reasonable to have a lot of statistical checks with tight bounds flagging many records. Generally respondents are very competent and the questionnaire is easy to fill in. Therefore statistical checks are focused on such big differences compared to the previous period which have a notable influence on the estimates.

Studies to set up reasonable bounds for the statistical checks indicate that a record should not be flagged unless the influence on the estimate is greater than approximately one percent.

2.3 The Macro Editing Application

The macro editing procedure focuses on the estimates of employment at the detailed industry level. Few important errors are expected in these data.

Most likely the outliers are due to high raising factors together with extreme natural increases (or decreases)

in employment.

To find suspect estimates is a key task in macro-editing. Today we are doing this manually by looking at the sample variance and the relative change, but we intend to do it automatically. We believe that the checks for detecting suspicious estimates should be based on time series analysis and sample variance. The suspicious estimates can be shown on the screen, and by clicking one of them the data of the individual observations of the selected aggregate can be displayed.

The industry code is entered for the suspicious estimate, and then data of the aggregate are plotted into a scatter gram. Each point represents an individual observation of the aggregate with the current quarter's reported data (the vertical axis) and the previous quarter's reported data (the horizontal axis).

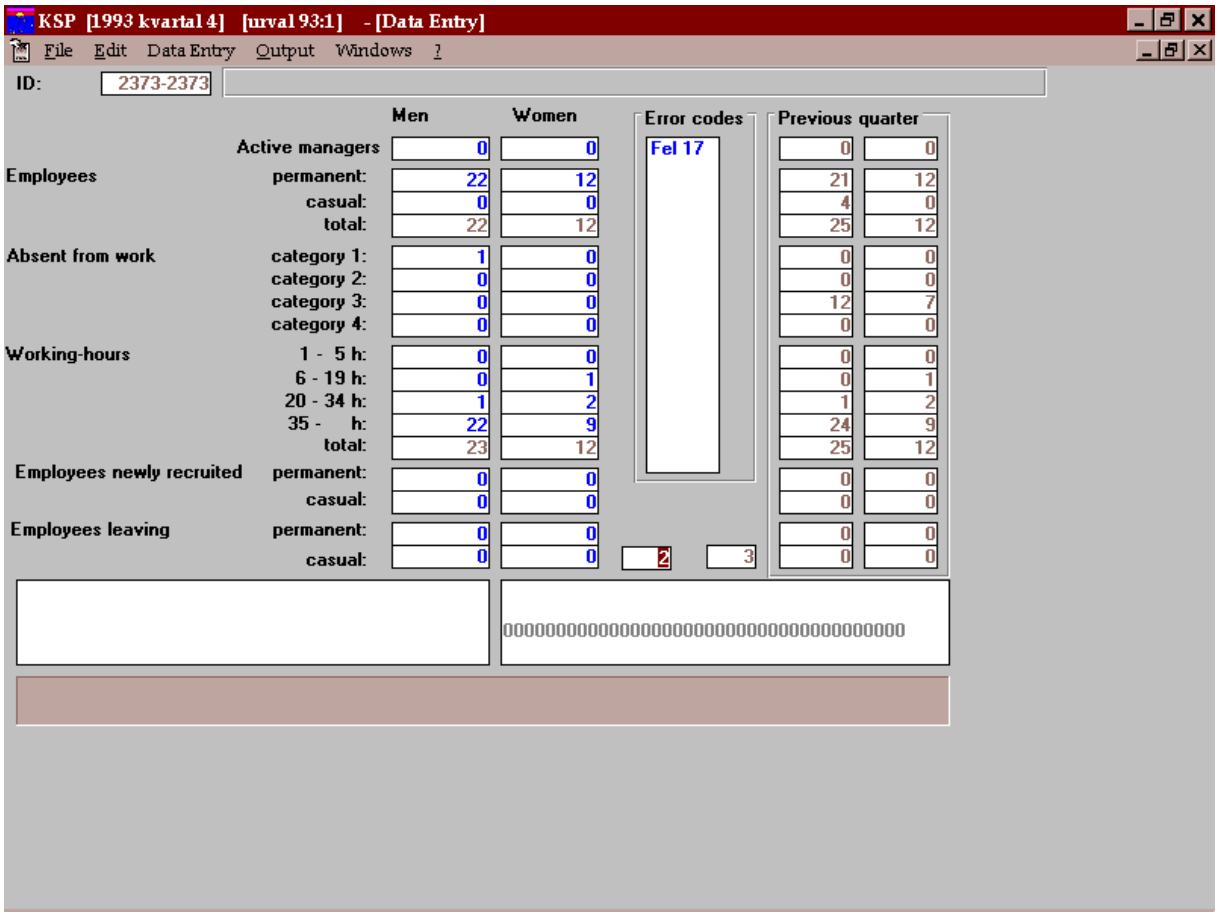
Clicking on an observation will cause the following data to be shown: current number of employees, previous quarter's figure, raising factor and a special parameter called Fraction Of Estimate (F.O.E.), that is the ratio between the weighted number of employees and the estimate.

Outliers which have an influence on estimate level are defined by relative interperiod change level.

Outliers are coloured red on the display. It is informative to interactively test various limits. Therefore it is possible to change the limits and to see how many observations are classified as outliers. When suitable limits are found the preliminary limits are substituted in each case.

To review a suspicious record, the editor double-clicks with the mouse on the record's point. Then the data entry application takes over and all the data of the record are displayed (the screen looks similar to Figure 1). When data are changed the input-edit procedure starts, to assure that no new errors are committed and to check that no other errors were hidden by the faulty data . When data are changed the scatter gram and all the influenced parameters are updated.

Figure 1. Data entry screen after error indication



It is possible to zoom into the observations. This is needed when the observations are tightly scattered. Zooming is done by drawing a box with the mouse around the area of interest.

2.4 Screen pictures

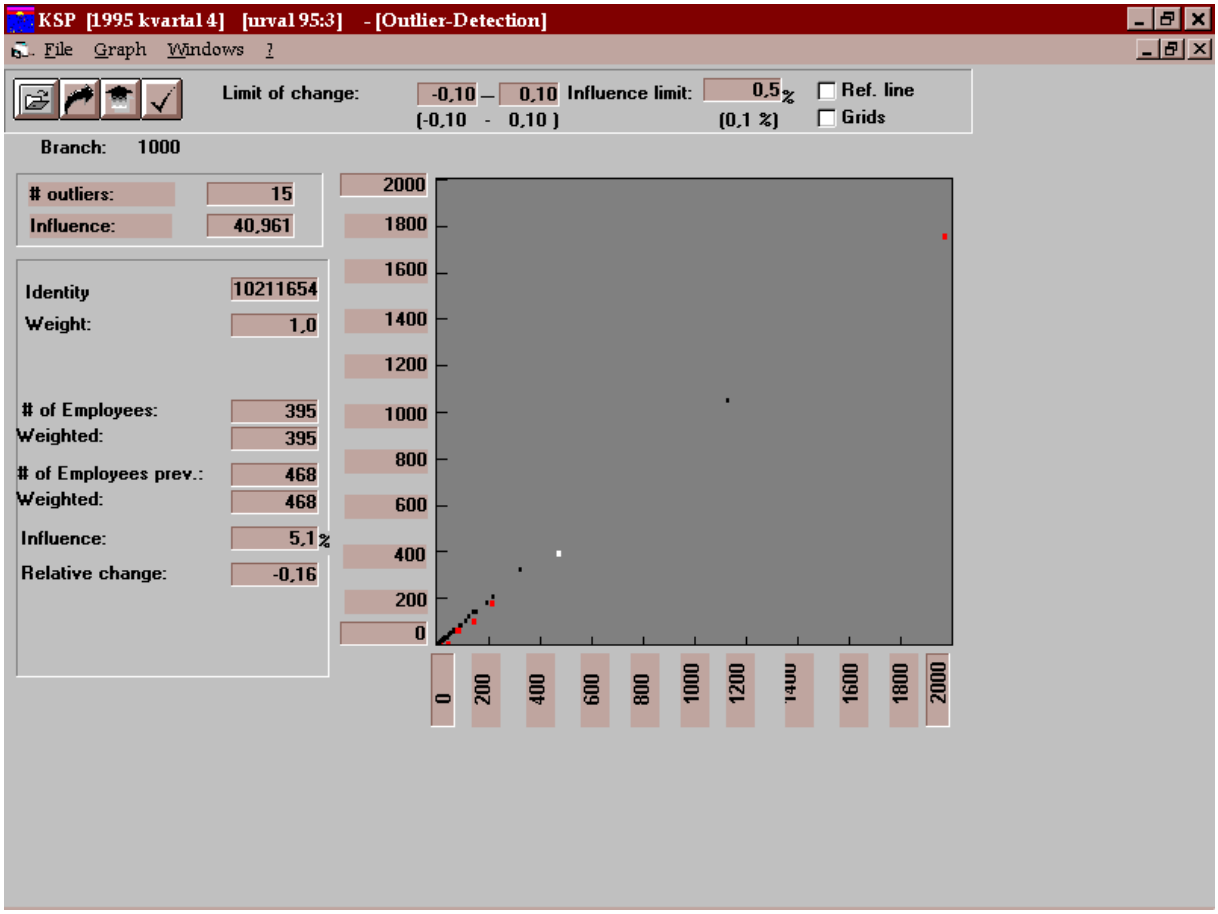
2.4.1 Input editing

Data editing is integrated with the data entry. When the application has indicated an error the screen looks like Figure 1 (from the start only the two left columns are shown). The two left columns are the current quarter's report. On the screen, the flagged items have red colour. The column in the middle shows error codes. The two right columns show the previous quarter's report.

2.4.2 Macro-editing

After the editor has selected the industries which data needs clarification, he starts the macro-editing application. Based on the industry code, the required data is retrieved from the SQL database. Based on that, the scatter gram is displayed (Figure 2). By selecting an observation, it is possible to get more detailed information about this observation. That information is shown to the left. The limits for identifying outliers are shown at the top of the scatter gram.

Figure 2: Scatter gram for a selected industry branch



2.5 Experiences with Visual Basic

Visual Basic version 3.0 (VB) for Windows has been used as developing tool for the graphical application and for almost the whole SEP system. With this tool it was possible to develop the graphical macro-editing application in a very short time.

VB proved to be fully sufficient for this type of applications. It is excellent for creating (and changing) a user interface.

A potential disadvantage is that there could be performance problems displaying a large number of observations (say more than 5000) with the present application. It is not reasonable to plot more than 1000 observations. In this case, it is better to macro-edit at more detailed levels instead.

3. CONCLUSIONS

Graphical applications are superior in reviewing data and looking for anomalies. For instance a scatter gram has the following positive properties (among many others):

- A total view of data is provided.
- Special patterns like a general decrease can easily be seen.

For repetitive establishment surveys it is a good idea to make a graphical macro-editing application. Various methods could be used depending on the specific conditions of the survey.

Our experience shows that graphical applications are not costly to develop. In fact with the right programming tools and some programming skills it is rather easy to introduce graphical methods into the survey process.

Because of the similarities between short periodic business surveys it is a good idea to build graphical programs in a modular way (of course that also goes for the data entry program). Then it is easy to reuse programming code for other surveys.

REFERENCES

[1] Esposito, R. The ARIES System in the BLS Current Employment Statistics Program, *1993 Proceedings of the International Conference on Establishment Surveys*, 1993.

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

[3] Gary Houston, Andrew G. Bruce gred: Interactive Graphical Editing for Business Surveys, Journal of Official Statistics, 1993, pp. 81-90.

[4] Hughes, Phillip J. , McDermid I. , Susan J. Linacre. The use of Graphical Methods in Editing, *U.S. Bureau of the Census: 1990 Annual Research Conference Proceedings*, 1990.

THE GRAPHICAL EDITING ANALYSIS QUERY SYSTEM

By Paula Weir,Energy Information Administration, Department of Energy
Robert Emery and John Walker, Science Applications International Corporation, USA

ABSTRACT

The Graphical Editing Analysis Query System (GEAQS) is 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 highlight 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.

Keywords: top-down editing; exploratory data analysis; Box-Whiskers graphs; dialogue boxes.

1. 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" [7]. The paper presented the subcommittee's findings that median editing cost as a percentage of total survey costs was 40% 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.

2. 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 selected 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 [3]. 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.

3. 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.

Figure 1

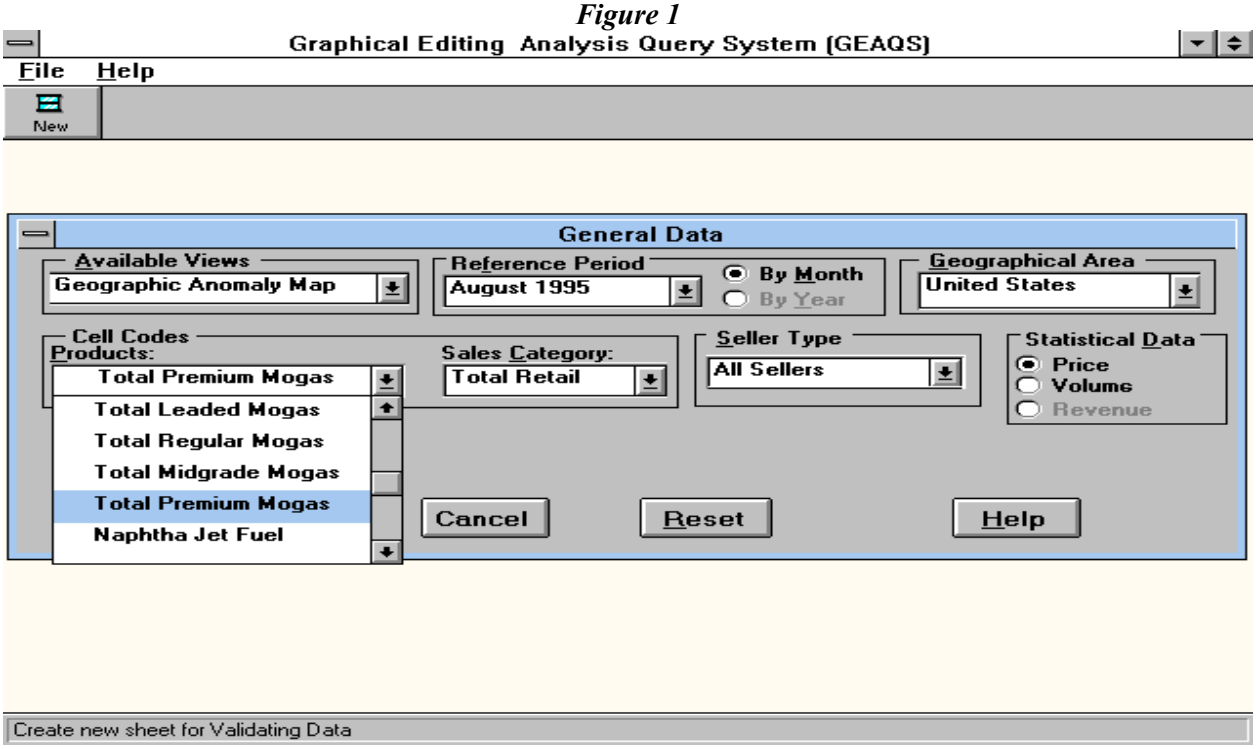
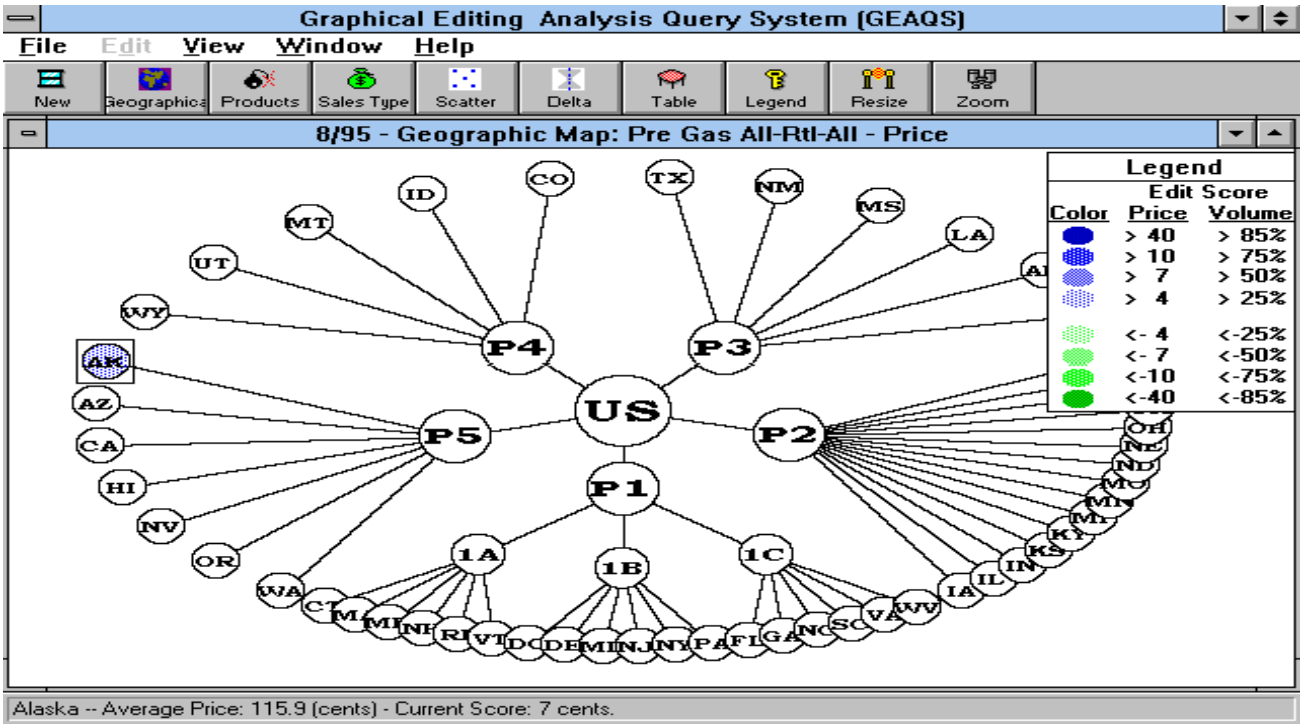


Figure 2



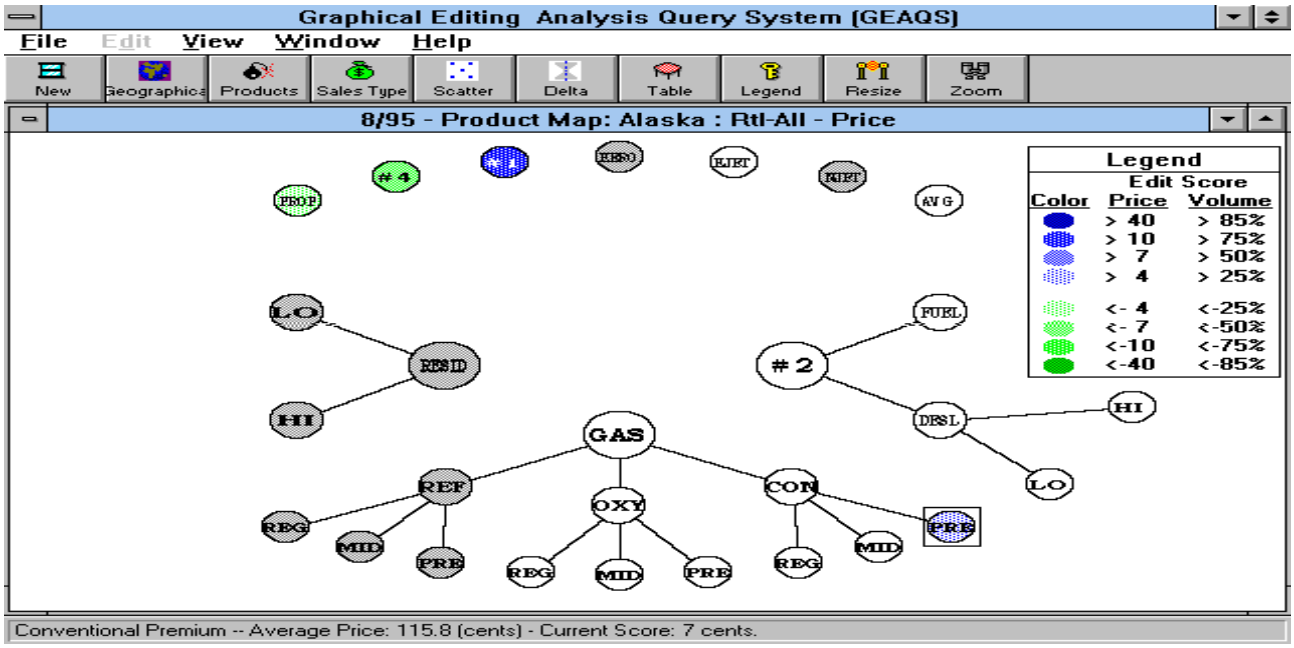
The geographical view requires the user to also select a product, sales category, seller type, statistical data type, and reference period from the drop-down lists provided by clicking within the respective boxes. As the user makes the view selection,the system adjusts the possible product selection, according to the combinations of aggregates calculated by the survey's processing system. Similarly, as the user selects the product, the list of 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:

$$(P_{k,t} - P_{k,t-1}) - (P^*_{\cdot,t} - P^*_{\cdot,t-1}),$$

where $P^*_{\cdot,t}$ is the PADD average price excluding state k.

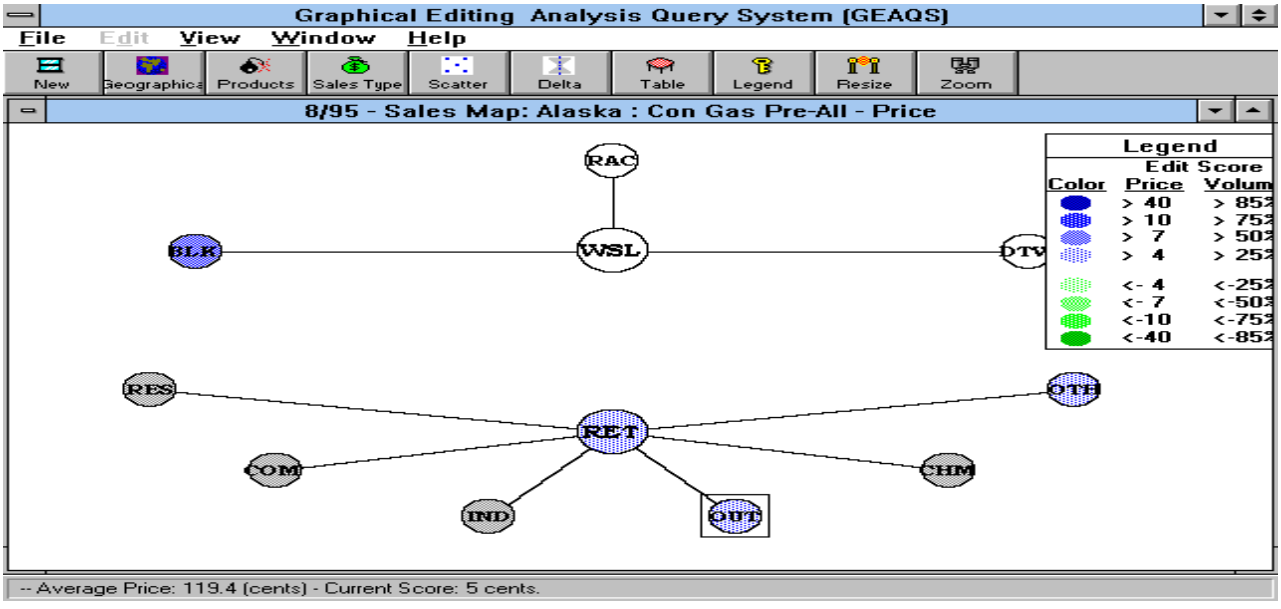
Volume and revenue current scores are similar, but use the difference in percent change between the state and the PADD or subPADD. The current scores for the U.S. and PADDs are just the price change between the previous and current period. Four shades of blue are used to represent scores that indicate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

Figure 3



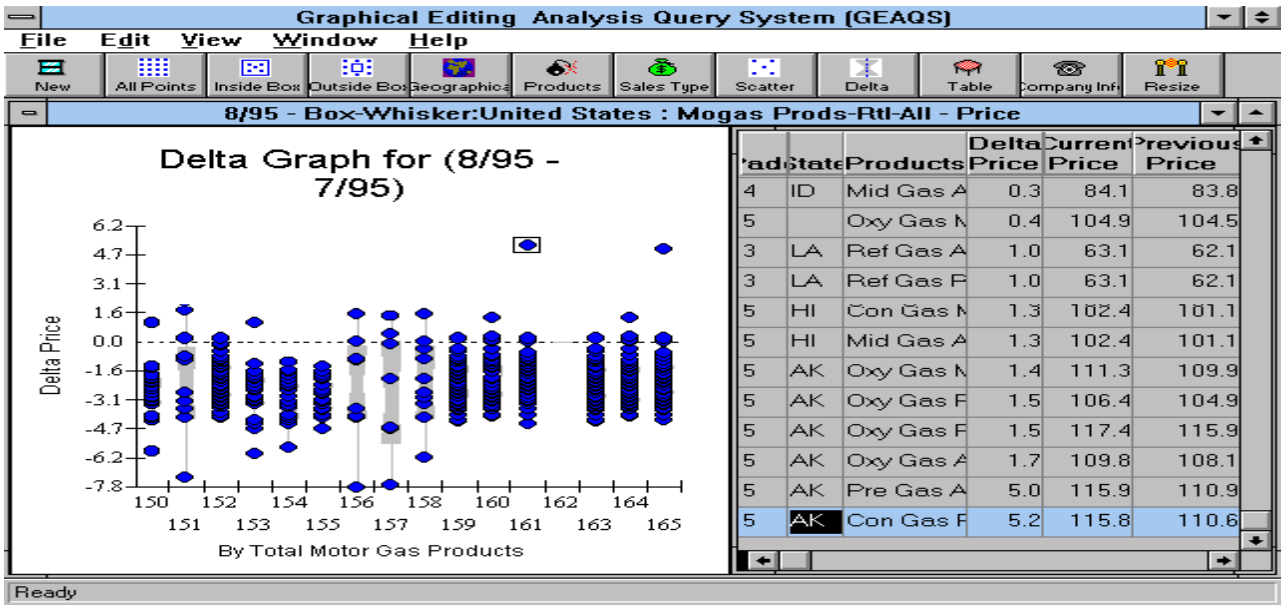
The user may drill down by either clicking on the products or sales type icon. If the user had previously selected a product that can further be broken down to the reported product level, the user would choose the product's icon. The window would be replaced by a new graphic, a product anomaly map (figure 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 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% of the values for the geographic changes, within the whiskers, and outside the whiskers, which are called outside values. 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

Figure 5



Pad	State	Products	Delta Price	Current Price	Previous Price
4	ID	Mid Gas A	0.3	84.1	83.8
5		Oxy Gas N	0.4	104.9	104.5
3	LA	Ref Gas A	1.0	63.1	62.1
3	LA	Ref Gas P	1.0	63.1	62.1
5	HI	Con Gas N	1.3	102.4	101.1
5	HI	Mid Gas A	1.3	102.4	101.1
5	AK	Oxy Gas N	1.4	111.3	109.9
5	AK	Oxy Gas P	1.5	106.4	104.9
5	AK	Oxy Gas P	1.5	117.4	115.9
5	AK	Oxy Gas A	1.7	109.8	108.1
5	AK	Pre Gas A	5.0	115.9	110.9
5	AK	Con Gas P	5.2	115.8	110.6

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 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% or more are depicted by red, values that represent 5% or more, but less than 50%, are yellow, and the remaining values, less than 5% 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).

Figure 6

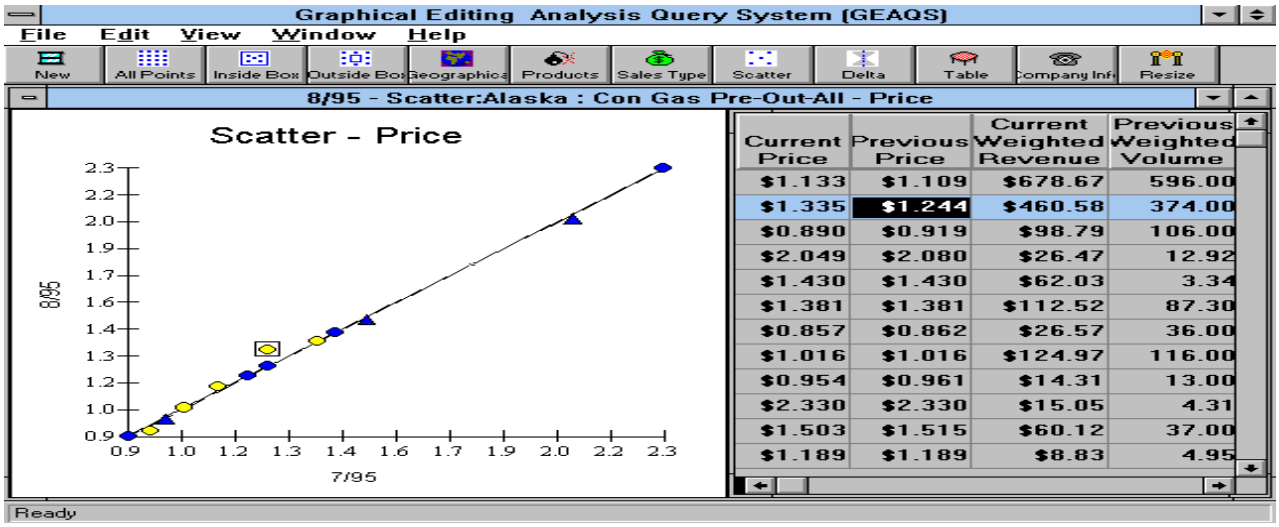


Figure 7

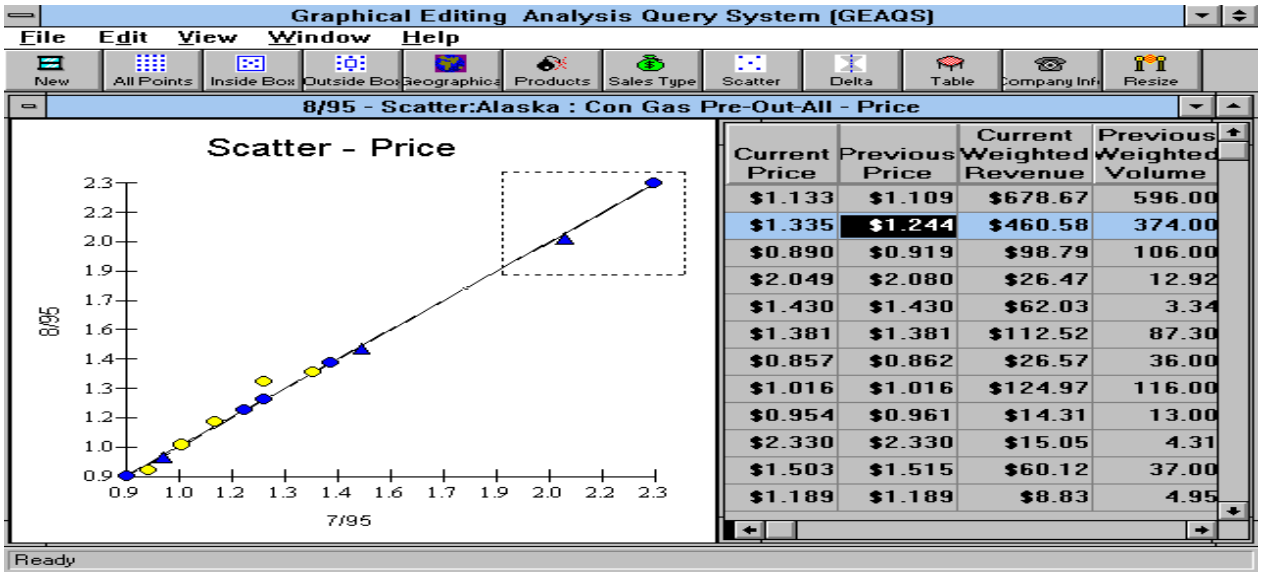
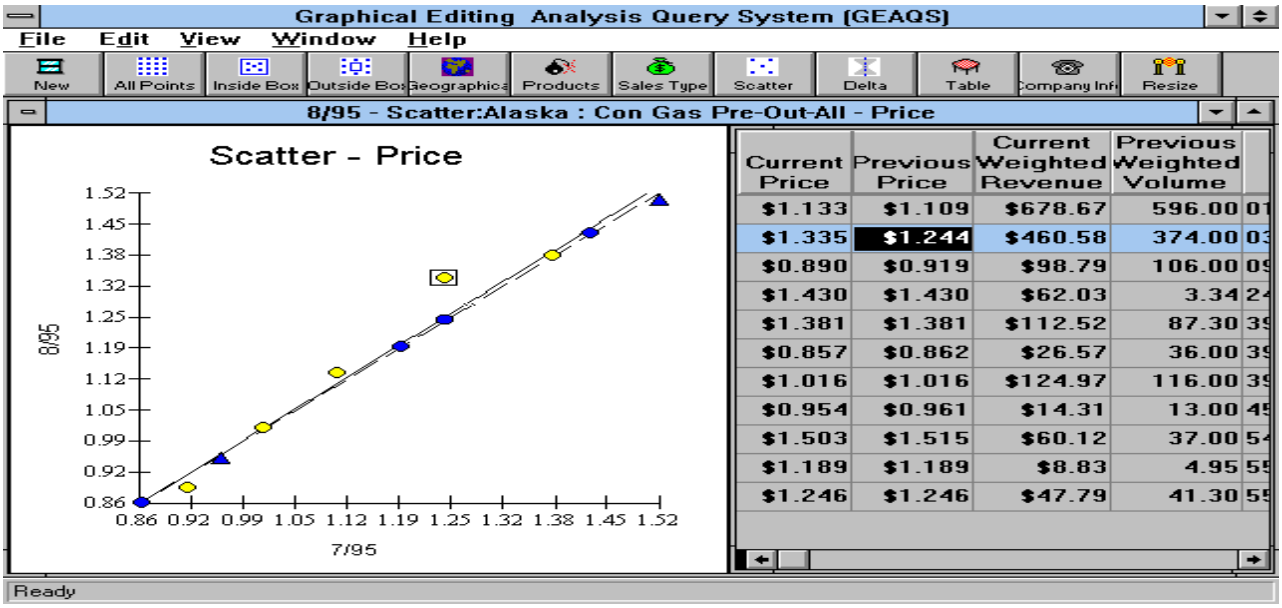


Figure 8



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.

4. 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 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.

REFERENCES

[1] Bienias, J., Lassman, D., Scheleur, S. And Hogan, H. Improving Outlier Detection in Two Establishment Surveys, *ECE Work Session on Statistical Data Editing, Working Paper No. 15*, Athens, November 6-9, 1995.

[2] Cleveland, William S. Visualizing Data, Hobbart Press, Summit, New Jersey, 1993.

[3] Engstrom, P., Angsved, C. A Description of a Graphical Macro Editing Application, *ECE Work Session on Statistical Data Editing, Working Paper No. 14*, Athens, November 6-9, 1995.

[4] Esposito, Lin and Tidemann. The ARIES Review System in the BLS Current Employment Statistics Program, *ICES Proceedings of the International Conference on Establishment Surveys*, Buffalo, New York, June 27-30, 1993.

[5] Mowry, S., Estes, A. Graphical Interface Tools in Data Editing/Analysis, *Washington Statistical Society Seminar presentation*, March 10, 1995.

[6] Subcommittee on Data Editing in Federal Statistical Agencies, Federal Committee on Statistical Methodology, 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, 1990.

Chapter 4

EVALUATION OF DATA EDITING PROCESS

FOREWORD

by Leopold Granquist, Statistics Sweden

This Chapter aims to serve as a basis to elaborate and finally establish standards for useful indicators and/or statistical measures on the rationality and efficiency of the data processing process, and indicating the problem areas of the data collection. What should be measured and how it should be done are the underlying issues throughout the Chapter.

An important step towards the mentioned goal is taken in the first paper. It outlines the general requirements for a computer system to measure and monitor the impact of data editing and imputation for the 2001 UK Census of Population. The UK Office for National Statistics identified the need when evaluating the 1991 Census operation. The paper documents the requirements for the "Data Quality Monitoring System" (DQMS), which have been gathered so far. DQMS will be developed iteratively on a prototype basis and the requirements will be enhanced or re-prioritized as work proceeds. One key requirement will be to allow data experts to check the assumptions built into the capture, coding, editing, and imputation of census data, and the impact of these assumptions on the data. This requirement is broken down into two parts, standard reports and ad-hoc enquiry facilities. The latter will allow intuitive and complementary analysis of the data. A number of standard reports and requirements for the ad-hoc reports are proposed. All of them will tell implicitly or explicitly what is recognized as important to measure in processing census data.

The second paper, written by Bogdan Stefanowicz proposes indicators on the rationality and the effectiveness of the set of edits in detecting all errors. The author suggests improvements to take into account that different types of errors may have a different impact on quality. It does not deal with errors introduced by editors. The efficiency indicator involves the number of undetected errors which cannot be found from studies of error lists. The author suggests that it might be estimated by simulations, but does not discuss this issue further. The second part of the paper discusses the role of error lists in evaluating editing processes and as a basis for improvements of the data collection.

The third paper is an overview of selected evaluation studies. Most studies use the error list method in different ways, and perform analysis with the aid of computers. The rationality indicator suggested by Stefanowicz is used by some authors and called the hit-rate. A number of evaluations are focused on the efficiency of editing processes and raise the question whether resources spent on editing are justified in terms of quality improvements. In some cases the question can be answered by studying the impact of editing changes on the estimates. How to carry out such studies is also described in Chapter 1. However, those methods cannot be used for measuring the data quality. To obtain measures on how editing affects quality, it is necessary to conduct reinterview studies, record check studies or simulation studies. Examples of such studies are presented in the paper, which also provides hints as to how the various methods can be evaluated. Some results from almost all of the studies are given.

The fourth paper is a description of two evaluation studies, each one consisting of a comparison of micro data collected from two different data sources: survey data and administrative data. This is a unique situation. Firstly, among the evaluation studies discussed in the second paper by Granquist, there is no evaluation of statistics collected from administrative data files. Secondly, in general, data from administrative sources are not available for comparing survey micro-data with administrative data except for a few items from the register used as the sampling frame. Although the paper does not provide details concerning the editing methods, it covers many aspects of editing and evaluation. The need for resources to measure or assess the impact of editing is stressed.

STATISTICAL MEASUREMENT AND MONITORING OF DATA

EDITING AND IMPUTATION IN THE 2001 UNITED KINGDOM CENSUS OF POPULATION

By Jan Thomas, Census Division, Office for National Statistics, United Kingdom

ABSTRACT

The Census Offices of the United Kingdom have identified the need to measure and monitor the impacts of the processing systems on census data. Requirements have been identified for a computer system to be known as the "Data Quality Monitoring System" to give this information.

The system will produce a series of standard and ad hoc reports, and will provide comparisons of distributions before and after editing and imputation and simple cross tabulations at various area levels. The data can be visualised, possibly by geographical area, to see where the error is occurring. Work has started on a prototype system and it is hoped that the prototype will be developed for use in the 1997 Census Test.

In addition, it is planned to appoint a team of "data experts" to analyse and interpret the results that will be reported from the system.

This paper outlines the general requirements of the system and those specifically relating to editing and imputation.

Keywords: editing; imputation; measuring and monitoring the data quality.

1. INTRODUCTION

The UK Census Offices have identified the need to measure and monitor the impacts of processing systems on census data. The evaluation of the 1991 Census operation highlighted the fact that the facilities which were in place to monitor the quality of the data were inadequate, and were employed too late to have any impact on problem resolution. Research is currently underway to produce a computer system which will measure and monitor the data as it is being processed in the 2001 Census.

It is planned that the use of this system, to be known as the "Data Quality Monitoring System" (DQMS), will be extended to cover data capture, coding, derivation of variables and sampling. This paper considers the editing and imputation requirements only as they are relevant to the data editing process.

It is recognised that to operate this system a team of people who are experts in data analysis will be needed. It is planned to appoint a team of six "Data Experts", with the responsibility for monitoring data as it is processed.

2. BACKGROUND TO EDITING AND IMPUTATION IN THE BRITISH CENSUS

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.

Imputation was only carried out on the 100% questions and not on sample (10%) questions; this was because most of the 10% questions have lengthy classifications, such as occupation and hence are difficult to impute with any accuracy. Automatic imputation on a record by record basis was first introduced in the 1981 Census, and was based on the work by Felligi and Holt in the 1970s, the so-called hotdeck methodology. This worked well in 1981, and so was carried through to 1991 with only minor changes for new questions.

This paper documents the requirements for the DQMS which have been gathered so far. It is a working paper, as the DQMS will be developed iteratively on a prototype basis and the requirements will be enhanced or re-prioritised as work proceeds. The general system requirements are shown in italics for ease of reference and are classified as either "standard" or "ad-hoc". The specific requirements for editing and imputation are then listed.

Although the role of the data expert is not yet fully defined it is anticipated that they will be in place sometime during 1998-99, and that they will become familiar with the data during the Dress Rehearsal for

the Census which takes place in 1999. One way of organising the team would be to give them topic and area-specific responsibilities. The Data Experts could get to know a geographical area and geographical displays/boundaries could be available in a digital form. It might be possible to feed in some information based on the enumeration district (ED) grading so that hard to enumerate areas are apparent as possible problem areas from early on.

3. THE OPERATION OF THE DATA QUALITY MONITORING SYSTEM

The DQMS will be responsible for monitoring the quality of data from the point at which data is captured. It will be an automated system with standard and flexible outputs and will work only on machine readable data with the ability to both print and provide electronic output. It must fit seamlessly into Census processing and not be the cause of any delay to the operation. Counts taken during processing will be compared with previous census data and with the data from external sources. The DQMS may need to link to ArcInfo (used by the geography planners), spreadsheets and statistical interrogation packages.

3.1 General Requirements of the Data Quality Monitoring System

The key requirement of the Data Quality Monitoring System will be to allow the Data Experts to check the quality of the assumptions built into the capture, coding, edit, imputation, derivation and sampling of census data, and the impact of these assumptions on the data, in order that corrective action can be taken if a problem is discovered, or advice offered to census data users if it is not possible to correct the problem.

This requirements are as follows:

a) Standard reports specifically designed to highlight problems associated with the census processes.

b) Ad-hoc enquiry facilities which will allow intuitive and complementary analysis of the data.

3.2 Standard Reports

The purpose of standard reports will be to profile the data in such a way that potential problems are highlighted. The standard reports identified are as follows:

a) Automatic monitoring of samples of the data during processing, with deviations (above pre-set thresholds) from the expected numbers, (from external sources), highlighted.

b) Standard distributions and simple cross tabulations at various area levels.

c) Counts of record types at the start and end of individual processes (i.e. coding/edit/ /imputation/derivation) showing at the start, missing and invalid fields, and at the end, changes to field values. All changes to be cross-referenced to the source of the change.

d) Analysis of the types of error and cumulative counts of errors, in broad groups.

e) Pre-defined outliers highlighted for review.

f) The geographical areas associated with poor coverage and poor form completion to be highlighted. This information will come form the field enumeration operation.

3.3 Ad-hoc Enquiries

The ad-hoc enquiry system will allow further analysis of data in order to identify and investigate problems. The system should allow easy interrogation, analysis and visualisation of the data. The system should provide:

access to the following datasets :

- Main Census Database
- Geography spacial datasets
- External reference datasets
- Counts database
- Processing control information
- Form control
- Problem Management System

The above information will allow the Data Experts to:

- Extract and aggregate data, (the lowest level information accessed may be an individual form. At the other extreme data may be aggregated to GB/UK level).

- Perform statistical interrogation and analysis, and to "dig down" to lower levels to investigate suspect areas.

- Visualise the data, possibly by geographical area,

to see where an error is occurring.

Specific requirements for these ad hoc reports are as follows:

a) the possibility to see the characteristics of a certain group of data;

b) The facility to create temporary simple derived variables and to cross tabulate them;

c) A test system on which to try out changes to the editing and imputation systems, and to review the results of these changes by using such DQMS tools as cross tabulation and distribution comparisons. Note: It is considered mandatory that a test system is available to allow the Statisticians and Data Experts to review proposed changes to edit and imputation programs, and to ensure that the changes give the required result, and do not cause inconsistency elsewhere. It is unclear at this point if the system needs to be a part of the DQMS or a stand alone tool;

d) The ability to link to Ordnance Survey maps to see that there is perhaps a prison or communal establishment which is giving a distorted figure;

e) Before a full database is reached, part way through processing, the ability is required to gross up and carry out predictions and simple modelling to show the likely outcome for a given area;

f) An algorithm or method to spot patterns in the data is needed, (these are already being used in the chemical and engineering fields - perhaps a neural network could help here.)

g) Tracking of the history of a particular record. Information will be required on what the various stages of the census processing have done to data items.

4. SPECIFIC REQUIREMENTS OF EDITING AND IMPUTATION

4.1 Editing

In 2001, the editing processes will identify missing, invalid or inconsistent items within the data. It is likely for the full editing system that the 1991 edit matrices approach will be used, and this will replace some inconsistent data with consistent, valid values and mark the remaining items for imputation. The DQMS will need to check the assumptions implicit in the editing rules.

It is a requirement that the DQMS system produce the following reports:

- The number of accesses to individual cells of edit matrices. This could give indications that the edit rules are not working or incorrect assumptions have been made.

- Access to every record containing an error, identifying the items in error and error type by geographical area.

- The totals of household, person and communal variables for each ED before and after edit and comparison with external sources. This could be an exception report showing those EDs which mismatch by a predefined amount with a facility to aggregate and interrogate.

- Details of EDs that fall below the pre-determined tolerance levels for errors.

- The totals of the variables of Communal establishments having nine or more people present, before and after edit, to provide a check against the communal establishment categories, so that a validation of establishment type code can be ascertained prior to imputation.

- Edit preference rules that are causing problems i.e. a high degree of secondary editing based around the acceptance of priority variables.

4.2 Imputation

The objective of Imputation will be to substitute appropriate data for missing, invalid or inconsistent values. The DQMS will need to check the assumptions implicit in the imputation rules.

DQMS will be required to produce the following reports:

- The overall number of imputed items, by variable.

- The overall number of household and people created under absent household imputation.

- The totals of household, person and communal variables for each ED, before and after imputation and comparison with external sources. This could be an exception report showing those EDs which

mismatch by a predefined amount, with a facility to aggregate and interrogate.

- Distributions before and after imputation.
- Predictions of final imputation distributions part way through processing.
- Details of any imputed outliers or rare occurrences.

4.3 Test of the Census 1997

Work has commenced on the development of a prototype system and a skeleton DQMS will be developed for the 1997 Census Test. This will be used to assess the prototype system in a semi-real environment and to refine the requirements. It has been suggested that it would be useful for some Data Experts to be involved in the 1997 test and for all to be involved in the 1999 Dress Rehearsal.

SELECTED ISSUES OF DATA EDITING

by Bogdan Stefanowicz, Warsaw School of Economics, Poland

ABSTRACT

The paper presents some proposals concerning estimation of the quality of data editing rules. An efficiency indicator and a rationality indicator of editing are proposed. They are considered as a function of detected and undetected errors in the input file.

The second part of the paper concerns information that can be derived from the error list accompanying the editing process. Special attention is paid to so-called "suspicious" cases signalled by the editing rules as the source of information for manifold applications.

Keywords: Evaluation of data editing; data imputation.

I. INTRODUCTION

This paper proposes some quality indicators to measure the editing process. The first two indicators are calculated based on the number of detected and undetected errors on the input file. Checking procedures detect some errors in the input file through editing rules, but a certain number of errors still remain undetected.

2. QUALITY INDICATORS OF EDITING RULES

2.1 The efficiency indicator

Let N_d be the number of signaled data items as detected errors in the input file F through editing rule R; let N_u be the number of undetected errors in the same file F. Then the indicator I_e:

$$(1) \qquad I_e = \frac{N_d}{N_d + N_u}$$

informs what fraction of errors can be detected by the rule R out of all ($N_d + N_u$) existing in F. The highest value of I_e ($I_e = 1$) is when R detects all real errors in F. In this case, N_u is equal to 0. The lowest value of I_e ($I_e = 0$) is when the number of detected errors is equal to 0 ($N_d = 0$) which means that no error in F is detected by R.

I_e may be used as a quality measure of R: if I_e is high then R is properly defined as a criterion for error detection; if it is low then the rule fails and must be improved.

I_e describes R from the point of view of its efficiency and says nothing about its rationality; interpreted as the ability to reject from F only those data which are real errors. In practice, any rule R may reject from F some number of "suspected" data which in fact are true. These data must be analyzed and are often "corrected" so as to fulfill R. Data of this kind will be called dummy errors.

2.2 The rationality indicator

The rationality indicator I_s can be defined as follows:

$$(2) \qquad I_s = \frac{N_e}{N_e + N_t}$$

where:

N_e is the number of real errors detected in F by the rule R;

N_t is the number of true data signaled by R as erroneous (dummy errors).

The higher the value of I_s, the more rational is R. Its highest value ($I_s = 1$) is when N_t equals 0 which means that all data items signaled by R are real errors. The lowest value of I_s ($I_s = 0$) is when no signaled data are erroneous or, in other words, all flagged data are dummy errors.

2.3 The combined efficiency and rationality indicator of a rule R

A comparison of I_e and I_s makes it possible to evaluate R in the light of its effectiveness. The indicator I_E of this characteristic can be defined as follows:

$$(3) \qquad I_E = I_e \times I_s$$

I_E takes on the lowest value ($I_E = 0$) when either I_e or I_s equals zero, and $I_E = 1$ when both I_e and I_s are equal 1.

To calculate I_e and I_s, it is necessary to have N_d, N_u, N_e, and N_t:

- N_d (the number of data items flagged by R) can be easily counted from the printed list of errors (E-list);
- N_u (the number of undetected errors with use of R) requires a simulation technique;
- N_e (number of real errors detected by R) can be counted from E-list after its manual verification by specialists;
- N_t (number of dummy errors) can be counted from E-list as well.

Some limited experiments were carried out by CSO in Warsaw at the end of the '70's in evaluating editing rules of annual reports on schools in Poland. Some editing rules were established concerning the number of classes (N_c) in school, number of teachers (N_t), number of class-rooms (N_r) and number of pupils in one class (N_p):

$$(4) \quad \begin{cases} (1) & |N_c - N_t| < a; \\ (2) & |N_t - N_r| < b; \\ (3) & c < N_p < d. \end{cases}$$

At the beginning, it was established that a = 1, b = 1, c = 20, d = 25. As a result, it appeared that the ratio R1 of dummy errors (to all flagged data) signaled by the rule (1) was 37% (R1 = 37%), the ratio R2 of dummy errors signaled by the rule (2) was 29% (R2 = 29%), and the rule (3) signaled 52% of dummy errors (R3 = 52%).

After some changes of the values a, b, c, and d in (4), the former values of R1, R2 and R3 also changed:

a = 3 ---> R1 = 9%,
b = 2 ---> R2 = 11%,
c = 20, d = 30 ---> R3 = 19%.

These facts inspired the consideration of the ratio of dummy errors as a measure of the rationality of editing rules.

2.4 The weighted efficiency indicator

The indicators shown above are calculated with the use of the total number of errors only. In fact, it is easy to modify them so as to take into account the influence of different groups of errors on the quality of output information. Such an influence can be expressed by appropriate weights included into formulae (1) and (2).

The idea of such a modification is based on the assumption that the numbers N_d, N_u, N_e, and N_t in the formulas (1) and (2) can be replaced by their weighted counterparts TN (e.g. N_d by $T_d N_d$).

For example, one can write:

$$I_e = \frac{(N_1^d \omega_1^d + \ldots + N_i^d \omega_i^d + \ldots + N_k^d \omega_k^d)}{(N_1^d \omega_1^d + \ldots + N_i^d \omega_i^d + \ldots + N_k^d \omega_k^d) + (N_1^u \omega_1^u + \ldots + N_j^u \omega_j^u + \ldots + N_k^u \omega_k^u)}$$

where:

- N_i^d (for i=1,...,k; k - number of established groups

of errors) - the number of detected errors in the i-th group;

- T_i^d (for i = 1,...,k; k - as above) - the weight of the i-th group;
- N_j^u (j = 1,...,k) - number of undetected errors (e.g. estimated with use of simulation) in the i-th group;
- T_j^u - the weight of the j-th group.

The weights w^d and w^u can be established with regard to the influence of the given group of errors on the quality of the output information.

It is assumed here that $N_1^d + ... + N_k^d = N_d$ and, $N_1^u + ... + N_k^u = N_u$. If all $T_i^d \Lambda T_i^d$ are equal 1 then the new formula can be transferred into (1).

A similar transformation of the formula (2) for N_s can be done in the same way.

3. ERROR LIST AS A SOURCE OF SUPPLEMENTARY INFORMATION

An analysis of the error list (E-list) shows some interesting characteristics and its usefulness for manifold purposes:

Manual correction of errors. This is the basic function of the E-list in practice in most verification processes.

Evaluation of input data quality. A comparison of the number and distribution of errors detected in the input file F with the user's requirements concerning information quality makes it possible to evaluate whether F meets these requirements. The type and kind of distribution of errors in F may be significant. This information can be used, for example, to determine if errors are systematic.

Analysis of sources of errors. A manual inspection of the E-list delivers information about sources which cause errors signaled in the list. This kind of information makes it possible to improve data collection and other procedures by eliminating or at least reducing errors in the next cycles of the statistical investigation.

Evaluation of verification rules. A manual inspection of the E-list makes it possible to verify signaled dummy errors and editing rules R which caused their appearance in the list. Such errors occur because of the improper definition of R. If the number N_t is high in comparison with all signaled errors then R should be verified on the basis of information contained in the E-list.

Analysis of new tendencies in the observed area. A correction of R in the case of large N_t is necessary because of its inconsistency with the real distribution of the observed variables. In other words, a large number of dummy errors may prove the existence of a new, unknown tendency in the distribution of the variables. Thus, the dummy errors listed in the E-list present interesting information about facts that may otherwise not be discovered. They may simplify detection of some unobservable characteristics of the piece of reality under investigation. In this sense, they can help to update to some extent the expert's knowledge.

Determining "untypical" cases. Logical editing rules R are defined so as to detect anything that is "untypical" or improbable. Any dummy error flagged by such a rule is a special case in comparison with all other input data. This is why dummy errors present important information for specialists who are interested not only in aggregated results but also in exceptional cases.

AN OVERVIEW OF METHODS OF EVALUATING DATA EDITING PROCEDURES

by Leopold Granquist, Statistics Sweden

ABSTRACT

An overview of a selection of available papers dealing with evaluation studies on editing processes and methods is given. The aim is to present those methods and techniques for evaluations of different kinds that have been used in practice. The focus is on the purpose of the evaluation undertaken, what was measured, and how it was done. For most of the methods presented examples of findings or tables are provided to illustrate the evaluation method used, and to furnish some data concerning the efficiency of the process or method under study. The review comprises analyzing error lists, reinterview studies, reviewing survey item definitions, simulations with artificially generated errors in Fellegi-Holt applications, measuring the impact of editing on reported data and performing analyses on raw data. Furthermore, a generally used technique of evaluating new editing methods is presented, along with examples of measuring the impact on estimates.

Keywords: Fellegi-Holt method; error analysis; measuring of impact of editing.

I. INTRODUCTION

The two studies carried out by Australian Bureau of Statistics (ABS) are briefly described together with a third study in Linacre and Trewin [9] as an introduction to a feasibility study of optimal allocation of resources in survey taking. The aim of the studies was to find more efficient editing approaches in the Agricultural Census and the 1985/1986 Retail Census.

1.1 1985-86 Retail Census of the Australian Bureau of Statistics

The main objective of the edit evaluation study was to analyze the frequency and effect of the manual data review to identify areas of potential rationalizing through improved questionnaire design and/or more effective processing designs. A sample of 336 short and 346 long forms of single-unit enterprises was selected for the planned full analysis. The most important components are listed below.

1.1.1 Data Content Queries

A remarkable finding was that the query rate for short forms was as unexpectedly high as 20% and close to the query rate for the much more complicated long forms (24%). Closer investigation of the queried short forms showed that very often suspected errors were insignificant, non-existent or could have been solved by referring to the values of other reported items.

1.1.2 Manual Intervention Before Data Entry

There were three types of manual pre-data entry tasks:

- coding, conversion of types, changes of name and address of the entity and so on, due to the design of the questionnaire or the processing system; (form/system changes);
- preparing the document for a fast, heads down data entry (rounding, deletion of decimals, insertion of missing totals); (tidy up changes);
- detecting and handling of high impact fatal errors causing problems for further data processing; (data changes).

The incidence of the three pre-data entry tasks is shown in table 1 along with the total number of data items changed (column 4):

Table 1

	Form/ system change s %	Tidy up changes %	Data changes %	Data items changed
Short forms (345)	65.5	56.8	43.2	444
Long forms (400)	55.7	56.0	74.2	1382
Total (745)	60.3	56.4	59.9	1826

The high frequencies of form/system and tidy-up changes for short forms were caused by low quality in

the register data, and the fact that clerks did much more than was necessary when tidying up data. Furthermore, it was found that respondents tend to omit totals and when they do provide them, tend to make errors in the summations. For the editing it was assumed that the components were correct and so, as a result of the evaluation study, it was suggested to let the system calculate the totals.

The number of substantial data item changes was higher than expected. The reason was that some clerks over-edited in the sense that they were making trivial changes. The number of ”clean” records was also calculated and found to be only 4.9% for short and 2.5% for long forms. Together, these two facts clearly indicate that the clerical intervention should be reduced.

1.1.3 Incidence of Computer Editing

After data entry all records were computer edited. The frequencies of fatal and query edits were calculated for short and long forms respectively. The main reasons for fatal edits are shown in the table in the preceding section. A table of the actions taken as a result of fatal and query edits was created and revealed that for query edits, 77% of the records remained unchanged. The conclusion from this evaluation study is that for ”similar surveys greater care has to be dedicated to the design of query edits to make them less wasteful and more effective”.

In general, the hit-rate, i.e. the percentage of flags resulting in changes to data, is proved to be a good indicator of the effectiveness of computer edits.

1.1.4 Impact on Key Data Items

Finally, the effect of the editing on values of five key data items was studied by an expert subject-matter statistician. The analysis resulted in a deeper knowledge of types of important respondent errors and of editor errors. The most common editor error was fixing totals to equal the sum of components to avoid flags from computer edits. It was called creative editing, which means that editors manipulate data in order to obtain the records to pass the computer edits without specifically being instructed to do so. Thus, editing may hide serious data collection problems, giving a false impression of the reporting capacity among the respondents.

1.2 1983/1984 Agricultural Census for Victoria, Australia

Editing in the ABS Agricultural Census accounts for about 23 per cent of the total budget of the program. It is also believed to have a substantial effect on the quality of data. Therefore, it was decided to carry out an editing study to investigate possible resource savings through reduced editing and/or increased use of automatic imputations, and to assess the effect of data editing in order to improve the efficiency of the editing process.

1.2.1 The Editing Process

The replies are streamed into ”big” and ”small” units. Big units are handled by experienced editors and extensively checked. Small units are checked by less experienced staff for legibility, use of correct type, and consistency. The consistency checks are comprehensive and can be somewhat complex.

After the manual pre-editing, data are entered heads down, and are then computer edited using predominantly consistency edits. Tolerance limits are based on historical experience, but can be changed according to available resources. Three types of failures are printed out for clerical handling: noticeable automatic adjustments; fatal errors; and so-called query edits, which mean that no change is necessary to clear the item for further processing. Changes to data are indicated on the error message, keyed in and submitted for computer editing. In the set of computer edits, all the edits used in the clerical pre-edit stage are included, thus causing considerably more work.

1.2.2 The Evaluation Study

For the study a SAS data base was created by entering original, pre-edited and computer edited values on a number of selected items from a systematic sample of 1000 units from the ongoing processing of the 1983/84 Agricultural Census in Victoria. Changes to data made in the pre-editing stage were assigned a code indicating the type of edit that caused the change. The following items were analysed:

A. Impact of Editing Changes on Totals

When all data of the sample were loaded into the SAS data base the study was conducted by tabulating the data and analysing the output. Tables of frequencies and magnitudes of editing changes and the impact on totals (aggregates) by item, by edit and by size of unit were produced both for the clerical pre-editing and the computer editing process. It should be noted that in this study it was not possible to obtain data on the hit-rates of the

edits.

B. Indicators of Distortion Effects

The issue was raised as to whether the applied consistency checks could distort underlying relationships between variables in the reported data. Correlations, considered a simple indicator of any distortion effects, were calculated for all pairs of items before and after clerical editing and again after computer editing. In the tables produced, no evidence of any significant distortion of relationships caused by editing was found.

C. Data on the Editing Process

The findings of all the analyses formed the basis for proposals of short- and long-term changes of the data collection and the editing strategy. The need for a sophisticated management information system was advocated; claimed to be the backbone for quality assurance in the processing system. The authors proposed that original value, edited value, type of error and data on failed checks should be written in a separate file for analysis of editing changes. Change analysis should provide the following types of information:

i) frequency and type of edit flags by item, by domain of study;

ii) magnitude of change by item, by domain of study;

iii) extent of redundancy in consistency checks;

iv) performance indicators by operator.

2. REINTERVIEW STUDIES

The only way to obtain reliable knowledge on how an editing process affects the quality of the estimates is to conduct reinterview studies with the respondents. Doing so through personal interviews is expensive, complicated and it is hard to obtain an acceptable response rate, especially in business surveys. Thus the number of reinterview studies presented in available papers is low. The only two available studies are described below.

2.1 1977 U.S. Economic censuses

Corby [2] presents the methodology and the results of a personal visit reinterview study aimed at measuring the accuracy of the published totals of employment, payroll and receipts for the 1977 censuses of retail trade, wholesale trade, selected services and manufacturers.

The questionnaires were edited for obvious errors, coded and keyed, and then carefully edited for coding and keying errors, for outliers and for suspicious answers. The reinterview data for each item were then constructed by changing the original data for each error made on a component included in the questionnaire. But this correction was only made provided the interviewer had classified the amount as a published value or reliable estimate. The records were then matched to census files to obtain the data values used for tabulation of the census publications. The final data records then contained original, reinterview and tabulation data, as well the component data used for determining the reinterview data.

Estimates and variances of the following ratios were calculated :

- the Reinterview total to the Tabulated total;
- the Reinterview total to the Reported total; and
- the Tabulated total to the Reported total.

Combinations of these estimates show how editing affects the quality of the data collection. It was found that the census editing changes the reported data:

- significantly too far but in the right direction for the items: retail sales, wholesale sales, services employment, and services annual payroll;
- in the right direction but not enough for the items: wholesale annual payroll, services first quarter payroll, manufactures annual payroll, manufactures value of shipments, and manufactures employment; and
- in the wrong direction for the items: retail employment, retail first quarter payroll, retail annual payroll, wholesale employment, wholesale first quarter payroll, and services receipts.

2.2 Content Evaluation of the 1982 Economic Censuses: Petroleum Distribution

In the evaluation studies of the 1982 Economic Censuses the same methodology as described above (Corby 1984), was applied. Estimates of totals for an item were computed from the sample of respondents to

the census using the originally reported value (reported data), the value used for the publications (tabulated data) and the reinterview value (accurate data), and then ratio estimates of pairs of these estimates were calculated. The three ratios:

- accurate data to tabulated data,
- tabulated data to reported data, and
- accurate data to reported data

can be looked at together to visualize what is going on in the processing. In this evaluation combinations of ratios indicated the following for sales and operating expenses, where strict inequalities indicate significant differences.

Types of errors were examined in order to obtain suggestions for possible improvements in the questionnaire design. It was found that errors made on individual components of sales and operating expenses are of little importance and cancel each other out. The largest types of errors originated from the reporting behaviour of using estimates instead of book values or from not reporting sales or operating expenses at all.

Tables were produced for each census item and for the components of the item showing:

- Eligible Number - applicable component where the respondent could make a mistake;
- Number of Errors - the component erroneously included or excluded;
- Number of Reliable Corrections - published figure or reliable estimate provided;
- Total change - sum of the published or reliable corrections provided.

The difference between the column of Reliable Corrections and Number of Errors shows the number of enterprises which could not provide a correction, or which provided changes assessed as unreliable by the interviewer.

2.3 Concluding Remarks

Evaluating editing processes by the methodology used in the referenced studies is highly recommended. It gives data on:

- the impact of editing on raw data and how quality is affected;
- possible error causes; and
- present differences in survey item definitions to definitions used in firms* book keeping systems.

3. REVIEWING SURVEY ITEM DEFINITIONS

Reviewing the way in which firms apply the survey item definitions can be recommended for evaluating the response quality and the effectiveness of the editing process in repetitive enterprise surveys. Results from such an evaluation process can be used to improve the data collection and to target the edits on important error types. Embedded in the survey vehicle it may replace parts of the edits or serve as an alternative to editing in the quality assurance process of the survey.

This conclusion was drawn from Werking et al. [13], a paper describing a Response Analysis Survey (RAS) conducted with the respondents to the U.S. Current Employment Survey (CES) and designed to review the survey item definitions against the firm*s book-keeping definitions. The RAS survey gives indications of the quality of the estimates and is aimed at improving the response quality in repetitive establishment surveys. Furthermore, it illustrates that editing cannot detect all types of errors, and that certain types of errors, when undetected, may seriously affect the quality of the estimates. In this case a number of enterprises erroneously include or exclude certain item components and do so consistently over time making it impossible to be detected by traditional edits.

It should be noted that traditional edit checks in repetitive surveys generally can only detect randomly appearing errors. The reason is that they are designed to discern those observations which have changed considerably since the previous report or are markedly different from most of the other units' aggregates for which the same edits are applied. Thus, over time, consistently erroneously reported figures or errors of the same type committed by a number of respondents cannot be detected by such range edits. That is why gathering intelligence concerning the definitions underlying reported data is so important for the quality assurance of a survey.

3.1 The Design Response Analysis Survey

RAS data are collected by a CATI-system and a fully automated touch-tone self-reporting system. For each principal CES item a number of components to be

included and excluded are presented to the respondents, who must answer with a yes or a no whether they include or exclude the component when providing figures to CES. For each negative answer component the respondent is asked whether he can separate the component and include/exclude it correctly in the future.

3.2 Results

In the referenced RAS study, the effect of the RAS in the CES was measured in two ways. First, separate estimates were made of the main items for the RAS group and the entire CES sample before and after conducting the RAS. Secondly, indirect measures of bias were obtained by tabulating the types and number of adjustments agreed to for the RAS sample.

The most notable findings were:

- over half of the enterprises studied did not entirely apply the survey definitions, primarily due to the record keeping system at the enterprise;
- the main item: ”Production Worker Earnings,” was significantly affected by non-application of the survey item definition. Applying the survey definition, the estimate for RAS units became 10.7 (standard error 3.2) to be compared with 1.6 for the control group; and
- the units changed their reporting habits, and the adjustments (samples too small to produce significant results) seem to be large relative to economic changes in the population.

3.3 Concluding Remarks

A suitable means of reviewing the application of survey definitions as outlined above is to use Touch Tone or Voice Recognition systems, unless the number of items under study exceeds about ten. When there is a need for studying many more items, a time schedule for evaluating groups of items may be set up allowing all items of interest to be covered under a given time period. Of course, such questions can be added to the survey instrument where only a couple of survey items are to be investigated. Then the evaluation is embedded in the data collection, a technique that is especially practical when self-reporting data collection modes are used.

4. STUDIES ON FELLEGI-HOLT SYSTEMS

A number of generalized systems based on the principles of the methodology introduced in Fellegi and Holt [5] have been developed: CAN-EDIT, AERO, DIA, GEIS, DAISY to mention those that have been the subject of many discussions at the Work Sessions of Statistical Data Editing and their predecessors (see Volume 1 of this series). Two evaluations have been selected:

- a simulation study using error planting that can be considered as a methodology of evaluating applications on categorical data, and
- an evaluation of imputations of missing values, where ”true” data are collected by reinterviews.

4.1 Simulation with artificially generated errors

The Fellegi-Holt imputation methodology as applied to the processing of the Spanish Labour Force Survey (LFS) was evaluated by simulation studies using real LFS data into which artificial random errors were planted [6]. Estimates based on the whole body of edited and imputed data were compared to estimates based only on the records accepted by the whole set of edits, assuming the item values of the erroneous records to have the same distribution as those of the approved records. These two methods of treating erroneous data were evaluated against the original LFS data file after having been processed by the edit and imputation system DIA. The quality was measured by calculating the distance between the relative distribution of the values of each item of each method to the relative distribution of the target data file.

The Spanish LFS is a complex survey which collects qualitative data. In the editing process, the generalized software for qualitative data (DIA) developed by the National Statistical Institute of Spain is used. For the study they selected those variables which were entirely edited by DIA. An original file was created using raw quarterly LFS data and data processed with DIA. The resulting file (FILE.GOOD) was considered as the target file, that is, the file of correct values to which the files resulting from the experiments were compared. Random errors were planted into FILE.GOOD by selecting records and changing the code values for a determined percentage of each item. For a selected item, the code value to change was selected by equal probabilities, and was changed to another code value of the item, to an invalid value or to a blank. This file, FILE.ERROR, was edited by DIA, resulting in FILE.GOOD2, consisting of the accepted records, and FILE.ERROR2, consisting of the erroneous records according to the edits applied in DIA. FILE.GOOD2 was used to obtain the marginal distributions of each item considered as

the estimates, called WAL, based on error-free records, i.e. the estimates resulting from discarding erroneous records assuming the distributions of the discarded records are the same as those of the error-free records. FILE.ERROR2 was imputed using DIA, and the resulting file was appended to FILE.GOOD2, obtaining FILE.EDITED. It was used to obtain the RLI estimates, i.e., the estimates resulting from imputing erroneous records using DIA.

These two methods of compensating for erroneous records were evaluated against the correct values, represented by FILE.GOOD, by using the item indicator Φ, defined as the sum of the absolute differences of the relative frequencies between the WAL and RLI code estimates respectively and the correct relative frequency of each code value.

$$\mathbf{\Phi WAL} = \cdot \text{FILE.GOOD(i) - FILE.GOOD2 (i)}\cdot$$

$$\mathbf{\Phi RLI} = \cdot \text{FILE.GOOD(i) - FILE.EDITED (i)}\cdot$$

where FILE.xxxx(i) is the relative frequency in the indicated file of code ”i” of an item, with ”r” code values.

The indicators ΦWAL and ΦRLI were calculated for each item. The sum of the item indicators showed that there was an overall increase in quality by using DIA for imputations. Discarding erroneous records resulted in ΦWAL=252, while imputing failed records resulted in ΦRLI=86. However, there were different results for different items. The increase in quality was high for items controlling the questionnaire flow, but low or negative for items with no or low correlation (redundancy) to other items. The items determining the routing in the questionnaire (flow variables) have the highest redundancy to other items, because answers to a number of questions indicate the same code of the flow variable.

To study this phenomenon more closely, another experiment was designed. An artificial file was created by constructing records so that all possible combinations of the item codes were represented (the Cartesian product for the items selected for this experiment was in excess of 200 thousand records). This artificial file corresponds to the original LFS file of the first experiment. Then the study was conducted in exactly the same way, first without two of the edits controlling the routing, secondly with the complete set of the ordinary edits with varying error rates of the items. The quality was decreased by using DIA without the two skipped edits, while it was notably increased with the complete set of edits for the flow variables.

The main conclusion drawn from this evaluation study was that there is an increase in data quality when applying automatic imputation according to the Fellegi Holt methodology. This occurs only in cases of random errors for items that are related to more than one or two items (that is, for cases where sufficient redundancy exists between items) and provided that the edit takes appropriate care of these relations.

For further results, conclusions and details concerning the studies, see Garcia & Peirats [6].

For evaluation of applications of generalized editing and imputation systems based on a determined methodology such as the Fellegi-Holt Methodology, the described method is useful. When implemented into a survey, it may serve as an excellent evaluation vehicle to improve the edits used, as data can be collected on the effects from varying the edits, error rates and error patterns (e.g. specific systematic errors). For example, the study by Garcia and Peirats shows how important for the quality it is to design the set of edits in a Fellegi-Holt based system to catch the redundancy that exists between answers to items.

4.2 Concluding Remarks

Fellegi-Holt based editing systems have difficulties in handling those systematic errors that arise from misinterpretation of an underlying concept and result in wrong but consistent answers to two or more items. In these cases there is a high probability that a correctly reported item is selected for imputation, thus increasing the number of errors. This evaluation vehicle is a tool used to see how the actual editing system will handle possible errors of this type. Thus, it should be noted that the method of conducting the described studies is very useful, although the same editing process has been applied to construct the file of ”correct” values. To obtain absolute measures of quality it is necessary to have a file of ”true” values, which will be much more costly to achieve.

4.3 Comparison with ”true” values collected by reinterviews

Werner [12] describes a study of the quality of the Fellegi-Holt imputation of missing values carried out by the United Kingdom Office of Population Censuses and Surveys (OPCS) in connection with the 1974 Test Census of Population.

An application of the Fellegi-Holt method is used to detect and eliminate any inconsistencies between the variables referring to a given person or household. The aim of the survey was to match imputed values with the true values. In order to achieve this aim, all missing values were first listed and subjected to clerical scrutiny to detect and eliminate cases of error occurring within the office, or for which follow-up was unnecessary. The remaining forms with missing items were returned to Census Office staff for interview with the households concerned, to discover the correct values of the missing items. These values were then matched with the values imputed by the computer system.

The study indicates that in a full census, where it is impossible to establish the values of missing items by follow-up, automatic imputation would be a more useful technique than proportional distribution of missing values. Werner concludes that the results from the 1974 Test Census have demonstrated that automatic editing is a practical and statistically sound method for use in census processing.

Table 2 shows how many of the computer imputations were correct for each item included in the survey. For comparison, the number of correct imputations which would have been expected by chance is shown.

Table 2

ITEM	TOTAL NUMBER OF IMPUTATIONS	NUMBER OF CORRECT IMPUTATIONS	NUMBER CORRECT EXPECTED BY CHANCE
HOUSEHOLD ITEMS			
Tenure	792	628	132
Number of rooms	284	160	47
Number of cars and vans available	723	393	181
Availability of fixed bath or shower	594	518	198
Availability of inside WC	819	631	273
All household items	3,212	2,330	831
PERSONAL ITEMS			
Sex	997	719	498
Marital condition	1,379	1,233	276
Age (5 year bands)	847	316	59
Economic position (status in labour market)	1,071	562	118
All personal items	4,294	2,830	951
ALL ITEMS	7,506	5,160	1,782

5. MEASURING THE IMPACT OF EDITING ON ORIGINAL (RAW) DATA

A major concern in the evaluation of the use of the SPEER edit and imputation system in the editing of the 1982 Economic Censuses, Greenberg and Petkunas [7], was whether the SPEER system in practice had an optimal strategy for determining which establishment records should be resolved by automatic procedures and which should be subject to individual review requiring special handling. Within this purpose the impact of the editing procedures on the respondents* original reported values was studied for the basic items of the Retail, Wholesale and Service Censuses. Initially they observed that the majority of editing changes were small in comparison with a small number of big changes, such as those due to reporting in units rather than thousands. For this purpose they developed a method which Statistics Sweden and Statistics Canada, among others, found to be a useful and easily implemented way of obtaining indications, whether the edits might be improved to focus only on big errors or

be more efficiently targeted on special error types for the survey under study.

5.1 Changes Relative to Total Change

The basic idea of the ”Greenberg-Petkunas” method is to create a change file for each record and item, containing the difference between the original reported value (usually denoted the *raw* value) and the final value (denoted the *edited* or *tabulated* value), to list all cases ordered by descending absolute difference, and to print out the cumulative percentage of cases and the cumulative percentage of change in graphs and tables. On the horizontal axis of each graph there is the cumulative percentage of cases in the change file, and on the vertical axis there is the cumulative percentage of change. The technique is carefully described as ”measure 1” in Lindell (see Chapter 1).

Greenberg and Petkunas found that approximately 5 % of the cases contributed to over 90 % of the total change and that the graphs of economic items were almost identical. ”Many of these large change cases were due to reporting in units rather than thousands; many seemed to be keying errors; and almost all were reviewed by a clerk/analyst,” they say in their report.

5.2 Changes Relative to Estimate

In a similar way the changes with sign can be related to the final estimates, making it possible to study in detail how the final estimate is approached by the cumulative sum of the editing changes taken in descending order according to the absolute size of the change. This is described as ”measure 2” in Lindell (see Chapter 1). Wahlström [11] used this method in her study on the impact of editing in the Annual Survey of Financial Accounts in Sweden. She shows that if only two percent of the largest changes had been made in the editing of the "Value Added" item, then the estimate would have been 98 percent of the published estimated total.

5.3 Changes relative to total change and estimate including missing values excluding sort errors

Boucher [1] uses both the described editing measures in the studies undertaken at Statistics Canada on the impact of micro-editing in the Annual Survey of Manufactures (ASM). The study was undertaken because most ASM records fail various edits and require extensive cleaning, and the impact of such processes had until then never been evaluated in Statistics Canada. The main object of that study was therefore to quantify the impact.

Boucher adjusted the change file for the bias introduced by a relatively high incidence of reporting in dollars when in fact responses are requested in thousands of dollars. Another notable difference to the studies mentioned above is that the records where item-nonresponse occurred were not excluded from the raw and the final data files. Graphs were produced with the two measures in the same diagram.

5.4 ”Largest changes needed” tables

In the Boucher (1991) study the editors and the subject-matter staff were interested in the graphs described in the previous section, but wanted the message more concretely presented. Their concern was how to translate the diagrams to formulate an operational goal of their editing task, since how much editing is needed to obtain the quality within a certain percentage of the ”correct” value. To them the correct value is the same as the final value. Therefore the concept of the largest changes needed was introduced. Tables were designed to represent the level of changes required (percentage and count of changes) to bring the estimate within a predetermined percentage of the actual ”final” data, had the changes been made to the corresponding largest contributors.

Table 3

Industry	No. of cases with changes	+/- 0.25% of final		+/- 0.50% of final		+/- 1.0% of final	
		per cent	number	per cent	number	per cent	number
Machinery	5	80.0	4	80.0	4	60.0	3
Sawmills	54	44.4	24	37.0	20	24.1	13
Petroleum	6	33.3	2	16.7	1	16.7	1

Table 3 is a reprint of the table presented in Boucher [1] and shows the largest changes needed as a measure of editing contributions for shipments.

6. PERFORMING ANALYSIS ON RAW DATA

To investigate whether editing has any impact on quality relative to the intended use of the survey, the planned analysis can be performed on raw data and edited data. When the analyses yield the same results or the decisions will be the same irrespective of which data have been used, then it can be stated that the editing was unnecessary. However, it should be noted that the method can be used to find out to what extent data should be edited.. The method was used in the evaluation of the machine editing of the World Fertility Survey, Pullum et al. [10].

6.1 The World Fertility Survey Evaluation

The World Fertility Survey conducted 42 large surveys in developing countries, using complex schedules, and did as much of the data processing as possible within the countries. The aims concerning editing were that WFS should be a hallmark of professional survey research and that WFS should serve as a vehicle to introduce modern editing to statistical offices in developing countries. Experts on editing were involved in developing the program for every country. The checking and updating were wholly computerised. All error correction was made by clerks, who examined the original questionnaire alongside the error printout and wrote out a correction statement.

Six countries were selected for the study, simply based on the availability of early raw data files. It is claimed in the report that "if these countries are unrepresentative with respect to data quality, it is probably because their data tend to be poorer than average". That is, the effects of editing may tend to be somewhat exaggerated with that choice of countries.

The study consisted of comparing the dirty (unedited) and clean (edited) pairs of files in diagnostic marginal distributions, two-way tables, fertility rates and multivariate analyses. The study is limited to the machine editing in WFS. This editing was preceded by two kinds of edits which were entirely manual, namely editing at the data collection phase (field edit) and input editing before the data entry (coding and checking of answers) when no re-contacts were possible (office edits). Structural editing was similarly excluded from the study. Thus the dirty file consisted of data which had been edited by field and office edits only. The clean files were the final versions of processed files where all cases with structural errors had been removed.

The main findings were:

- the rather elaborate logit regression and multiple regression analyses differed surprisingly little between the dirty and clean files;

- the multivariate analyses were relatively insensitive to the editing. Specifically, the changes in inferences about the magnitude of effects and their statistical significance were almost always less than the differences that would exist between two independent, clean samples;

- the cost of the machine editing is crudely estimated to be an average delay of approximately one year: compared to the benefits, such delays are excessive.

The study led to the following overall statement:

Consistency is desirable more because of later data processing convenience than because of its value for analyses, but it should be achieved quickly and practically never by referring back to the physical questionnaires.

7. EVALUATING NEW METHODS

Generally, new or alternative editing methods or procedures are tested by applying the method to raw (original reported) data and measuring the impact of those changes made in the current editing process, those which are identified and those which are not identified. The editing changes made in the ordinary data processing are considered as ”true” errors and the editing with the tested procedure is simulated by giving flagged items the tabulated value. Thus the test can never yield higher quality, but can show that the loss in quality is negligible or insignificant. The aim of the test is to find out whether the new method is more efficient to flag data than the current method, as measured by the impact of changes on the estimates.

7.1 Calculating Pseudo-bias

Latouche & Berthelot [8] use this technique when conducting studies on score functions for limiting the manual review of flagged records to the most important ones. The data collection and data capture process was simulated using data from the 1987 Canadian Annual Retail Trade Survey (CARTS). The score functions were evaluated against the final 1987 data. After a record has been flagged as suspicious, a manual review is possible. A manual review was simulated by replacing all the 1987 reported data of a questionnaire flagged for manual review with the corresponding 1987 tabulated data. For any given number of records flagged (review rate), an estimate $Y_{i,87}$ of the total for the full sample was computed using 1987 reported values for non-flagged units, and 1987 released values for flagged units. The absolute relative discrepancy between $Y_{i,87}$ and the total $Y'_{i,87}$ that was released in 1987 constituted the pseudo-bias:

$$\text{absolute pseudo-bias} = 100 * (*Y_{i,87} - Y'_{i,87}*)/(Y'_{i,87})$$

Different review rates, 0, 17, 34, 50 and 100 per cent, using the same score function, were applied to the sample.

They found that for all review rates the pseudo-bias was small for the frequently reported variables and that there was a significant decrease in the standard error in going from a 17% to a 34% review. But little is gained in going from 34% to 100%. Thus, as a result of the study, a 34% review rate was recommended. Table 4 (taken from Latouche & Berthelot [8]) shows the overall pseudo-bias obtained with a 34% review rate and the percentage of time the variable is reported.

Table 4

VARIABLES	% PSEUDO-BIAS	% TIME REPORTED
Net sales	-0.63	100
Gross commission	13.91	5
Receipts from repairs	-1.43	46
Receipts from rentals	-0.08	10
Receipts from food services	2.68	2
Other operating revenue	-13.20	15
Total net sales and receipts	0.18	100
Non operating revenue	-1.46	22
Opening inventory	-0.87	100
Closing inventory	-0.34	100
Purchases	-2.23	99
Salaries	-0.92	99

7.2 Producing Tables of the Number of Domains of Study by Size of Pseudo-bias

In the studies of the aggregate method (Volume 1, section C of this series) the pseudo-bias is calculated for each published domain of study. To see whether there is a number of estimates with an unacceptable pseudo-bias tables were produced showing the number of aggregates for each item by the absolute size of the pseudo-bias (see the table below). Domains of study showing high pseudo-bias were identified to compare the bias with the standard error of the estimate to see whether measures had to be taken to modify the editing method.

Table 5 shows the number of aggregates by the total relative difference in per cent of the estimates in a study of the aggregate method on data from the Survey on Employment and Wages in Swedish industry. The figures within parentheses show the outcome of another study on the same data.

Table 5

DIFFERENCE	WORKERS/HOURS		PAY-ROLL	WAGES/ HOUR

statistics. The effect of an edit is visualised by the vertical change of the curve. Big changes imply a steep drop or rise, while small changes do not change the vertical position of the curve. Every change can be compared to the confidence interval on the vertical axes.

7.3 Using Graphs and Confidence Intervals

Van de Pol (see Chapter 1) presents studies on selective editing in the Netherlands Annual Construction Survey (ACS). A score variable is developed to prioritise the edits. To apply selective editing to survey data, a critical value of the score variable has to be determined; only records above this critical value are followed up. Here graphs were used to determine critical values. Comparisons between edited and unedited data were made by plotting the estimates of the mean for decreasing values of the score variable. The vertical size scale reflects the confidence interval around the mean for the ACS

REFERENCES

[1] Boucher, L. Micro-editing for the Annual Survey of Manufactures. What is the Value-added?, *Proceedings of the 1991 Annual Research Conference,* U.S. Department of Commerce, Bureau of the Census, March 17 - 20, 1991, pp.765-781.

[2] Corby, C. Content Evaluation of the 1977 Economic Censuses. *SRD Research Report No: CENSUS/SRD/ RR-84/29*, U.S. Bureau of the Census, Statistical Research Division, Washington DC: U.S. Department of Commerce, October 1984.

[3] Corby, C. 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, 1987, pp. 27-60.

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

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

[6] Garcia, E. and Peirats, V. Evaluation of Data Editing Procedures: Results of a Simulation Approach. *Statistical Data Editing, Volume No. 1, Methods and Techniques*, Conference of European Statisticians, Statistical Standards and Studies, No. 44, 1994, pp. 52-68.

[7] Greenberg, B., and Petkunas, T. An Evaluation of Edit and Imputation Procedures Used in the 1982 Economic Censuses in Business Division, *1982 Economic Censuses and Census of Governments, Evaluation Studies,* Washington DC: U. S. Department of Commerce, 1986, pp. 85-98.

[8] Latouche, M., and Berthelot, J.-M. Use of a

Score Function to Prioritize and Limit Recontacts in Editing Business Surveys, *Journal of Official Statistics,* Vol. 8, No. 3, 1992, pp. 389-400.

[9] Linacre, S.J., and Trewin, D.J. Evaluation of Errors and Appropriate Resource Allocation in Economic Collections, *Proceedings of the US Bureau of the Census Fifth Annual Research Conference*, U. S. Department of Commerce, Bureau of the Census, March 19-22, 1989, pp. 197-209.

[10] Pullum, T.W., Harpham, T. and Ozsever, N. The Machine Editing of Large Sample Surveys: The Experience of the World Fertility Survey, *International Statistical Review*, Vol. 54, No. 3, December 1986, pp. 311-326.

[11] Wahlström, C. The Effects of Editing - A Study on the Annual Survey of Financial Accounts in Sweden, unpublished report, Stockholm, Sweden: Statistics Sweden, F-Metod No 27, 1990-02-26 (in Swedish), 1990.

[12] Werner, B. The development of automatic editing for the next Census of Population, *Statistical News No 37*, UK Central Statistical Office, 1977, pp. 3710-3715.

[13] Werking, G., Tupek, A., Clayton R. CATI and Touch-tone Self-Response Applications for Establishment Surveys, *Journal of Official Statistics,* Vol. 4, No. 4, 1988, pp. 349-362.

EVALUATING DATA EDITING PROCESS USING SURVEY DATA AND REGISTER DATA

by Marius Ejby Poulsen, Statistics Denmark

ABSTRACT

The system of socio-demographic statistics in Statistics Denmark is mainly based on administrative data, but some surveys are also carried out. One of the main strengths of the system is the possibility to integrate or match data from the two sources. This is used in almost all the statistics production in Statistics Denmark. In this paper, two alternative applications will be presented.

The first application concerns a micro-based comparison of survey data and register data within the same area. The purpose of this presentation is to show how this kind of comparison can be used to highlight areas where the editing processes can be improved. The system of labour market statistics, which is based on both survey data and register data, is used as an example.

The second application concerns a formalised data quality and editing project, where a continuous industry survey is used to correct individual data in the Central Business Register. One of the interesting aspects of this survey is the future plans, where results from one survey are used in organising the sample plan for the following survey.

Keywords: survey data; register data; quality assessment; micro data; error correction; Industry Survey.

1. INTRODUCTION

Statistics Denmark is well known for its strategy of using administrative data in the production of socio-demographic statistics. One of the crucial problems of this strategy is that Statistics Denmark has to use the administrative data provided and has very little influence on the collection of these data. Regarding data quality, it is argued by some statisticians that registers cannot provide adequate precision of data and that data are not reliable. This is because statisticians have no control over the content of registers, in contrast to their own surveys.

In Statistics Denmark the administrative data covers most of the statistical areas, but a few surveys are also carried out. The Danish system can in general be illustrated as follows:

Figure 1. The Danish system of socio-demographic statistics

Source: Eurostat and Statistics Denmark (1995).

The information sources are data from the administrative registers and from a few surveys. These raw data are subsequently more or less "edited". The edited data from the administrative registers are included in either a statistical register or in a classification module and the edited data from the surveys are included solely in a statistical register. Editing in this connection includes every process from the moment when the data are received to when they are either included in a statistical register, a classification module or published as tables or model data. Model data are normally used in connection with research projects.

2. COMBINING SURVEY DATA AND REGISTER DATA

There are numerous theoretical possibilities for combining survey data and register data, but of course - in practice - the legislation puts some restraints on the practical possibilities. The Danish system of statistics is built on the idea of using the different unambiguous identification numbers in the administrative systems, enabling linkages of statistical information and thereby enriching the statistical output.

The case where register data and survey data within the same area are available, represents a straightforward opportunity to assess both data quality and the methodology behind the production of statistics. In addition, when the data from the two sources are unambiguously identifiable, the assessment can be done on a micro level, evaluating individual data.

The purpose of this paper, is to give examples of the combined use of survey data and register data, with special emphasis on assessment of quality and consequences on editing. Two empirical studies will be presented as examples.

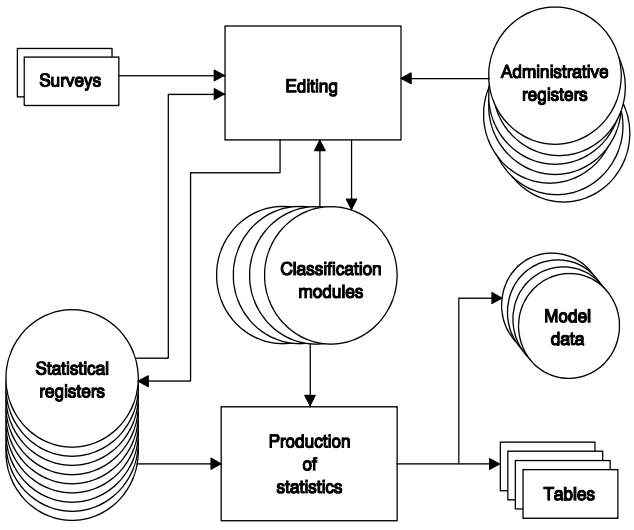
The first study is about micro-matching of survey data and register data in the labour market statistics, with the aim of detecting discrepancies in similar statistics from the two sources. Examples of how this kind of study can be used for evaluation of the collection and coding procedures will be given.

The second study is a presentation of the so-called Industry Survey. The purpose of this survey is to detect errors in the Central Business Register on the basis of questionnaires addressed to the single units in the register - a sample of all Danish establishments - and to correct these errors. The Industry Survey has been carried out continuously for some years, but this year the survey plan has been changed. The new Industry Survey is methodologically even more interesting from an editing point of view.

3. EDITING IN THE LABOUR MARKET STATISTICS

Labour market statistics are an example of a statistical area where both register data and survey data are used in production. The Labour Force Survey (LFS) represents survey data and on the register side several statistical registers are available, e.g. the Register of Employment Statistics, which forms the basis of the register-based labour force statistics (RLS).

Both LFS and RLS are based on individuals, and each of these are identified by an unambiguous number called the *personal code*. Therefore it is not only possible to compare the statistical output from the two systems but also individual data, i.e. to compare the data in the records for each individual included in both systems. This kind of linkage is only allowed for statistical analysis. In this connection register data are normally linked with survey data when drawing samples for the surveys and when enumerating the survey results. In the future, however, the combined use of the two types of data will probably be more and



more frequent in the production of statistics.

The methods of data editing in the two systems, LFS and RLS, and the amount of (and details in the) data available are essentially different, but in some areas both systems aim to measure the same statistics. Therefore, it is important to assess whether they actually do so and, if they do not, to explain why. In short, to assess the quality of the systems on the basis

of the statistics they produce, and to explain the reasons for possible discrepancies.

The presentation in this section will not go into details about all editing procedures that are carried out in the two systems. First, the systems will be presented in general. Second, as an example of differences regarding editing, the collection and coding of data on line of industry will be used. The two systems that will be presented are the CATI-based Labour Force Survey from 1992 (LFS92) and the Register-based Labour Force Statistics from 1991 (RLS91).

3.1 Register-based labour market statistics - RLS91

RLS91 is a good example of a typical system of register-based statistics. The main source of information are data from different administrative registers. These data are sometimes used directly, but frequently a lot of different editing is undertaken before the statistics are published. When the editing is finished, the data are included in a statistical register. Some other statistical registers (the statistical registers of education, population, unemployment, etc.) are also used to create the Register of Employment Statistics.

The purpose of RLS is to give a description of the labour market situation in the last week of November each year. RLS is based solely on the primary activity for each individual in the reference week (in other words, it is based on persons not on jobs). All statistics are based on the total labour force and examples of published statistics are the distribution of the total labour force by employment status, labour market status, etc., the employed population by region, educational level, etc., changes in distribution, commuting statistics and finally also several combinations of the examples mentioned.

One main characteristic of RLS is that it is a total account, where every individual is identified through the personal code. Another essential characteristic, common to most of the register-based statistical systems in Statistics Denmark, is the system itself, i.e. the route from administrative data to statistical data, which includes, among other things:

- error correction;
- imputation of missing values and how to treat the problem of insufficient information;
- delimitation of initial subgroups, e.g. regarding labour market status;
- setting up rules of priorities between different subgroups and data;
- setting up rules for choosing the most essential economic activity for each individual in the reference period;
- the use of auxiliary information from other administrative registers or statistical registers.

3.2 The Labour Force Survey - LFS

The typical purpose of the surveys is to provide extra information about social conditions which are not (directly) covered by data in the statistical registers, and the LFS is no exception. The LFS is voluntary and based on a mixture of CATI and questionnaires. The respondents not reached by phone receive a questionnaire.

BLAISE is used to collect and register the responses. Afterwards, all textual answers are coded automatically or manually. For the latter, auxiliary information on branch, education, etc. is used.

3.3 Conclusion

The methods and principles used in the data collection are essentially different in the two systems, mainly because of the fact that the information sources are different. As a result, it would be surprising if the resulting statistics were consistent, i.e. every individual classified equally in the two systems according to industry, sector, etc.

As will be shown in section 4, there are some discrepancies at micro-level when it comes to data on industry. However, the resulting distributions are not alarmingly different. It is very difficult to point out the main reasons for the discrepancies, but the description above gives an indication of possible "error sources". The question is how much of the discrepancy is due to actual errors (registration of answers, coding etc.) and how much is due to differences in the methods used.

4. ASSESSMENT OF DATA QUALITY IN THE LABOUR MARKET STATISTICS - STUDY 1

To assess the differences regarding editing, the collection and coding of data on industry is used as an example. As mentioned earlier, the existence of the personal code makes it possible to analyse all kinds of data and to compare similar statistics from two different sources. The presentation below will focus on the discrepancies on an individual basis, and not whether the distribution of the population is different in the two systems, according to any variable or

characteristic, although these two angles are interrelated.

4.1 Comparing data on industry

In this first study data on industry coding from RLS91 and LFS92 are compared. First of all the inconsistency rate between the data in the two systems is estimated. The discrepancies can then be analysed/distributed according to different characteristics, the main target being to highlight areas where the number of discrepancies are above average.

One of the crucial problems in the comparison is the existence of a difference in the reference period. To solve this problem, that is to try to identify a group of individuals assumed to be consistent with respect to the register data and the survey data, individuals who reply to specific questions as shown below have been picked out.

Question:	**Answer:**
What was your main activity in the reference week?	I was employed
Were you employed for a limited period of time?	No
Is your place of work cited in Denmark?	Yes
When did you start working at your current place of work?	Before 1991

Consequently, the population is characterized by being employed in Denmark for a non-limited period and presumably by having a job in the reference week of RLS91. The number of individuals fulfilling these conditions was 8087 (the stable population). The industry code at one-digit level is then compared between this group of individuals and the group of people classified as employed salary earners in RLS91. The result is that 90.3 per cent of individuals have the same code in the two systems.

Another problem when comparing the data in question is that neither the LFS-data nor the RLS-data are without errors. Consequently, there is no error free checklist. Nevertheless, let us assume that the industry code is correct in RLS, and look more closely at the 9.7 per cent discrepancy in order to highlight on possible error-sources in the survey.

First of all, Table 1, when looking at the horizontal dimension, illustrates that some industries are better coded than others. Considering the numerical size of the different industries, *Manufacturing* and *Public and other services* particularly have an inconsistency rate below average.
Above average we find the industries *Wholesale and retail trade* and *Financial institutions and insurance*.

The first error-source investigated is the interviewers carrying out the interviews. In Table 2 the rate of consistent and inconsistent pairs of data from Table 1 are distributed according to the person who carried out the interview.

Table 1. Comparison of industry code, for the stable population, number of individuals [1]

	Industry code in LFS92										
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Total**	**pct.**
Industry code in RLS91											
1	77	0	4	0	4	1	3	1	20	110	30,0
2	0	4	1	0	0	1	0	0	0	6	33,3
3	3	3	1511	0	27	39	9	23	17	1632	7,4
4	0	0	1	73	4	4	0	0	2	84	13,1
5	5	0	18	0	307	10	2	16	7	365	15,9

[1] Content of industry codes on 1-digit level:

1. Agriculture, forestry and logging, fishing; 2. Extraction of oil, coal and gas and mining; 3. Manufacturing industry; 4. Electricity, gas and water; 5. Construction; 6. Wholesale and retail trade; 7. Transport, storage and communication; 8. Financial inst. and insurance; 9. Public and other services.

6	6	0	137	1	14	780	17	26	63	1044	25,3
7	0	1	10	0	6	10	570	0	13	610	6,6
8	3	0	31	0	7	16	5	706	29	798	11,5
9	13	0	33	7	13	26	17	47	3188	3344	4,7
Total	107	8	1746	82	382	887	623	819	3339	7993	9,7
pct.	28.0	50.0	13.5	11.0	19.6	12.1	8.5	13.8	4.5	9.7	

On the whole, interviewers who do more interviews have better results. This is confirmed when looking at the total inconsistency rate (pct.) for the interviewers who did less than the average number of interviews (about 223 interviews = 7148/32). For this group the inconsistency rate is 10.7 pct, and for the group of interviewers who did more than the average number of interviews the rate is 9.6 pct.

Table 2. The stable population, distributed according to the interviewer who carried out the telephone interview (only respondents interviewed by phone are included)

Inter-viewer	Consis-tency	Inconsis-tency	Total	Pct.
1	1	0	1	0.0
2	5	0	5	0.0
3	11	0	11	0.0
4	27	2	29	6.9
5	54	3	57	5.3
6	115	22	137	16.1
7	133	11	144	7.6
8	127	18	145	12.4
9	132	22	154	14.3
10	163	33	176	7.4
11	177	20	197	10.2
12	175	26	201	12.9
13	197	20	217	9.2
14	190	35	225	15.6
15	189	43	232	18.5
16	219	15	234	6.4
17	226	19	245	7.8
18	231	20	251	8.0
19	227	30	257	11.7
20	242	19	261	7.3
21	250	21	271	7.7
22	253	24	277	8.7
23	256	28	284	9.9
24	259	37	296	12.5
25	286	22	308	7.1
26	291	23	314	7.3
27	296	28	324	8.6
28	297	28	325	8.6
29	337	31	368	8.4
30	338	33	371	8.9
31	340	36	376	9.6
32	402	53	455	11.6
Total	6446	702	7148	9.8

Another possibility is to distribute according to the type of interview. As mentioned earlier two types of interviews are performed; telephone-based and questionnaire-based. The result appears in Table 3.

Table 3. The stable population, distributed according to the type of interview

	Consis-tency	Inconsis-tency	Total	Pct.
Telephone Interview	6446	702	7148	9.8
Questionnaire	770	75	845	8.9
Total	7216	777	7993	9.7

The consistency rate is better for the questionnaire-based interviews. This is somewhat surprising given the possibilities of obtaining more specific information in a telephone interview. Furthermore the interviewers carrying out the telephone interviews are trained to be specific when an answer is ambiguous. When handling a questionnaire the coder has to use the textual answers as they are.

A third possibility is to distribute according to type of respondent.

Table 4. The stable population, distributed according to the individual interviewed

	Consis-tency	Inconsis-tency	Total	Pct.
The respondent him/herself	6473	681	7154	9.5
The respondent's spouse	681	83	764	10.9

Others	62	13	75	17.3
Total	7216	777	7993	9.7

The picture is clear. The information on which industry the respondent is working in is best given by the respondent him/herself.

An interesting result appeared when cross-tabulating the data on consistency in industry code with some of the variables in the data set. It appears that there is a significant difference in the consistency rate between males and females, see Table 5.

Table 5. The stable population, distributed according to the respondents' sex

	Consis-tency	Inconsis-tency	Total	Pct.
Male respondent	3566	522	4088	12,8
Female respondent	3650	255	3905	6,5
Total	7216	777	7993	9.7

A plausible explanation to this result is that the females are concentrated in fewer industries. About 60 pct. of the females are employed in the sector *Public and other services*, which is normally fairly easy to define. For the male population the problem of defining where their job should be assigned, e.g. in production, sale or construction occurs to a larger extent, thereby presumably negatively affecting the consistency rate.

As mentioned earlier, it was assumed the industry code was correct in RLS was only an assumption. In RLS data the industry code is actually created in many different ways. The basic key when classifying an individual according to industry is the workplace. In most cases the workplace is linked to the individual automatically by matching the data from the information sheets mentioned earlier (personal code and identification number for the establishment/workplace in which he or she works) with the Central Business Register. In some areas, however, it is difficult to define the workplace appropriately and therefore the workplace is either imputed (the same workplace as last year, the workplace closest to the individual's domicile), set as a fictitious workplace (cleaning personal, etc.), set as the individual's domicile (insurance agents, etc.) or set as the employee's municipality (roadworkers, etc.). In RLS data on the method used for the placement of workplace for every individual is included, and in Table 6 the stable population is distributed according to whether the placement of workplace in RLS91 is done automatically or manually and, as in the tables above, whether the data on industry code is consistent or not in the two systems.

There is no significant difference in inconsistency rates between automatic and manual placement of workplace in the RLS. It seems that the editing undertaken regarding the manual placement provides equally good results as the automatic placement, the latter being the workplaces where manual editing is assumed to be unnecessary.

Table 6. The stable population, distributed according to method used in the placement of workplace

	Consis-tency	Inconsis-tency	Total	Pct.
Automatic placement of workplace	6778	731	7509	9.7
Manual placement of workplace	438	46	484	9.5
Total	7216	777	7993	9.7

4.2 The results seen from an editing point of view

The type of analysis presented above is not part of the daily work in the office of labour market statistics, but was conducted exclusively for this paper. It is obvious that the ability to integrate statistical information in Statistics Denmark enables micro-based assessment of data quality and editing processes, but unfortunately the available resources limit the practical possibilities. Nevertheless, analyses of the kind presented above (and others) could be an integrated part of the methodological work when further developing the different statistical systems. At the least, the analyses can be used to highlight areas where, for example, editing processes should be given special attention.

Specific comments on editing aspects of the different results (tables) are given below. The fact that some industries do better than others regarding

consistency between industry codes from the two systems is not taken into consideration, although in some areas this is an underlying explanation. The fact that errors occur in both the survey data and the register data - there is no checking list - is not taken into account either (for further discussion of evaluating this aspect, see section 5).

Consistency rate between industries (Table 1)

Table 1 illustrates the basic result of the comparison, the magnitude and relative consistency rates for each industry. This table can obviously be used to highlight those industries where collection and coding of data seems to be more difficult. In other words, special attention should be paid to the relatively large industries where the discrepancy is over the average. Assessment of possible improvements for these industries in, for example, the collection and coding procedures could be a possible line to follow.

Interviewers (Table 2)

This table could be viewed as a ranking of the interviewer's performance. On the other hand, considering the fact that both LFS and RLS are influenced by errors, it would be controversial to use this kind of information directly, e.g. to confront interviewer who has a rate of inconsistent responses over the average. Nevertheless, the information from the table is useful. As an example, it might be the case that there is a connection between the consistency rate and the interviewer's experience (amount of training). A straightforward suggestion would be to let fewer interviewers carry out the interviews and/or to put more resources into pre-training of the interviewers.

Type of the interview (Table 3)

The questionnaire-based interviews turned out to be relatively more consistent than the telephone-based ones. This result should not, however, lead to a suggestion that questionnaires be used to a higher degree, considering the cost-efficiency of the telephone-based interview.

Looking at the results from tables 2 and 3 together, the most rational suggestion would be to improve the editing of the telephone-based interviews.

Who is the respondent? (Table 4)

In view of the poor consistency rates for interviews not carried out with main respondents, the importance of trying to get in touch with the main respondent should be emphasised. On the other hand, if it was compulsory for every interview to be carried out with the main respondent, this would probably increase the rate of non-response.

Editing the register statistics (Table 6)

The results show that in those cases where manual placement of workplaces was necessary, the consistency rate was as good as in those cases where automatic placement was done. This was a positive result, which implies that there is little improvement needed in this special editing procedure.

In the study presented here, emphasis has been on the collection and coding of data on industry. It would be preferable if more data on the editing processes were produced. In LFS data the useful information for this purpose could be: whether automatic or manual coding has been used, who carried out the manual coding and finally the number of records coded manually due to spelling errors or alternative spelling[2]. In RLS more information/data is needed on the formalized but also the ad-hoc based editing processes which have an impact on data values.

5. QUALITY CONTROL AND ERROR CORRECTION OF REGISTER DATA USING SURVEY DATA - THE INDUSTRY SURVEY

In section 3 and 4 an example of the possibilities for combining register data and survey data in order to assess data quality on a micro level was presented. In this section, the background and some results from an alternative and more formalized application, combining the two types of data, will be presented. The purpose of this presentation is to show how editing of register data can be organized based on the questionnaires, confronting a special sample of the units in the register, in order to capture more cost-efficiently any erroneously registered unit.

The application in question is *the Industry Survey*. This survey has for some years been used as a method to correct errors of registered variables in the Central Business Register. In February 1996 the survey was described in an internal report which, among other things, contained a general description of the survey and presented some results from the 1995 survey, and some plans for future industry surveys.

---

[2] If it is a significant amount of records, the work on manual coding could be reduced.

The Central Business Register information is widely used, inside as well as outside Statistics Denmark. Updating the variables describing the enterprises is an important element of the work on the Central Business Register. This is mainly done by updates from administrative registers.

However, the experience has shown that changes which may have a considerable bearing on the quality of the production of statistics are not being recorded or updated in the administrative registers, which are the primary source of the Central Business Register. The industry of an enterprise is an example of such a variable, and is also one of the most widely used. The quality of all statistics broken down by industry is, to a certain degree, dependent on the reliability of this variable.

In the presented study an attempt is made to readdress the general data quality problems caused by these shortcomings in the Central Business Register by conducting questionnaire-based surveys mailed to enterprises. In addition to a targeted error correction, quality measurement is becoming an important element of industry surveys.

5.1 Results regarding data on industry code

For the analysis of the results regarding industrial changes presented below, the starting point is exclusively mailed questionnaires which have been completed; the data are for 2,436 workplaces.

A total of 380 workplaces, i.e. 15.6% had an incorrect industry code. Table 7 shows the share of units which fell into different industry groups compared with their previous coding. It should be noted that the industrial changes recorded can be partly attributed to imprecise or incorrect registrations and partly to actual changes in the industry.

It appears from Table 7 that units in *Manufacturing; Wholesale/retail trade, hotels, restaurants* and *Financial intermediation, business activities* have an above average rate for industrial changes. At the opposite end are units within the industries *Electricity, gas and water supply; Construction* and *Transport, storage and communication*. In numerical terms, it is particularly enterprises within *Financial intermediation, business activities* and *wholesale/retail trade, hotels and restaurants* which predominate. Moreover, it can be noted that there is a high rate of industrial changes for enterprises under *Activity not stated* - a natural consequence of the survey.

Looking at this result together with the results from the study presented in section 3, at least two things are notable. The industry *Manufacturing* has a high share of industrial change but the inconsistency rate in Table 1 is below average. The industries *wholesale/retail trade, hotels and restaurants* and *Financial intermediation, business activities* also have a high share of industrial change, corresponding to the result from Table 1. Without going into further details, but bearing in mind that other aspects occur in a comparison of the two studies, it is necessary to take several things into consideration when drawing conclusions, especially from the study presented in section 3.

Table 7. Industrial changes by standard industrial grouping

	Sample survey	Industrial change		
			Share	Distribu.
			per cent	
1. Agriculture, fishing and quarrying	551	87	15.8	22.9
2. Manufacturing	174	39	22.4	10.3
3. Electricity, gas and water supply	28	0	0	0
4. Construction	142	9	6.3	2.4
5. Wholesale/retail trade, hotels, restaurants	567	95	16.8	25.0
6. Transport, storage and communication	91	3	3.3	0.8
7. Financial intermediation, business activities	525	95	18.1	25.0
8. Public and personal services	307	33	10.7	8.7

9. Activity not stated	51	19	37.2	5.0
Total	2 436	380	15.6	100.1

5.2 Future plans

To make maximum utilisation of resources in the future, it has been decided to send out the industry survey based on fixed criteria reflecting the experience and knowledge about erroneously registered units. It has been planned that the "capture capability" of the criteria, understood as the share of returned questionnaires forming the basis of changes in the register information, must be continuously assessed. The weighting of the fixed criteria is immediately effected according to their share in the total respondent population. The 1995 sample survey, which is excluded from these criteria, will form the basis of an assessment of the "capture capability" of each individual criterion, and it will also provide the statistical/analytical basis for setting up any new criteria in the future.

The work on the industry survey will be a permanent part of the statistical division's tasks after 1996. The Industry Survey will be carried out on a quarterly basis, thus continuous assessments of any possible changes in the stated criteria will be made and the total "capture share" should increase over the year.

The idea of continuously assessing the “capture share” for each of the stated criteria and consequently letting this assessment influence the future sampling frames is a new one. The cost-efficiency in this way of conducting the survey is obvious. One aim of the survey is to correct erroneous data in the Central Business Register. Using information from earlier surveys is obviously a more efficient way of reaching this aim rather than drawing a random sample of establishments each period.

The result from the Industry Survey underlines the statement from the first study that register data can not be used as a checking list for the survey data. Nevertheless, the two studies can be used together to improve the editing procedures. The results in the first study could be used in establishing new criteria for the sampling frame. This study showed industries with high rates of inconsistencies, and these industries could be included as a criteria, e.g. the criteria being “industries in which the rate of consistency, when comparing LFS and RLS, are under x per cent”. Behind this suggestion lies an assumption that the responses in LFS are correct.

Finally, an improvement of the evaluation in study 1 would be to match the register data on industry code for the stable population in study 1 with data from the Industry Survey. The dataset in both studies contain unambiguous identification numbers for the establishments, and the results from study 1 could consequently be further analysed, for example, on whether the discrepancies are due to survey errors or errors in the Central Business Register. In this way the background information used for suggestions and/or decisions on whether, and where, to use additional resources to improve the editing process will be of a higher quality.

6. CONCLUDING REMARKS

It is obvious that if the staff resources available in the statistical offices were not so limited, more evaluation studies would always be preferable. As in other statistical offices, Statistics Denmark has limited resources and the timeliness of the statistics is a high priority. This does not, however, imply that the efforts made regarding validation, evaluation and quality assessment have low priorities. The marginal costs of using more resources should be taken into consideration in this connection. Work on this topic has been going on in Statistics Denmark for some time, and will also be part of the future work plan.

In the first study the straightforward possibilities of comparing individual data between sources, in order to highlight error sources and potential areas where editing processes could be improved, was presented. In this connection it is worth mentioning a new project in the area of labour market statistics, which started in the autumn 1995, called Labour Market Accounts. This is built on the ideas from CBS Holland regarding their development of a Labour Accounting System. The objective is to develop a system which produces consistent and coherent measures of the labour market situation. Some of the sub projects will most certainly include detailed assessments of the methodologies used in the different areas of the production of labour market statistics, e.g. by combined use of survey data and register data. In this way the efficiency of resources used for validation, quality assessment and evaluation of data editing processes will increase.

REFERENCES

[1] Eurostat and Statistics Denmark. Statistics on

persons in Denmark - a register based statistical system, statistical document, 1995.

[2] Poulsen, M. E. The LFS and the register based labour force statistics - a quality assessment, contributed paper to the SMPQ-conference, Bristol, April 1-4, 1995.

[3] Statistics Denmark. Statistiske efterretninger - Arbejdsmarked (1992:20)*, Arbejdsstyrkeundersøgelsen 1991,* 1992.

[4] Statistics Denmark. Statistiske efterretninger - Arbejdsmarked (1993:17)*, Registerbaseret arbejdsstyrkestatistik ultimo November 1991,* 1993.

[5] Statistics Denmark. Gross national product; comparison of labour force surveys with data from administrative registers with special emphasis on the full coverage of economic activity, Unpublished report, 1994.

[6] Statistics Denmark. Industry Survey, Sample Survey 1996, Working Paper no. 1, 1996.

Chapter 5

IMPACT OF NEW TECHNOLOGY ON DATA EDITING

FOREWORD

By Ron James, Electronic Data Systems Ltd, United Kingdom

This chapter focuses on the impact of applying new technology to data editing processes. In this interesting age of rapid technological advancement and increasing importance of statistical information, statistical agencies are looking at ways of improving the speed and relevance of statistical production and, above all, of reducing their costs.

What is the true impact of the new technology? Is efficiency improved without negative impact on the quality of estimate? Are time series discontinued due to major changes in methodology brought about by the new technology? Are cost reductions realized in the first year or in several years? Is the initial reduction cancelled by hidden technology maintenance costs in later years? Are purely processing cost reductions properly set against possible increased support costs? The hope of this chapter is to begin a dialogue within statistical organizations on these issues.

In principle, the following technological improvements can be observed in the practices of statistical agencies:

(i) introducing an electronic questionnaire via Computer-Assisted Personal Interviewing (CAPI); Computer-Assisted Telephone Interviewing (CATI); and Computer-Assisted Self Interviewing (CASI);

(ii) new software capabilities enabling applications to be created more easily and quickly - GV-S, BLAISE, GEIS, etc.

(iii) implementation of electronic data interchange in the data collection and data editing process;

(iv) using optical reading;

(v) new ways of exploiting administrative data, e.g. relational databases (RDBMS);

(vi) cheaper processing power PC/LAN, CLIENT/SERVER.

In the case where an **electronic questionnaire** can be completed by the respondent directly and interactively, a potential error or problematic entry could be checked directly at the time of entry. This has a potentially large impact on the quality of the data and subsequent editing processes. Orientation towards an electronic questionnaire is reported in almost all contributions to this chapter. Provided the work has been put into the **CAPI/CASI/CATI** software in the first place, there could be considerable savings in the human and computing effort of editing.

The production of **generalized software packages** for part or all of the editing processes and more, such as BLAISE (developed by Statistics Netherlands) and GEIS (developed by Statistics Canada), almost removes the need for local software development for the processing of surveys. It does allow the statistician better control over the content and timing of surveys, removing the often lengthy, expensive and inflexible software development previously necessary. There is no excuse for tailor-made software, specific to a survey and unique in most ways.

A significant influence on the whole data collection and data editing process proved to be **electronic data interchange** (EDI). The extensive use of this technology in statistical agencies permits closer communication with data suppliers and increases their involvment in the data editing process. Implementation of EDI in many statistical offices is still, however, in its infancy. Valuable experiences in this respect could be learned from the contributions presented by Keller (Statistics Netherlands) and Clayton et al. (U.S. Bureau of Labour). Also in this chapter, the contribution from the Statistical Office of Slovenia considers EDI to be an integral part of its data collection policy.

We have been promised the paperless office for some years and many using **scanning** and workflow

techniques have almost succeeded in achieving this goal. **OCR** or **ICR** systems that attempt to read and interpret both typed and hand-written characters respectively from paper forms are now being used to great effect. The primary object is to remove the need for manual keying of data. Although some of the papers that follow [see Dumicic, Statistics Croatia, and Blom et al., Statistics Sweden] describe considerable success with this technology, keying or manual intervention is never completely avoided. Furthermore, new problems arising from systematic errors can appear, having an impact on validation and possibly even autocorrection or imputation. The use of such techniques requires the entire process of data capture, validation and vetting to be redefined, including different ways of designing questionnaires in both physical and logical ways.

The advent of the **relational database** for use in administrative systems facilitated by both software and hardware performance improvements should enable the extraction of data in much greater volume and with almost infinite flexibility without impacting operational development and use. The consequences for editing may be considerable in that the solution to such problems may lie in the examination of trend information and the use of random techniques to ensure that the resultant statistical collection of microdata have been populated so that they have the correct statistical properties.

Most of the potential effects of adopting cheaper processors and presumably different operating systems are both enabling, in terms of the products available, and problematic, in that any organization will want a controlled and standard environment with a minimal set of skills required for its support and use. For example, the **Client/Server paradigm** has much to commend it in both cost and flexibility but it is intrinsically more complex than a single layer Mainframe-terminal population as a computing environment. The obvious advantage of cheaper processing power may be lost because of the lack of a corresponding reorganization of the whole data production process. To introduce such new rules could very often be a rather complicated interdisciplinary process.

It is therefore important that national statistical offices, when considering their strategy on statistical data editing, take into careful consideration both the possibilities offered to them by recent technological development and the necessary organizational measures enabling the efficient introduction of new technology into statistical production.

EVALUATING DATA QUALITY WITH COMPUTER ASSISTED PERSONAL INTERVIEWING

by Tom Pordugal and Roberta Pense, National Agricultural Statistics Service, USA

ABSTRACT

The Technology Research Section of the National Agricultural Statistics Service (NASS) has been investigating two methods of improving the data quality and timeliness of survey data. These involve the integrated use of Interactive Editing (IE) and Computer Assisted Personal Interviewing (CAPI), including the use of telecommunications software to transmit data. The 1994 June Area Survey research project was conducted in Indiana and Pennsylvania as a pilot test to evaluate procedures. The 1995 test involved Indiana. As a secondary consequence of this investigation, the effect on staff functions has also been evaluated.

Keywords: batch edit; CAPI; CATI; check-in; data capture; data collection; error correction; error suppression; editing instrument; hand edit; IE; routing instructions.

1. INTRODUCTION

NASS conducts an omnibus survey of farmers and ranchers during the first two weeks of June. The questionnaire design and survey specification functions

for this survey are centralized in Headquarters. The actual administration of the data collection however, is distributed to each of NASS's 44 field offices. The surveys use samples from both list area frames, with face-to-face interviewing for the area frame samples, or segments. Area frame segments are identifiable land units, most commonly about 640 acres in size. Information is obtained concerning crops and livestock associated with these land areas. About 1500 interviewers throughout the country obtain information concerning approximately 50,000 farms and ranches through these area frame samples, while screening out another 60,000 people. Official estimates based on these survey data are released approximately two weeks after data collection. Thus timeliness and optimal resource allocation are of major concern to NASS for this survey.

NASS's traditional method of collecting and editing data for face-to-face interviews is via pencil and paper with batch processing on a timeshare mainframe computer. After the interviewers collect the data on paper, their supervisors review the questionnaires. Then the questionnaires are sent into the State Statistical Office. With batch processing, errors discovered after data entry involve re-handling the questionnaire, transcribing the corrections, additional data entry and another batch edit run. Therefore NASS staff spend a great deal of time reviewing the questionnaires prior to data entry to avoid corrections after data entry. Typically there are at least three reviews of the questionnaire by different staff members in the office prior to data entry.

Interactive editing and CAPI are inter-related, but can be separated. CAPI refers to the data collection method. NASS interviewers use sub-notebook computers loaded with computer programs which contain the questions, routes, and edit checks. Because there are edit checks during the interview, interactive editing does occur during the interview. However, after the data are sent to the State Office, more stringent edits can also be applied interactively using the same software. This phase of editing in the State Office is NASS's definition of interactive editing.

Interactive editing eliminates many of the pre-data-entry review functions and streamlines the post-data-entry editing of paper questionnaires. Only two staff members reviewed the questionnaires prior to data entry, compared to the traditional three reviewers. Each of these reviewers performed only a fraction of the functions they originally performed.

Computer assisted personal interviewing moves most of the edits to the data collection phase of the survey, thereby further reducing post-data collection review. All edits related to routing through the questionnaire are handled during data collection. Edits displaying invalid answers (out-of-range) are triggered during data collection. Data capture also takes place during the interview avoiding the need for key entry in the office. NASS is in the stages of CAPI development, with applications limited to a small scale (only involving one State). As implementation is expanded, more issues will arise, such as the need for an infrastructure to support hardware management, data collection "hotlines", etc. NASS's decentralized data collection structure poses challenges in maintaining consistency in high quality CAPI survey administration, CAPI training, and staffing. The impact of these issues are not addressed in small scale tests.

On the other hand, many issues that must be dealt with in a development phase, such as the high cost of initial training, procurement, and system development, will be stabilized once procedures are fully implemented.

2. EFFECT ON HEADQUARTERS STAFFING

With NASS's traditional paper and pencil method of data collection, one person in Headquarters designs and prepares the paper questionnaires, while another person writes editing programs for the batch mainframe system. NASS has combined these two functions when designing CAPI instruments.

Blaise software was used to develop both the editing and data collection instruments. Statistics Netherlands developed this software to provide an integrated system for surveys. By programming one set of code, the program can be compiled into either CAPI, CATI or IE survey instrument. This reduces a programmer's time in writing and maintaining code, as well as insuring consistency and compatibility between data collection and editing. The staff involved in writing and maintaining the code estimate that they spend about 60 percent as much time preparing both the data collection and editing instruments simultaneously as they would if these instruments were in different media and therefore needed to be prepared separately. However, this magnitude of time savings will not be realized until there are no paper questionnaires being prepared for this survey. As long as both CAPI and paper questionnaires are used for the same survey, staff resources are being duplicated and

there is a small increase in overall workload.

NASS uses other techniques to improve efficiency in instrument preparation. To maximize both the programmer's time and consistency across surveys, the instrument is modular in nature. Thus, parts of the questionnaire can be re-used in other surveys. The "shell", or front and back questions which are administrative in nature, was developed by another unit in NASS. Although a few modifications were needed for this survey, the bulk of the work had already been done.

Two modules of the instrument, the grain stocks section and field crops table, incorporate a different set of questions (crops) for each of the 45 State Offices. To allow for generalization of the code so that it can be used for all States, the instrument uses external files with State specific information. Question text, unit of measurement, and edits are triggered according to the external file.

3. EFFECT ON DATA COLLECTION STAFF

The June Area Survey interviews are often conducted outside, in bright sunlight, so computer screen visibility is crucial. At times, the farmer is so busy, the interview is conducted in the farmer's field, beside the tractor, so computer speed and instrument maneuverability are imperative. Interviewers' must juggle the computer and an aerial photograph of the segment (about 2 feet by 2 feet in size). From the interviewer's perspective, this survey is NASS's "worst case" scenario for using CAPI as a data collection tool because of the stressful environment.

In 1994, seven interviewers in Indiana, two interviewers in Pennsylvania, and their supervisors were trained in CAPI. Interviewers were selected by their supervisor to participate in the project. In 1995, all 27 field interviewers and their 6 supervisors in Indiana were trained. The interviewers' computer backgrounds and aptitudes ranged widely. The typical NASS field interviewer is a woman about 45-65 years of age, with little computer experience. Because these interviewers had not used CAPI before and paper and pencil was their backup data collection method, training was intense and covered skills needs in both data collection methods. These included such CAPI skills as: understanding the screen layout, navigation within the questionnaire, and simple error corrections and suppressions; keyboard skills; telecommunications procedures; survey management protocol; and hardware familiarity. For paper and pencil data collection, NASS focuses on interviewing techniques, definitions, routing instructions, and program material. Therefore training costs increased significantly. From 1992 (all interviewers used paper and pencil) to 1995 (all interviewers used CAPI), interviewer costs for Indiana increased about 50 percent, with 38 percent attributed to CAPI training for first-time CAPI interviewers, and 6 percent attributed to inflation over time. It is hoped that the increased training costs are a one-time start-up cost for new CAPI interviewers.

The number of completed interviews are listed below. The interviews classified as non-agricultural, were basically screen-out interviews. The agricultural interviews are those where actual farm data were collected. In 1994, 23 percent of all agricultural interviews in Indiana were collected by CAPI.

Table 1: Number of interviews collected via CAPI

	1994			1995
Interviews	Indiana	Pennsylvania	Total	Indiana
	Number			
Non-Agri-cultural	619	195	814	2464
Agricultural	363	95	458	1510
Total	982	290	1272	3974

Feedback from the interviewers and their supervisors indicates that they were able to use the computer effectively but felt they needed more practice. In-class practice time was considered the most important aspect of the training, although a few interviewers enjoyed practicing in the privacy of their own home.

When asked on a scale of 1 (hated it) to 10 (loved it) how the interviewers liked CAPI, 75 percent "liked it" (scores in the 7-9 range). Two "loved it"; one "hated it"; and the rest were neutral (scores of 5 and 6). The aspects the interviewers liked best about CAPI (mentioned by at least 20 percent of the interviewers) were:

1) less paper to keep track of;

2) questions asked in the correct sequence;

3) the importance and challenge of a new experience; and

4) less questionnaire review by the interviewer after the interview.

The aspects they liked least were:

1) poor screen visibility in the sun/heat; and

2) interviews took longer.

Most interviewers thought that they could have completed the interview faster on paper. Part of the perceived problems with speed in CAPI was the questionnaire complexity. In 1994, as the interview progressed, things slowed down, especially if many rows of the large field table were filled in. Another "problem" with speed is that interviewers were required to fill in questions they may have left blank in a paper interview. This may have been a bigger factor in 1995, when weather forced farmers into a later planting schedule and the farmers had less time to spare.

Analysis of 1994 time data from Indiana indicate that the CAPI interviews were slightly faster than the paper interviews. Because of problems with the data, only 128 out of 363 CAPI interviews (35 percent), and 552 out of 1239 paper interviews (45 percent) were summarized. The CAPI interviews for agricultural tracts averaged 18 minutes while the paper interviews averaged 19 minutes. The median was 15 minutes for both data collection methods, with a range of about 3 to 55 minutes. CAPI time stamps were collected automatically by the survey instrument. The interviewers using paper were asked to record starting and ending times on the paper questionnaire, and most of them rounded to the nearest 5 minutes. Thus, the data seem to demonstrate that CAPI should not take any longer than the paper questionnaires.

Interviewer cost data for 1994 were also analysed. The data collection costs for interviewers using paper questionnaires averaged $36.60 per agricultural tract. The CAPI interviewers' costs averaged $33.78 per agricultural tract. Thus, CAPI interviewers' cost averaged $2.82 per agricultural tract less than paper interviewers. This difference may be attributed to less post-collection review by the interviewers and supervisors for CAPI data. However, the difference in costs may be due to other factors, such as number of interviews, geographic area covered, etc. This analysis also does not include training costs, nor does it include hardware costs.

Based on this limited data analysis, it appears that CAPI does not warrant hiring additional interviewers, nor does a different type of interviewer need to be hired. However, training is very important.

4. EFFECT ON OFFICE STAFF

The CAPI interviews were to be transmitted electronically to the State office on a daily basis, avoiding a large workload towards the end of the survey. In 1994, interviewers transmitted about 105 times during the 16 days of data collection, averaging slightly less than once a day. In 1995, about 40 percent of the interviewers said they transmitted at least every day; 25 percent transmitted every other day; and the rest transmitted as needed. The length of time needed to run programs to prepare the data for transmission, along with the need to recharge batteries and long workdays, was the most common reason for not transmitting every day. Because the transmitted interviews were not physically removed from the sub-notebook computers, interviewers could re-edit the completed questionnaire if new information was received. With paper questionnaires, once the paper questionnaire is sent to the office, the questionnaire is no longer available to the interviewers in the field. Interviewers therefore tend to keep the paper questionnaires in the field longer.

Thus, it was hoped that the use of telecommunications with CAPI for all samples in 1995 would distribute the office workload more evenly throughout the survey period. This was to be evaluated in 1995. Unfortunately software "bugs" surfaced during the survey itself. Procedures had to be changed in mid-survey. These changes required a significant amount of the office staff's time to communicate both problems and fixes to the appropriate staff. Therefore, they did not have time to perform their usual office functions until late in the survey period, negating the analysis.

In 1994, it was possible to compare the amount of time in the office spent processing CAPI collected data to the amount of time spent on paper collected data. In-office interactive editing of CAPI-collected data averaged 31 percent less time per agricultural interview to process than paper interviews. This savings is in addition to the time savings due to interactively editing data. Previous NASS research estimated that interactively editing paper questionnaires reduced staff time about 15 percent as compared to batch editing

paper questionnaires. The average office time spent per agricultural interview for CAPI was 19.3 minutes, while it averaged 28.1 minutes for paper questionnaires. The following office processing functions were eliminated by CAPI: check-in, hand edit (arithmetic, routes, etc.), and key-entry, totalling 12.3 minutes. The time spent reviewing the questionnaire for within questionnaire consistency checks was cut in half. However, time needed for file management increased significantly with CAPI.

There was concern that CAPI was decreasing or eliminating the functions performed by lower-paid employees, while increasing the higher-paid employees' time. However, when weighting the 1994 time savings by salary, there was still a 27 percent savings in processing CAPI interviews. More savings would have been realized if it were not necessary to re-suppress all soft error signals. Detailed cost data were not collected in 1995. However, the software problem which required a mid-survey change in procedures certainly affected the distribution of costs. Delegation of responsibilities to lower-paid employees could not occur until procedures stabilized, which was late in the survey period. Thus the workload was centralized among a few higher-paid staff in 1995. In addition, the survey management procedures did not function as planned, resulting in a labor-intensive follow-up.

5. OTHER COMMENTS

Several Headquarters units worked together to allow field interviewers to use their sub-notebook computers to conduct telephone interviews from their home for surveys which are traditionally collected by computer assisted telephone interviewing (CATI) from the office. The Blaise software features an integrated design which allows conversion from CAPI to CATI with a simple recompilation of the instrument. CATI data collection is cheaper than CAPI (no mileage cost), so data collection via CATI is preferred. However, if the office interviewers can not contact the respondent by CATI, the assignment is sent to a field enumerator to use CAPI. This is called Computer Assisted Telephone Home Interviewing (CATHI). The field interviewers collect the data in person in certain situations (no known telephone or previous refusal) and by telephone in other situations. This allows the field interviewers, who tend to be more knowledgeable about agriculture, to have heavier involvement in more surveys. It also allows higher utilization of the computers so that training and equipment costs are amortized over more interviews.

The role of supervisory interviewers in NASS is changing. With paper and pencil interviewing, supervisory interviewers have traditionally spent most of their time reviewing and evaluating their interviewers' work after it is completed. With CAPI, many of these functions are handled by the instrument during the interview. In fact, with interviewers transmitting the data directly to the office, the supervisor's post-interview review is eliminated. Thus, their focus is shifting more toward training interviewers and testing instruments prior to the survey. Because they have not had as much first-hand experience using CAPI to conduct interviews, the supervisory interviewers need as much time as possible to prepare for the training. During the survey, supervisors still need progress reports and objective measures by which to evaluate their interviewers and keep track of their progress during the survey. As stated earlier, the Survey Management procedures targeted for implementation in 1995 did not function as expected, and therefore many modifications are needed.

Table 2 summarizes some of the issues affecting the survey staff when comparing the paper and pencil data collection method to CAPI.

Table 2: Summary of Some Issues Affecting Staffing

Issue	Paper and Pencil	CAPI
1. Procurement	Printers, Envelopes	Computers, Software
2. HQ Development collection	Separate jobs for data collection questionnaire and computer edits	Integrated development of data and edit instrument
3. Enumerator Training	Proper routing, Required answers, Within-record consistency checks	How to use computer
4. Enumerator Responsibilities		
a. Survey Management	Keep track of papers	Keep computers/batteries functioning
b. Data Transmission	Mail	Via modem and telephones
c. Post-Interview Edits	Routing, Completeness of data legibility, Make notes	Make notes
5. Supervisory Responsibilities	Review after data collection	Training and testing
6. Office Staff		
a. Pre-Survey	Labelling questionnaires	Prepare files for computers
b. Survey	Check-in and Data Entry	Not needed
c. Questionnaire review	Several staff reviews	Not needed
d. Post-Data Capture Editing	Many errors to review	Very little editing other than handling refusals or inaccessibles
e. File Management	Minimal time required	Increased time requirements

TRENDS IN TECHNOLOGY FOR DATA COLLECTION AND EDITING AT STATISTICS SWEDEN

By Evert Blom, Statistcs Sweden

ABSTRACT

This paper presents the results of a survey on the current situation in data collection methods in Sweden based on a questionnaire sent out to survey managers. As more than half of the production costs of statistics are taken up by data collection, entry, editing and coding this is the key area for rationalization and automation with the use of modern EDP methods. Three different methods Computer-Assisted Interviewing (CATI/CAPI), Computer-Assisted Self Interviewing (CASI) and traditional Paper and Pencil Interviewing (PAPI) are analysed.

Keywords: computer-assisted interviewing; computer-assisted self interviewing; scanning.

1. INTRODUCTION

The use of modern data communication tools is one of the most promising areas to make the collection and editing part of statistics production more efficient. Attention should be drawn to cost reduction, timeliness and quality improvement due to simplification, fewer production steps and the possibility of direct feedback to the respondents during the collection process itself. Furthermore, the integration of questionnaire design, computer-assisted interviewing, data entry, analysis and summarization into the survey process improves

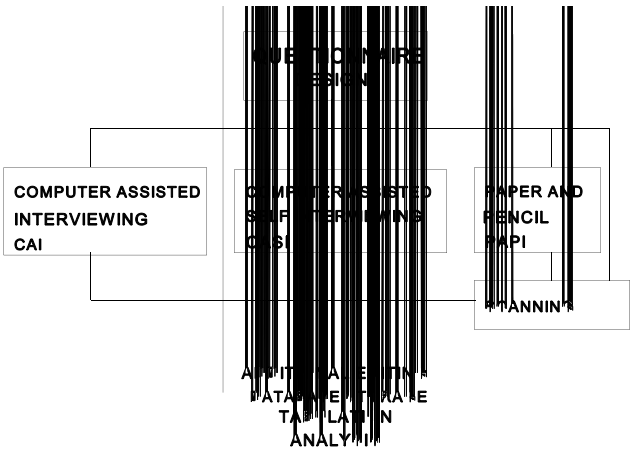
Survey Management, reducing duplication in data definitions.

Using Electronic Data Interchange (EDI) and diskettes, questionnaires can appear at the respondents own computer for further work. Computer-Assisted Self Interviewing (CASI) can provide a front-end support (instead of traditional back-end support) in the same way as computer-assisted telephone interviewing (CATI) and computer assisted personal interviewing (CAPI).

To realize such a vision is a challenge which calls for continuous development work on rationalization and automation of the data collection process. The questionnaire itself is constructed using an integrated software package. Not only the questions are formulated but also the answer categories, skip patterns, instructions, edit rules and failure messages. As an output from the questionnaire design the survey manager can choose the transformation into three different applications (see Figure 1):

- Computer-assisted interviewing (CAI) which includes CATI/CAPI mode.
- Electronic Form for Self Interviewing (CASI).
- Ordinary Paper Form for mail out/mail back.

Figure 1



Due to the overall survey design, and respondent's desire, it is possible to choose more than one output. In the case of paper and pencil output, a data entry program is automatically generated. To make scanning and optical character recognition (OCR) possible there will also be a feature for scanning/interpretation set up and the image of the questionnaire will be available during the edit process. The sketch below envisions the idea.

Furthermore EDI between different data systems could be integrated in the model, and complemented by an Electronic Form for data not found in the respondent's own application.

2. SURVEY ON CURRENTLY USED METHODS AND TECHNIQUES

To find out the current situation in data collection methods used, an in-house questionnaire was sent to survey managers (Statistics Sweden 1994). From a list of surveys conducting their own data collection, about 270 answered, giving a good overview and a basis for further follow-up. The main purpose was to assess the potential scanning volume, but some information of more common interest is listed here:

- The annual volume of questionnaire pages is about 5.7 million; on average 3.6 pages per questionnaire (the population census is not included);
- About half the surveys have less then 500 respondents and about 20% have more than 3000 respondents. Only four surveys have more than 50 000;
- Traditional mail out/mail back is most frequently used (53% of cases but often in combination with other collection methods). Interview surveys are usually conducted in CATI/CAPI mode;
- Data is spontaneously delivered by fax in 20% of the surveys. The amount is increasing;
- About 40 % of the surveys are provided with pre-printed information (not including labels with personal number, identity, etc.);
- Data entry is usually done using personal computer (80%) and supported by RODE/PC, Excel or Paradox, often as "heads up" applications;
- Half of the surveys are manually edited before data entry;
- Respondents are recontacted in about 60% of the surveys; to what extent is not clearly reported. Some output editing (macro editing) is carried out in about a third of the surveys;
- A bit less than 60% of the surveys report there is no coding at all. Automatic coding is done in 5% of the surveys.

3. TECHNOLOGICAL ENVIRONMENT

All employees at Statistics Sweden have their own personal computer connected to a local area network

(LAN). This includes the interviewers who are dispersed throughout the country. A high-speed bridge between the two locations (Stockholm and Örebro) ensures good performance in data communication. The mainframe can be seen as the most powerful LAN server. There is a trend towards downsizing production to personal computers, and a rapid change of platform to that of client/server is going on. The most production, however, still takes place in the traditional way, with personal computers used only as terminals, emulating IBM 3270-protocol.

4. COMPUTER-ASSISTED INTERVIEWING (CAI)

Computer-assisted interviewing, which means telephone interviewing (CATI) and personal interviewing (CAPI), is today standard in all professional interview surveys. Statistics Sweden has developed an in-house system called DATI, which is the Swedish abbreviation for CAI. It is used both by the telephone group as a central CATI application and by the field interviewers for decentralized CATI and CAPI. This mode of survey application produces "clean data" for downstream analysis and production, which means that comprehensive object edit routines are built in.

Current software standards for the mainframe and micro environment are:

Software:	**Mainframe:**	**Micro/PC/Network:**
Operating System:	MVS	IBM LAN Server and DOS/Windows for the workstations
Database management:	MIMER	PARADOX, ACCESS and Microsoft SQL Server
Production tools:	SAS Easytrieve TAB68 (in-house tabulation package)	SAS DATI (in-house for CATI and CAPI) RODE-PC for data entry and editing
Dissemination:	AXIS	PC-AXIS
Graphics:		Charisma
Word processing:		MS Word
Desktop:		PageMaker
Spreadsheet:		MS Excel (micro)
Presentation:		MS PowerPoint
Fax, Telex, Teletex:		SST TELEMATIC
E-mail:		MS Mail
Scanning/OCR		Eyes&Hands (ReadSoft)

A new CATI system named WinCATI is under development and will replace DATI and push the degree of computerization to almost 100%. The new system will contribute to enforced competition by the

quickest possible flow of interviews under strict control. Centrally the system will be used by some 80 interviewers with access to a common database. By letting the system automatically select the interview to be performed next, a greater flow of interviews is expected. It is based on Windows and Client/Server with SQL based data handling. Gösta Nilsson [6] outlines the ongoing development in this area.

With regard to realizing the vision of integrated software support for all types of questionnaire design, mentioned in section 2, the program IFK (Swedish abbreviation for Interactive Questionnaire Construction) is one of the candidates for general use. For example, IFK could be used to generate electronic forms (programs) to be run on the respondent's PC and to generate forms for traditional mail out/mail back. The latter will then be prepared for scanning as the form's definition and many edit rules are already done during the questionnaire design.

5. COMPUTER-ASSISTED SELF INTERVIEWING (CASI)

5.1 Electronic questionnaire

The electronic questionnaire created is sent to the respondents either by mailed floppy disks or by some other form of electronic transmission.

The respondents enter their data on a computer using interactive questionnaire-based data entry screens with built-in edits. They are aided by help screens, function keys, pop-up windows, menus, data edit failure messages and other instructional features. There are in principle two ways of presenting the form on the computer. It can appear in the same shape as an ordinary paper-based questionnaire (full screen form) or one question at a time is presented on the screen, similar to what usually happens in a CATI/CAPI application. A respondent's support for data entry, instructions and edit checks is integrated into the program.

The KYBOK Study:

As reported by Granquist [5], an experiment with electronic questionnaires was carried out for a sample of 110 parishes, among a total of 1200, in the 1992 KYBOK census. A commercial electronic mail system called MEMO was used for distribution and collection. MEMO was originally developed by Volvo, and many companies within and outside Sweden are connected to the system.

The questionnaire was constructed using a program called FORMAX, developed by Statistics Sweden. The program generates an electronic questionnaire, and shows the questionnaire as one or more screens on the respondent's monitor (each screen corresponding in principle to one page of the questionnaire). After answering the questions, it is possible for the respondent to edit the information and thus correct any error therein.

The results of the KYBOK study were rather poor. Only 33 parishes filled out the questionnaire. However, there were far fewer errors in these questionnaires than in the others, and above all we gained a great deal of experience and a deeper insight into problems connected with this mode of data collection. Problems occurred at the following points:

- delivery of individual forms to each parish;
- some respondents were not familiar with the technologies used;
- data editing needed to be better coordinated and adjusted to the new mode of data collection. Survey processes should not be transferred to a new collection mode without re-engineering them to the new mode;

The findings determined that this mode should include the following features:

- help-function keys that can provide definitions of concepts and terms;
- further and extensive information on error messages through help function;
- automatic conversion of reported data formats to survey item formats (e.g. data are reported in kilograms when reporting in pounds was required);
- data transfers (copying of data) to other items summarizing subtotals and so on should be executed by the program;
- the respondent should be allowed to comment on changes or provide explanations of why failed or suspicious data have not been changed;
- warnings like "did you include/exclude a certain component of the item...." should be included to help the respondents to acquire a deeper understanding of the concept and to give

information to the survey management on the quality of the reported data;

- original respondent values should be stored for audit trails.

5.2 Touchtone Data Entry (TDE)

TDE is an established technique used in many applications (e.g., banking and mail-order). The U.S. Bureau of Labor Statistics reports excellent respondent acceptance of TDE [8]. Questions are answered using the telephone keypad. One of the advantages compared to CATI is that the respondent provides survey answers at his/her own convenience and that instant corrections can be made by edit rules built into the conversation. Furthermore, the interview goes quickly and is cheaper to implement than CATI. On the other hand, this technology limits the complexity of both the questions and the answers. Response occurs at the respondent's initiative, and if motivation is lacking, conventional follow-up is necessary, which reduces the savings in time and costs that this mode otherwise provides. The telephone costs can be less than those incurred for regular postage, and data entry costs disappear.

An inventory at Statistics Sweden indicated that six or seven surveys could employ TDE techniques. This was considered sufficient for investments in the equipment and systems development. During the tests, a number of data, counters and time meters have been registered for use in the evaluation. The respondents have also been asked their opinions of the new data collection mode.

The TDESOS survey:

A test using Social Service Statistics was completed. In this survey, every municipality in Sweden (286) was asked to report the amount of social welfare it had paid out for each quarter. In this first attempt, for the 1993 fourth quarter collection, 46% answered in TDE mode. The rest used ordinary paper and pencil, but a follow-up study indicated that the TDE level could be increased to about 70%. However, the actual TDE level is not more than 50 - 60 %. The TDE system is adopted as a regular mode of data collection since the first quarter 1994.

The TDE trial showed the following positives:

- Keyed-in information is repeated by the computer, making it possible for the respondent to check whether he/she has committed a keying error;
- The respondent immediately receives comparison data calculated from earlier responses. In this way the respondent can determine how reasonable the newly reported data are;
- The respondent is told which of the municipality's data will be published;
- Reported data, approved by the respondent, can be put directly into the production process.

Furthermore, TDE offers some time savings since traditional data entry, checking and editing are eliminated. It offers more complete information about the production process itself because everything that happens is automatically logged, calculated and presented. Finally it creates potential for further rationalizations such as automatic faxing of questionnaires, reminders, etc..

TDE PPI Study:

The Producer Price Index (PPI) was used in an initial small acceptance test carried out in May 1993, followed by a three-month production test in October - December 1993. The results were very encouraging, but due to reorganization of the survey the TDE mode was not implemented until 1996. The ongoing implementation seems to be as promising as the earlier test indicated.

Some of the main findings from the study 1993 are listed below:

- 86% of the experiment group used TDE at least once and in most cases did so without difficulty;
- In debriefing sessions, respondents expressed a high degree of acceptance of TDE;
- Response rates increased in the experiment group, and exceeded the control group by about five percentage points. The final response rate for the TDE-mode was 75%;
- Production deadlines were kept;
- As an added advantage, TDE generates excellent data on the production process for monitoring purposes and successive quality improvement;
- The costs of using TDE were easily shown, but not the savings potential. To evaluate the savings potential, we need greater documentation of the current production process. We believe that we can reach the stated savings.

Further use of TDE:

A new potential survey for TDE is a survey on Domestic Trade, which involves about 5000 enterprises. Three different forms are used. Sales turnover figures are collected monthly from about 2000 enterprises and quarterly from the rest. Contrary to experiences from the other surveys mentioned, the first implementation results indicated a very low answering level (less than 10 % using TDE). We are actually examining this and planning for a redesign of the information package and further implementation actions.

Anyway, careful implementation plans must be made and follow-up actions corresponding to actual results must be prepared.

6. SCANNING

One way of improving the processing of paper-bound information is to use scanning and optical character recognition (OCR), which means that the forms are photographed in a scanner and the information is interpreted by a computer. Traditional data entry is thus rendered unnecessary and editing can be done during the scanning process. The document can be stored as an image accessible during the editing and correction process and the scanned image can also be used for long-term archive purposes. One reason that we are using this technique is the mandate that all specific data entry work (heads-down) shall be eliminated.

The information that requires interpretation is:

- Numerical information (figures 0-9), which can be handwritten or typed;
- Handwritten or typed alphabetical signs (regarding e.g. occupation);
- Bar-code and various types of marking (cross, line, etc.).

There are at present no systems that can interpret free handwriting letters (cursive handwriting). But controlled handwriting (block letters) can be interpreted. Free figure panels can be interpreted by the most highly developed systems (which is to say that no special boxes are required for the various characters).

Statistics Sweden has been using scanning for data capture since 1993. The overall goal was to obtain shorter production times and less expensive statistical products while maintaining data quality.

Experiences from tests and regular production have been reported by Blom [2]. A more extensive report is given by Blom and Friberg [3].

Based on the experiences from the testing and the production of nearly 500,000 scanned pages in different surveys, it was recommended to further extend the use of scanning and OCR. A purchasing procedure was started, aimed at placing a couple of locally oriented systems at the subject-matter departments. Suppliers were invited to tender, based on a comprehensive documentation, including requirements and evaluation criteria. The requirements consisted of 85 critical points, complemented by 56 questions regarding different points which must be answered by the suppliers. Quoted systems had to go through a practical test to verify their capacity and quality using different forms from the production line. Some examples of requirements from the tender document are mentioned below.

Purchasing is for what we call *a Basic System*. In *general* the system must be for use in the existing EDP environment. It must be a multi-user system with an interface for Windows and must be capable of handling multipage forms containing several sheets with double-sided print, namely sets of forms.

The system must be capable of recording forms in any format from A5 up to A3, of reading recto-verso copy, of handling divergences in the number of pages in a set of forms, of simultaneously recording different types of forms and automatically indexing; reading not less than 30 A4 sheets per minute in the production.

The system must be easy to handle when it comes to *defining forms* for recording, including an easy check function, without special EDP skills being needed. It must have functions which make changing the format of fields, placing them, etc. easy; and it must have a function that filters away irrelevant information such as separation and other special characters. Examples of such characters are / - + :- , .

The system must be capable of *interpreting* TIFF files, CCITT Grp 4; marks in boxes, with at least a 99% recognition rate on a character basis; numerical pre-printed characters (OCR-B), with at least a 95% recognition rate on a character basis; numerical machine-printed characters, with at least a 90% recognition rate on a character basis; numerical handwritten characters, with at least a 90% recognition rate on a character basis. The system must also be able to interpret document files received via fax modem.

The system must include *editing, correcting and coding* capabilities comparable to the data entry/correction system currently used at Statistics Sweden (RODE/PC). This means duplicate controls, validation of values and intervals of values, checking against external tables outside this system, controls between all fields in the form, making equations using fundamental rules of arithmetic, force controls with retained erroneous values, etc. The system must also cope with reject correcting and other editing/correcting with multiple simultaneous users, and must have functions for manual supplementary keying of fields which are not being interpreted. In the process of editing/correcting, the system must take no more than 1.5 seconds for changing image, when the system is operated in stand alone mode and utilizes DDE and DLL functions.

On the *output* side, the system must offer the possibility of transferring images to an electronic image database. Then the image is immediately accessible for further editing. For *workflow information and production statistics*, the system must include a process control and have statistical functions showing the recognition rates on a character level, field level and form level.

Finally the tender document should also include requirements concerning s*afety and protection against unauthorized access.*

Some results from the tests

The figures show the interpretation levels achieved in the test runs. Interpretation level is the share of characters that the OCR system manages to interpret automatically. That is, the number of interpreted characters divided by the total number of characters in the test material. The share is expressed as a percentage.

Survey	**System A**	**System B**
Balance statistics	97,3 %	98,3 %
Salaries and wages statistics	94,9 %	92,2 %
Answer Sweden statistics	not available	97,7 %

There is a risk that the OCR-system misinterprets a character and misjudges the confidence level to be OK. This leads to a substitute in the resulting data file. Substitutes occurring in the test runs were measured for the different surveys and systems. On character level, the substitution rate varied between 0,1 % and 2,8 %. On field level substitution rate varied between 0,4 % and 3,7 %.

The scanned values were finally compared to a carefully controlled key file leading to the results. Results indicated that only a few substitutes may change values significantly as a substituted digit in a high numeric value may have a big impact on the results. We conclude that users of OCR systems need to use different control systems to secure high quality in data capturing; e.g. check number controls for ID-numbers, comparisons to external tables with previous values which are expected not to deviate more than to a limit set by the user, etc.

We finally regard Scanning and OCR systems to be promising alternatives to traditional keyboard punching with a high potential of cost reduction for data entry while maintaining high data quality.

7. THE USE OF FAX

FAX equipment is becoming increasingly common, not just in companies. Today about 30,000 form pages are delivered to Statistics Sweden by fax from respondents and it is expected that the number will increase in the future. This development may also depend on the interpretation of characters; the received picture could go directly to an interpretation unit in the recipient's computer. The technology exists, but does not appear sufficiently developed as yet. Some studies are planned to be undertaken in this area.

Fax can also be used for reminders in a survey. It was found in the TDEPPI study that using fax had a big impact on the respondents' answering behaviour. 50% of the respondents answered the same day they received the fax reminder. This was very promising, but it turned out that the final response rate was not raised as a result of this new method.

REFERENCES

[1] Blom, E. Building Integrated Systems of CASIC Technologies at Statistics Sweden. *Proceedings of the 1994 Annual Research Conference*, U.S. Department of Commerce, Bureau of the Census, March 1994.

[2] Blom, E. Data Collection and Editing at Statistics Sweden, Working Paper No. 14. *Work Session on Statistical Data Editing*, Cork, Ireland, 17-20 October 1994.

[3] Blom, E. , Friberg, R. The Use of Scanning at Statistics Sweden. *Proceedings of the Conference on Survey Measurement and Process Quality*, Bristol, April 1-4, 1995.

[4] Ferguson, D. P. Review of Methods and Software Used in Data Editing, *ECE Work Session on Statistical Data Editing*, Working Paper No. 4, Stockholm, 11-15 October, 1993.

[5] Granquist, L. Report No. 7 on Data Editing Activities in Statistics Sweden, *ECE Work Session on Statistical Data Editing*, Cork,17-20 October, 1994.

[6] Nilsson, G. Computer-Assisted Telephone Interviewing, Statistics Sweden´s new system - WinCati, Statistics Sweden PM, 1996.

[7] Statistics Sweden. Produktionsenkät, Undersökning av 1992/93 års produkter med egen datainsamling, June, 1994, (in Swedish).

[8] Winter, D. L. S., Clayton, R. L. Speech Data Entry: Results of the First Test of Voice Recognition for Data Collection, *Paper presented at the Joint Statistical Meetings of the American Statistical Association*, Anaheim, CA, August,1990.

TECHNOLOGICAL INFRASTRUCTURE USED FOR DATA EDITING AT SLOVENIAN STATISTICS

By Milan Katic and Pavle Kozjek, Statistical Office, Slovenia

ABSTRACT

The Statistical Office of Slovenia is a centrally organized institution which collects data directly from reporting units. The majority of surveys have a complete data collection and thus there is a relatively high basic data input.

The data editing process is technologically based on minicomputers linked to a mainframe computer system installed in the Government Centre for Informatics. Different methods and techniques are used in the data editing process. On-line and batch processing as well as the combination of both methods of work are used. Special data entry package DCR-5000 is mainly used for key data entry on minicomputers. Optical Character Reading (OCR) technology and interactive entry are used for some surveys. For data checking and correction different software is used, mainly in-house developed software package for index correction (INDXPO, batch and online version) and Godar/Vega-STAT (GV-S) system for interactive data editing. In future, focus will be on the implementation of Electronic Data Interchange in the data collection process.

Keywords: Electronic questionnaire, “heads-down” entry, OCR entry, integration of different data editing phases.

1. INTRODUCTION

Data editing is still the most central and extensive operative task in the whole statistical production process. The level of complexity of the data editing process depends primarily on the two following basic factors:

- Technical and technological approaches to the data editing process in statistical production as well as their integration;

- Organizational issues, that is, contact with reporting units, decisions about sampling, the possibility to obtain data from other sources (e.g. administrative registers and other public evidence).

The Statistical Office of the Republic of Slovenia (hereafter: the SORS or the Office) is a centrally organized government institution and therefore statistical surveys are carried out in direct communication with reporting units. Data collection is usually undertaken using the classical method on paper-based questionnaires, but for some surveys data are obtained on computer-readable media from existing administrative evidence and registers and are further processed without any data editing process. A general characteristic of the majority of surveys is a complete coverage of reporting units, which results in a

relatively high amount of new data input each month. On average, about 55 thousand statistical reports with about 7.5 million different source data are collected each month.

In accordance with the above-mentioned facts, technological and organizational approaches are set up to perform data editing for each individual survey.

2. TECHNICAL INFRASTRUCTURE

The existing technical infrastructure is of basic significance for the design and implementation of data editing applications. The main components are as follows:

- Mainframe. This is a main governmental server and the SORS is only a user. Here are stored all statistical data, classifications and other code-books (directories), Address Registers and main administrative registers including the most important Business Register, as well as applications and procedures. Apart from this, the mainframe provides production reliability and central functions concerning data security and complete back-up, which is very important for the data editing process.

- Minicomputers with Unix operational system linked to the mainframe. Five systems are dedicated to data entry and the data editing process and one to remote job submitting.

- Local Area Network.

3. DATA EDITING PROCESS, METHODS AND TECHNIQUES

As mentioned before, the majority of data collected at the SORS is still processed by traditional methods. In the data entry phase, high-speed manual data entry and OCR techniques are used in most of the surveys (interactive data entry is not yet widely accepted). At the moment several methods and techniques are used at the SORS. They are implemented in accordance with their matching the conditions of individual statistical surveys. If needed, a combination of individual methods is possible as well as changes in working techniques in order to increase the productivity of work and the quality of basic data sources. Figure 1 shows an extended work flow with phases involved in the existing data editing process.

Figure 1 BLOCK SCHEME OF DATA EDITING PROCESS

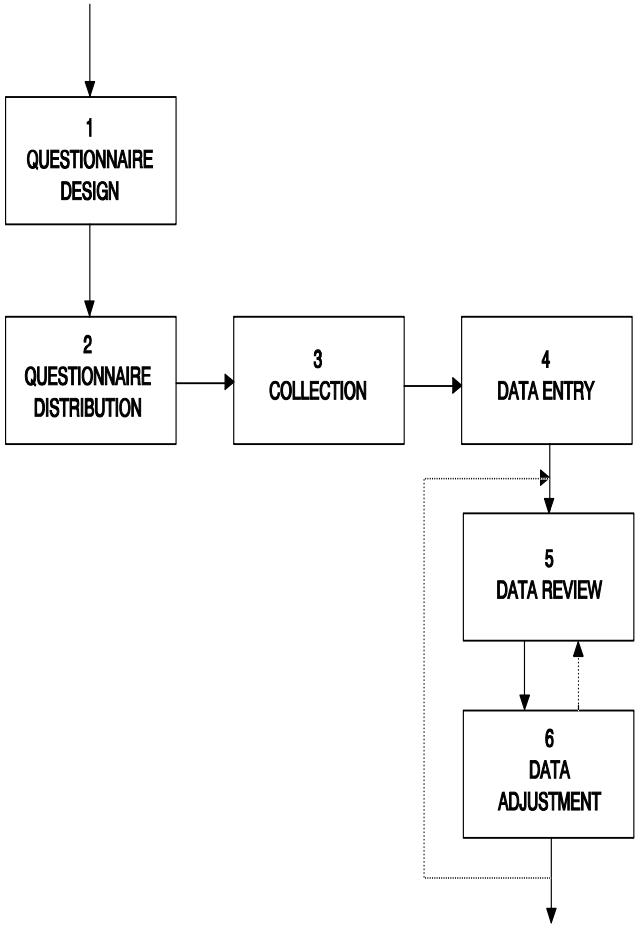
The main steps of the data editing process are as follows:

- **Questionnaire design**
  - Design of paper-based questionnaires
  - Design of electronic questionnaires with included editing rules

- **Distribution of questionnaires**
  - Address Register preparation
  - Addressing and numeration of questionnaires and mailing them to the reporting units

- **Data collection**
  - Completeness checking of questionnaires and survey
  - Manual data review
  - Urgent calls to non-responding units

- **Data entry**
  - Heads down and verification
  - OCR



- CAPI by Interviewers
- Storing survey data on data base (VSAM file) of source microdata

- **Data review and adjustment (correction)**
  - Batch processing (Cobol, PLI)
  - Interactive processing (GV-S, BLAISE)
  - Mixed batch and interactive processing

3.1 Questionnaire design

At this phase an important impact of new information technology can be recognized. Today the design of questionnaires is done exclusively on microcomputers (PC). The paper-based forms are prepared in a desk-top publishing environment with appropriate commercial software packages. Especially important is the design of electronic questionnaires (e-questionnaires) using the BLAISE system. The possibility to include certain rules and instructions for data entering and control in e-questionnaire is of great importance for speed processing and quality results. Because of these new possibilities the design of e-questionnaires is now closely connected with the whole data editing process.

3.2 Distribution of questionnaires

In this phase as well we can recognize a big improvement due to new technology. The distribution of questionnaires is based on address registers which are prepared individually for each survey. All questionnaires are properly addressed before they are sent to the reporting units. The address also contains the identification number of a reporting unit from the address register. Furthermore, in 1995 we began setting identification numbers for these questionnaires, written in bar-code technique, which enabled us to obtain quicker insight into and control of the collection process by scanning of questionnaires' identification. The distribution process is mainly automated. The e-questionnaires are distributed by portable computers and interviewers who also collect the survey data.

3.3 Data collection

The majority of statistical data are collected using a classical method, which means on paper-based questionnaires. For a small number of surveys, data are entirely or partly collected on computer-readable media, mainly on diskettes. Collection and concentration of data is performed on the basis of individual address registers. This phase of work is relatively critical from the operative point of view as the number of reporting units which do not send their data in due time (non-response) is increasing. Therefore, this phase of work is performed in parallel to the data entry and editing phase, although it should normally follow the concentration phase.

3.4 Data entry

*3.4.1 High-speed manual data entry (*heads-down*)*

For high-speed manual data entry we use a commercial applicaton package running on the UNIX operating system. The package enables entry and validation of source data under the control of user-created format programs. It is parametrically controlled and generates the mask for data entry on the basis of requests defined for each individual statistical survey, and within this for each different record structure. Thus different checks in connection with content, format and length of data fields can be performed at entry. The extent and types of control have been standardized for our needs, which means that there is a minimal but equal level of control for all surveys.

3.4.2 OCR data entry

At the moment only 6 statistical surveys (the most extensive ones regarding the number of reporting units) are processed using the OCR method. At present, about 30 percent of the total number of questionnaires collected annually are entered using the OCR technology. This technology was used for the entire data entry of the Census of Slovenia 1991, which was also a good opportunity to introduce this technology into the statistical data editing process. However, each survey covered by the OCR can be back-stored in a classical way. Forms which are not processed by the OCR system for whatever reason are later entered in a classical way. Because the work is now more rapid and rational, a formal control of data entry is not performed during the optical reading.

3.4.3 Interactive data entry

This method of data entry is based on different dedicated applications as well as application packages developed for other data editing functions (data review, data correction), where the possibility of data entry is built in as an additional function. The most important system used is the GV-S system developed by the SORS (Godar/Vega-STAT system). It was presented at the Work Session on Statistical Data Editing in Stockholm in 1993.

Data entry using the Blaise system is a newer approach which is expected to affect development due to the fact that the system supports integrated survey processing, and not only its individual phases as is the case in the existing software and technological process of work. The Blaise system was installed in 1993. Since then, remarkable progress has been made. Now about 15 surveys are processed using Blaise. Some of them are completely covered in all phases by Blaise, and in others Blaise covers some phases only of the survey, e.g. data entry and editing. CADI mode of data entry is often used in combination with traditional high-speed data entry on minicomputers, while CAPI mode was already successfully applied in some surveys.

3.5 Data review and correction (adjustment)

Batch or interactive/on-line approach is operatively used in this phase of the data editing process. A combination of both approaches is also often used. Usually the process begins with batch processing and continues in an on-line mode. In the past, this approach has proved to be very successful for extensive and complex surveys, where the initial ratio of clean raw data was very low. In this way, regarding the existing technological work process, the duration of the most time-consuming phase of work is reduced to a minimum.

The majority of surveys are, for the time being, checked in batch mode by Cobol or PL/I application programs. After performing data review, the process of correcting source data is continued. Data review routines produce error messages, which give the identification of erroneous data in the source survey data file. In some surveys these data can be corrected directly on video terminals. All further procedures of data correction run on a special standard software package for index entry of corrections INDXPO, which works in batch or in on-line mode.

More and more surveys (at the moment about 40) are edited using the GV-S system, which is particularly suitable for surveys with a high number of reporting units and a relatively small number of data per unit. The GV-S system enables, as well as data correction, the following functions:

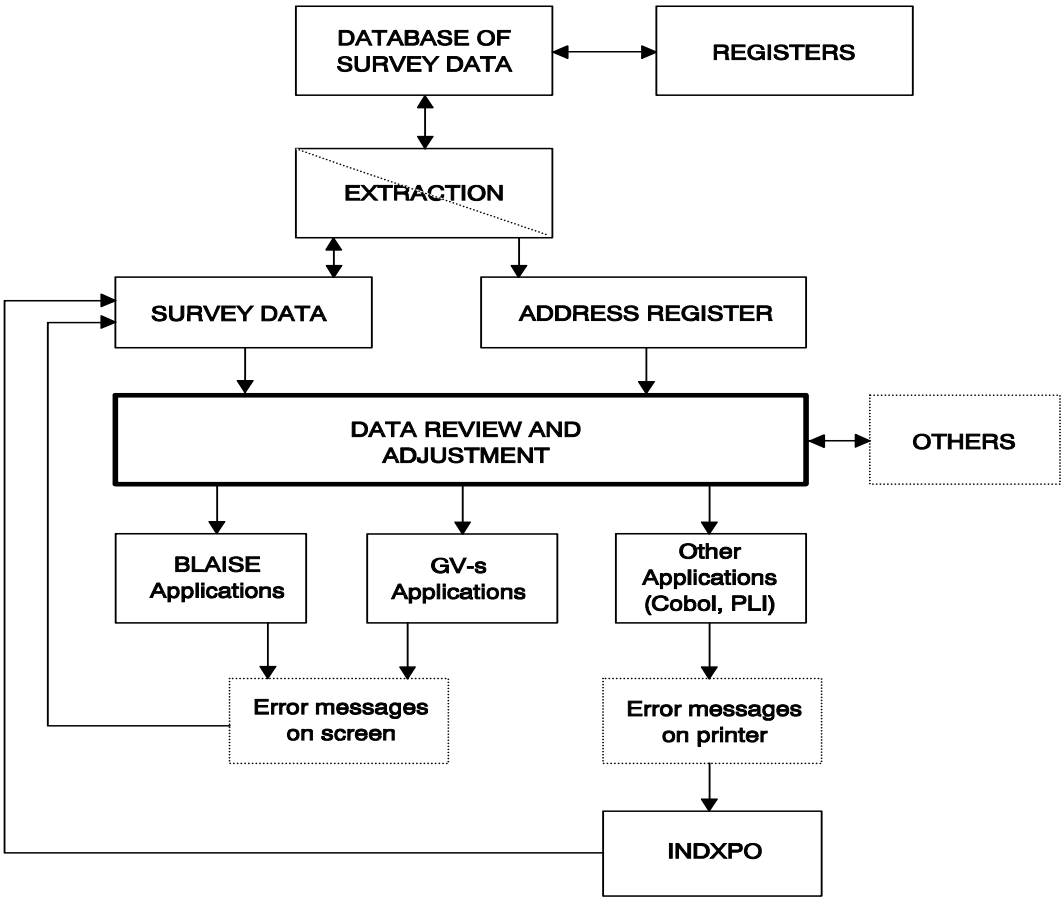
- statistics of errors with their frequency distribution;
- integrity of basic material and consultant tables;
- interactive performing of different queries and control tabulations;
- interactive performing and activating of final processing;
- automatic archive, etc.

The system is generalized in relation to individual surveys as well as for different levels of users (project development, data entry, data review and correction, analyses of results, final processing, etc.). The GV-S system enables, due to this generalized approach, the relatively rapid and standardized development of applications, as well as rapid and effective data correction.

Figure 2 shows the process of data review and adjustment.

The importance of data review and correction using LAN and the Blaise system is rapidly increasing at the SORS. The possibility of a gradual and adjusted introduction of Blaise into the existing system of data processing is one of its great advantages. The smooth and user-friendly transfer of different types of data files under the Blaise system enables import, review and correction of data entered by some other systems, as well as data export to the system where the data are finally processed.

Figure 2 DATA REVIEW AND ADJUSTMENT



4. THE IMPACT OF NEW TECHNOLOGY ON DATA EDITING

In the past, computer equipment and system software were strongly centralized and rigid, so that only batch processing was possible. Recently, we have recognized a tremendous development and change which consequently generate new possibilities to change the old strategy, methods and techniques of the statistical production process. Using these new possibilities, we have reached significant improvements in different directions such as:

- development and implementation of interactive applications promoting the data editing process, which brought us quicker survey results;

- introduction of personal computers and communications (LAN) to statisticians, integrating them properly into the data editing process; this brought us further improvement of data quality;

- integration of particular phases and subphases in the data editing process, which brought as further rationalization, etc.

The impact of new technology is better reflected in data editing. The GV-S system was the first system to introduce interactive data editing at the Office. Although it is a mainframe system it was a good source of experience for the new development of interactive applications on microcomputers and LAN. Experience shows that people who started to work interactively on the mainframe need less time to adapt to interactive applications on LAN.

Recently, more and more data editing activities at the SORS are being performed on networked or stand-alone microcomputers. The development is based on the Blaise system for integrated survey processing. Main benefits and advantages were observed in CAPI and CATI surveys, where data entry is integrated with editing, and the respondent is available during the interview as a source of correct data. Computer-assisted interviewing reduces the time required to obtain clean data and provides a better quality of answers. There are, however, disadvantages as well. Different approaches to survey processing can not always be easily integrated into existing standards at the Office, new methods and techniques are often ineffective without radical reorganization and they can

usually not be generalized to all kinds of statistical surveys.

CADI-based surveys (Computer-Assisted Data Input and editing), although less effective than CAPI or CATI, are more flexible and they can be more easily integrated into the traditional (batch-oriented) organization of statistical production. This is especially important in the process of downsizing, where some parts of the applications are already based on the new (LAN) platform and the rest is still dependent on the old one (mainframe). Our experience shows that a combination of CADI and mainframe applications can be very effective in the surveys, where some activities (e.g. back-up, archiving or data entry) still need to be performed on mainframe. In the "combined" approach, an interactive mode of work as well as batch processing can be used.

The application of computer-assisted methods and techniques introduces many advantages, but also some new problems. Thus the effects need to be analysed carefully and results studied and used for future planning. New technology of data editing also impacts internal organization of work at the Office. New possibilities of decentralization and the moving of some data processing activities to the subject-matter departments introduce some new needs such as training, coordination, reorganization of departments, etc. - all important issues and requiring close attention.

Another very important subject is a design of user interfaces. Tools and utilities of the minicomputer environment enable the better and more flexible design of user interfaces, which can be defined as more user-friendly. It results in better control and overview of all the phases of data processing. The majority of instructions are moved from the paper to the screen, which results on the one hand in faster informing and processing, but which on the other hand can cause lack of written documentation.

5. FUTURE WORK AND ORIENTATION

Our main policy is the standardization of individual phases of the production process, such as collecting data electronically, replacing the paper questionnaires with electronic ones, in which the source data should already be at least partly controlled. The introduction of electronic forms and inclusion into Electronic Data Interchange (EDI) is, however, our important future task in this area.

In the transitional period, which will, in our opinion, last at least 4-5 years, we are planning to extend the use of generalized applications for data editing also to all other statistical surveys, which are at present processed by Cobol and PLI applications.

At the moment the data editing process is still very much mainframe-oriented, which could, in the near future, be a limitation for further development and reducing real costs of work. Therefore, the data editing process has to be planned for future rightsizing to lower and more flexible system platforms.

In the following transition period we are planning changes in computer equipment at a lower level, mainly minicomputers and existing asynchronous terminals with intelligent terminals and personal computers.

6. CONCLUSIONS

The data editing process is organized and performed on the basis of different and mutually complementary approaches. Generalized and non-generalized approaches are included in this process as well as interactive and batch modes of work, which can be mutually interacted according to the concrete objectives. The use of different methods of data entry, data review and correction is the consequence of the gradual introduction of new standards, technologies, and solutions in this area. The changes and new solutions are being planned and introduced gradually, which is of essential importance for the smooth continuation of the work.

ELECTRONIC DATA INTERCHANGE FOR STATISTICAL DATA COLLECTION

By Wouter J. Keller and Winfried 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 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.

Keywords: EDI, Meta-information

1. INTRODUCTION

Statistics Netherlands is at present under the influence of several developments. As everywhere else, it no longer operates as an untouchable organization 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 data suppliers. The outcome should be a product which, although not actually sold on a market, our clients eventually want.

We are confronted with new developments in Information Technology (IT) which will give us the opportunities to construct the necessary tools to meet the new demands. In a situation like this, a national statistical institute (NSI) needs to make the right strategic choices.

2. DEMAND-PULL

The production process is influenced by the growing demands of our clients and respondents. There is a strong political demand for a decrease in the response 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".

Furthermore, budgets are shrinking so there is a demand for higher efficiency and higher productivity. Concerning our output we see a demand for greater user-friendliness. One particular aspect is a demand for an improvement of the coherence of the totality 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.

3. TECHNOLOGY PUSH

However, we are blessed with information-technological (IT) 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, on paper. Our suppliers produce their data by electronic means and will want to use those means to deliver their data directly to us in order to minimize 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.

4 STRATEGIC CHOICES

New demands and new tools will affect all aspects of our production process. To describe them let us first discern three stages within this production process. The input-phase is where the data are collected in contact with the respondents. In the throughput-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-phase, the collecting of data. First, data collection among individuals and households. It is not an exaggeration to state that a major step forward has already been taken at Statistics Netherlands. We have introduced all kinds of Computer-Aided Interviewing (CAI) and have developed BLAISE to do this. (It is needless to say that BLAISE does more than develop and present electronic questionnaires.) The advantages 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. Still, there is even more to gain. First of all in the efficiency of the production process itself. But also in the statistical sphere improvements are still possible: new ways of interviewing: CASI, computer-aided self-interviewing and, although not directly an IT matter, 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 represents for interviewing among households, EDI (electronic data interchange) will represent for data-collection among enterprises. Later in this paper we will go deeper into EDI with enterprises.

In the throughput 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 of handling 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, the output phase. Here 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 the more professional user particularly will want to select and receive his data by electronic means. Statistics Netherlands is producing or developing these means: data on CD-ROM, data on Internet.

More important, and perhaps more difficult, is the way data should be presented using these new media. The amount of information will be much greater than that in our paper publications. At this point the management of the meta-information becomes crucial.

For this purpose Statistics Netherlands is developing STATLINE. This should lead to a data-base intended for the end-users that should give access to "all" our data. As could be expected, structuring these data is the main problem. At the same time we are confronted with a lack of coherence due to a lack of statistical co-ordination. 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 decision has been made to aim for that structure wherein all publications and all other dissemination of data goes through STATLINE.

5. 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 accept these 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 made. This is particularly inefficient on the input side.

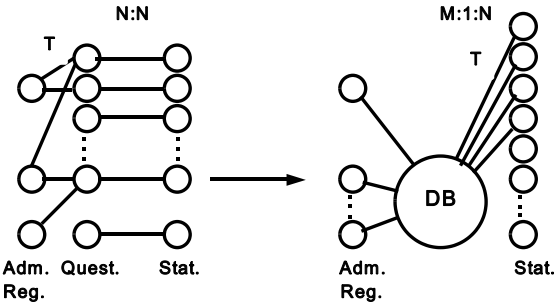
In the new situation (we are talking more than 10

years from now) data collection especially will be re-ordered. No longer the demand for information but the supply, the available actual data-sets, will dictate the organization 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, the integrative systems such 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:

Figure 1. Process: old vs. new (2000+)



On the left we see the old situation with a separate production line for each individual statistic. On the right, 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.

6. ELECTRONIC DATA INTERCHANGE (EDI)

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

A NSI collects data to produce statistical output. A translation needs to be made 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 initial 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. Most particularly, we and our respondents would like to avoid data-entry.

7. 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. For each possibility we will indicate the nature of the translation and especially who is going to make it. We concentrate on the conceptual translation.

7.1 EDI on centrally-kept registers

We do not approach the individual respondent here 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 way of data collection creates no response burden.

There are, however, disadvantages, the most important being 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% 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%?

In the Netherlands there are several examples of usable registers. There are centrally-kept registers of enterprises with the chambers of commerce. The tapes of these registers feed our own register of statistical units. Statistical data can also be had from fiscal (company tax, VAT) or social security sources. For several sources (chambers of commerce, company tax and VAT), the possibilities are being either used or researched.

7.2 Commercial bookkeeping bureau's

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

Having said that these offices often hold much of the information required by the NSI, 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% of the enterprises in one particular branch. In this case it is profitable for the NSI to make the necessary translation. In other cases, we propose providing software by which the bureau itself makes the necessary translation.

7.3 EDI on individual respondents

When the possibilities described above 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. In the Dutch situation, the financial records and those on wages are strictly separated.

Here we classify by the translator of the information.

7.3.1 The NSI translates

One of our EDI-projects, EFLO, works along this line. 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. What is important here is that we are dealing with a limited number (600) of respondents.

7.3.2 The respondent translates to a standard record

Here a standard record of information is defined. The standardization regards both the conceptual and the technical aspects. Producing the record and 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 standardized to a certain degree among respondents. Furthermore, to make a standard record possible the NSI may sometimes have to move towards the concepts of the respondent. In such a case a larger part of the total translation to the final statistical output has to be done by the NSI.

This mode of EDI has a clearly favourable effect on the respondents' burden, particularly when the standard record is available in the bookkeeping software the respondent uses and regularly updates. 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 regards wage information. This subset of company records is highly regulated in the Netherlands, thus making it possible to define a standard record.

7.3.3 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.

7.3.3.1 Paper questionnaires

This clearly is no form of EDI. We mention it as a possibility in order 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.

7.3.3.2 Electronic questionnaires

Although, strictly speaking, this is at the most only 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.

7.3.3.3 "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 financial records and described in the next section.

Before we go into this, we give below 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

8. EDI-PILOT 2

We will now describe the project EDI-Pilot 2 directed at the financial records of individual enterprise 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 form only a part of the accounts of an enterprise. Accounts on wages and employment are particularly excluded. This is not a choice voluntarily made by Statistics Netherlands but one forced upon us by the way the bookkeeping systems are organized 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 a 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 organization 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 provided explicitly by the respondent, meaning by 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.

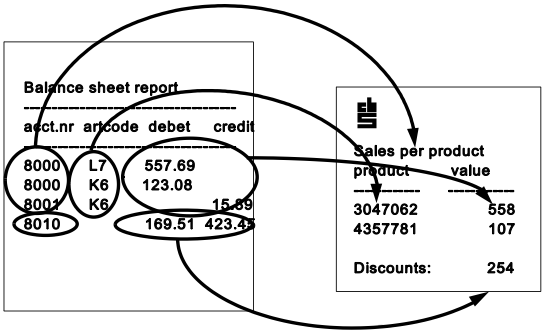
Financial accounts also differ in their technical lay-out. A large number of bookkeeping software systems is in use. There is no standard record for information to be selected electronically from the software and it is not expected that it will be possible to define such a record within the near future.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

This means that some ingenuity was needed to create the automated link we were looking for. This was done by using the reports or print-outs of the software system. Instead of printing them, they were sent to a file, a print-file, to be read by the translator, the main part of the software module that will run on the respondents computer that 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 lay-out 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.

We see, then, the two parts of the translation scheme. The first part lays down the lay-out 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.

Figure 2 THE TRANSLATOR



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. The second time the translator needs to produce statistical information, he can use the already available translation scheme. 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.

9. SCOPE OF PILOT 2

As already mentioned, Pilot 2 is directed towards the financial accounts. The principle is that all the information that is tapped from the financial accounts by any statistician of Statistics Netherlands will go through Pilot 2 if automated retrieval of that information is possible. In practice this means that several large areas of 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 in this area in IRIS, but also the possibilities of obtaining enough foreign trade data whilst aiming in the first place 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 production statistics. They cannot be addressed using 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 relatively small amount of information asked for here, more is expected from centrally-kept records (VAT, corporate tax) and from the practice of book-keeping offices to often keep books for hundreds of smaller enterprises. The very large enterprises are also excluded. Because of their complexity, they need an individual approach also by means of EDI but, in this case, "tailor made".

Regarding the number of respondent participating in this kind of EDI, we should mention that in pilot 1 a number of 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 hundreds of respondents. Pilot 2 will also be used to approach the book-keeping offices. 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 a number of 25,000 units to be approached with this instrument, partly through the book-keeping offices.

The revenue of Pilot 2, if successful, will represent 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. In the second place, new activities arise in the form of a growing help-desk and a field-service that will not only have to cope with book-keeping problems but also with technical automation problems.

10. CONTROLLING PILOT 2: THE META-SYSTEM

Eventually Statistics Netherlands aims to reach several thousands of respondents. This, of course, requires 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 this part.

Constructing the combined questionnaire involves the coordination of the approach of the different statistics, aimed at the financial records, but also with the book-keeping practices of the respondents. EDI will be 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 this 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 into 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 book-keeping 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 receives its own combined questionnaire.

11. 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 access them 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.

We now see that the character of meta-information has changed. In most of the literature we often find meta-information as a merely 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 is now 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) co-ordination. 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. What is remarkable is the fact that this growth in statistical co-ordination 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, in other words all the problems of statistical co-ordination, 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.

NEW DEVELOPMENTS IN AUTOMATED DATA COLLECTION: ELECTRONIC DATA INTERCHANGE AND THE WORLD WIDE WEB

By Richard L. Clayton, Tony M. Gomes, and Louis J. Harrell Jr., Bureau of Labor Statistics, USA

ABSTRACT

A growing number of automated data collection methods have been developed for establishment surveys since the advent of the microcomputer. These include Computer Assisted Telephone Interviewing (CATI), Computer Assisted Personal Interviewing (CAPI), Touchtone Data Entry (TDE), and Voice Recognition (VR). Survey organizations have exploited this new technology to improve data collection operations, namely: data quality, timeliness, costs, and respondent burden. Furthermore, the continued growing availability of sophisticated technology such as Fax, Electronic Data Interchange (EDI), and the World Wide Web now available to the survey *respondent* offers new opportunities for further improvements in survey data collection. This article shows experiences gained during the implementation of the last two technologies.

Keywords: EDI; data quality; Internet; World Wide Web.

1. INTRODUCTION

The U.S. Bureau of Labor Statistics (BLS) has made effective use of these automated collection methods in many of its surveys. Now, two relatively new automated collection methods are being developed and field tested by the BLS—EDI and the Web. EDI targets large, multi-establishment firms while the Web represents the movement of TDE approach onto the "information superhighway." The Web offers an intuitive, visually appealing interface and it is suitable for reporters with limited numbers of establishments.

The use of EDI as a data collection method for large, multi-establishment firms is discussed in Section II. This section reviews EDI as an industry standard transmission method, its potential use in survey data collection, and its implementation in two BLS establishment surveys. Section III describes the prototype Web collection system and its relationship to TDE, and identifies considerations in the development of a Web survey data collection system. Costs of both EDI and Web collection methods are discussed in Section IV. Section V contains some concluding remarks.

Two Establishment Surveys

The Current Employment Statistics (CES) is a voluntary monthly survey of 390,000 non-farm business establishments yielding estimates of employment, average weekly hours, and average hourly earnings at the national, state, and metropolitan area levels. With preliminary estimates published after only 2 weeks of data collection, the CES provides one of the first indicators on the health of the U.S. economy. *The Multiple Worksite Report* (MWR) is a quarterly report of employers with 10 or more business locations used to collect employment and wages to supplement statewide data provided by employers through State Unemployment Insurance (UI) covering virtually all U.S. businesses. The establishments covered under UI laws provide the population frame for the CES survey. Both surveys traditionally have been conducted by mail and in a Federal-State

cooperative agreement whereby the States collect the data for their use as well as for the Federal Government.

Overview

The bulk of the CES sample is comprised of single unit reporters and small multi-establishment firms. Approximately 90% of the CES sample respondents report for 6 or fewer establishments and are therefore eligible for TDE or VR reporting. However, as the Internet becomes more accessible to these respondents, the Web could eventually replace TDE/VR. Of the remainder, 1% will be eligible for EDI, and the rest will use mail or FAX.

2. ELECTRONIC DATA INTERCHANGE (EDI)

2.1 The Use of EDI for Normal Business Transactions

In its most formal sense, EDI is the use of accepted industry standards for data formats, called transaction sets, for transmission of routine business transactions. Less formally, EDI is the transmitting of machine-readable files—the so-called "flat files" or ASCII files—using electronic transmission, tape, or diskettes.

There are two major industry standards governing "formal" EDI. In the U.S. the Accredited Standards Committee X12 standard is maintained by the American National Standards Institute (ANSI). The other standard is the United Nations Electronic Data Interchange for Administration, Commerce and Transport (UN/EDIFACT) which is used in most other countries but is gaining acceptance in the United States. The two organizations are working together to merge towards using EDIFACT as a single standard for worldwide use. A transaction set specifically designed for reporting statistical data was developed by the U.S. Bureau of the Census and has been approved by the X12 Committee[1].

2.2 The Use of EDI in Establishment Surveys

The benefits of EDI to the firm are reduced burden and costs for reporting. The benefits for statistical agencies are increased volumes of data at low costs, reduced errors, and other potential efficiencies such as timeliness. The need to integrate EDI into the U.S. federal survey environment has been realized at the highest levels. In 1993, the U.S. Office of Management and Budget has released instructions requiring that every justification for data collection include a statement describing consideration of electronic data collection. Under these guidelines, every government agency will have to evaluate EDI. Eventually, the wider use of EDI by the government agencies will lead to greater use in the private sector.

The use of EDI for surveys stems from two major forces. First, many firms in the U.S. are *demanding* low-cost, low-burden solutions to survey requests. A few firms have actually refused participation in some BLS surveys based on the high burden levels found in overall survey reporting. Secondly, the need to find low-cost collection methods within the survey organization is pushing us towards EDI and the other computerized methods listed earlier.

EDI's potential as a survey methodology takes advantage of central databases in large, multi-establishment firms for extracting survey responses in a single transmission. The databases must be sufficiently accurate and timely to satisfy the survey agency's data quality and deadline requirements. Also, to avoid specification errors, it is important that data concepts and definitions of the survey agency be carefully matched to those of the firm. Only then will this paperless approach lower respondent burden and, if implemented carefully, improve data quality.

2.3 EDI Implementation

Collecting data electronically requires a single collection site, rather than 50 State collection agencies, as is the case in the Federal-State environment. Thus, the BLS established an EDI Data Collection Center in Chicago, Illinois in early 1994. The Center is now responsible for solicitation, data collection, editing, dissemination to the States, and represents a single point of contact with the firms. The group for which the return on EDI investment will be greatest consist of the largest firms in the survey. The definition of "large" may vary across surveys, but will generally mean those firms providing large volumes of data. Large, multi-establishment firms providing employment and payroll data, for example, would meet the size criterion. Also, survey collection which occurs frequently, such as monthly or quarterly, are good candidates for EDI.

2.3.1 Statistical Issues

The current number of firms meeting the EDI size criterion is small. For example, in the U.S. there are about 5.7 million firms covering approximately 7.5

million individual establishments. Of these firms, approximately 650 have 50 or more establishments and report for both the CES and MWR. However, these firms account for 9.5 million employees, or 8% of the total U.S. employment. Thus, with a relatively small number of contacts, vast amounts of data are potentially available. Most of these firms will likely be included in all surveys' certainty strata, making them very important to each survey and to the survey community as a whole. The current collection methods continually overburden these reporters and may lead to withdrawal from all voluntary surveys and reduced cooperation with mandatory surveys and censuses.

To minimize reporting burden, we expect the EDI firm will report for all its establishments rather than only the ones included in the sample. By receiving all establishments, the survey agency may need to develop sub-sampling routines or separate estimating cells for this portion of the universe. A file for all establishments may also contain new establishments (births) since the last transmission. Usually, identifying new births and soliciting their participation are difficult and expensive activities. For EDI respondents, this can be relatively inexpensive and immediate. For employment surveys, the capturing of the surge of new employment, new establishments, and deaths is critical to accuracy and economic understanding.

The switch to EDI may also change the level or detail of reporting. For example, if a firm previously provided a single company-wide report, the switch to EDI may allow more detailed reporting at much smaller levels. The individual establishments may be classified differently thus affecting key variables such as industry, location, and size which will ultimately affect estimation series.

A detailed review of the firm's data elements will identify the suitability of EDI for recurring extracts of pre-defined data elements. This review must encompass all three dimensions of data definitions: timing, content and method. Timing refers to the correct reference period; content is the inclusion or exclusion of various components making up the definition of a concept; and method is the means for calculating the desired measure. All three dimensions must be carefully reviewed prior to the programming of fixed extracts.

The next step is to test incoming datafiles for usability before inclusion in estimation. A brief overlap period (3 months for CES and 2 quarters for MWR) during which data are reported by both paper and EDI should be used to ensure data quality. The emphasis on data quality from the outset is essential for ensuring accuracy and preventing future problems. To use the vernacular of the Total Quality Management community, "do it right the first time." During the overlap period data can be reviewed for consistency.

Timeliness is also a concern. EDI respondents will likely provide a file containing all establishments. However, if waiting for all units to be included in the transmission jeopardizes the survey's timeliness, then arrangements must be made to receive partial files for preliminary estimates and the remainder for revised estimates.

2.3.2 Format and Transmission Options

Many large firms have established organizational units, EDI shops, for handling necessary programming and communications with their business partners. As more and more of these EDI shops expand into firms' human resources and payroll operations, which are the source of most establishment data, EDI becomes a viable data collection method. For the time being, the BLS offers employers the choice of using formal EDI (X12 format) or flatfile. The BLS has developed a standardized flatfile format that allows employers to report data for both the CES and MWR using a single format and communication protocol. In addition, to the two format options, the respondents are offered two transmission methods—either through a Value Added Network (VAN) or directly to the BLS via a bulletin board system. The Internet may also be used for such transmissions.

2.4 EDI Implementation Results

2.4.1 Solicitation Results

Unlike CATI and TDE, the conversion to EDI is a rather slow process because it requires that the firm use resources for programming and transmission. Since most firms do not place responding to surveys at the top of their priorities, conversion to EDI can take as long as 12 months. Since solicitation began in 1994, over 200 firms have been contacted, of which 28 have been converted to EDI. The 28 firms now reporting through EDI account for 14,000 establishments and 1.6 million employees. The conversion target is approximately 650 large firms comprising 135,000 establishments and 9.5 million employees. So far 32 firms have refused EDI, however, none has refused the concept itself. The

most common reasons for refusing EDI are the inability to justify programming costs and the lack of data processing support.

2.4.2 Quality Improvements

The careful review of data concepts, contents, and timing conducted with the respondent prior to EDI conversion has produced numerous quality improvements, some of which are listed below:

- Respondent corrected employment counts where, previously, all active employees for the entire month were reported, which incorrectly included employees who did not work during reference period.
- Respondent began reporting employment, hours, and earnings for non-supervisory employees.
- Respondent erroneously included employees who were in an "active" status but did not receive pay for the pay period which includes the 12th of the month.
- Respondent previously reporting data at county wide level rather than at worksite level.

Table 1 shows a typical firm "before and after" conversion to EDI. Prior to EDI, the respondent spent 40 hours per month preparing reports transmitted to 35 States using three different reporting methods, covering 55 establishments and 58,000 employees. With EDI, the BLS now receives a single transmission covering all establishments and employment in the firm, 1,171 and 103,735 respectively. In addition, the firm's reporting burden has been reduced by 90% to only 4 hours per month, while both timeliness and item-nonresponse have improved.

Table 1. CES Electronic Conversion Experience

	Prior to Conversion	After Conversion
RESPONDENT BURDEN		
Cost	40 hours/month	4 hours/month
States/Contacts	35	1
Collection Methods (Mail, CATI, & TDE)	3	1
COVERAGE		
Employment	58,000	103,735
Establishments	55	1,171
TIMELINESS (Mean Closing)[1]	1.6	1
DATA QUALITY		
Accuracy	Unknown	Correct reference period
Item non-response	21%	0%

[1] Mean Closing of 1 indicates that on average the report was received by the first deadline.

Recommendations and Lessons Learned: *Electronic collection is preferable to other methods of collection.* EDI offers both large, multi-establishment firms and the survey agency the most efficient way of collecting data. Once the up-front work is completed, significant gains in efficiency will be reaped by the firm. With the entire data compilation and transmission process automated, little month-to-month intervention will be required. From the agency's perspective, the standardization of electronic data format enables the agency to design a single processing system capable of handling large amounts of data from multiple respondents.

Standardization of approach to EDI is necessary. In the U.S., where there are several major statistical agencies, the need to establish standards for data formats, documentation, coding, etc., is critical to EDI's acceptance by the business community. Standardization and consistency across surveys will reduce the likelihood of errors and increase the likelihood of respondents embracing electronic reporting. User guides should follow standard guidelines and styles to the extent possible to facilitate the participation of the EDI team in the firms.

Establishing EDI reporting with a firm requires a significant commitment of resources by the firm. The conversion to electronic reporting requires an up-front investment by the firm, so the decision must usually be made by a middle level-manager, such as the payroll manager, or someone higher. A formal request for programming from the firm's Data Processing (DP) department is typically required. The collection of employment data, especially if voluntary, is not one of the firms' priorities. Therefore, the first step is to approach payroll and tax personnel who actually complete the survey forms and who are more like to see the benefits of electronic reporting.

3. THE WORLD WIDE WEB

Electronic mail (E-mail) and World Wide Web services are increasingly available within businesses and may be exploited for survey data collection. The BLS has developed a prototype Web collection instrument for a pilot test of Web collection in the monthly CES survey. Respondents receive E-mail requesting that they enter their data in a Web page. The data are immediately edited and transmitted to the survey agency's computer. The Web offers an intuitive interface, low cost, and a standard, easily managed data record format. Cost reductions are obtained through automated editing, also allowing improved data quality. Links to other related sites can be provided, giving the respondent access to survey data products.

3.1 Web Survey Methodology Compared to TDE

TDE respondents receive a monthly advance notice message sent by postcard or FAX. This message replaces the arrival of the survey form as a reminder. Data collection is performed by dialing the TDE system and entering data as requested by the digitized verbal prompts. Non-respondents receive telephone or FAX prompts on specially designated days conforming to the availability of their payroll records [6].

The CES Web survey collection cycle parallels the TDE respondent contact process. It begins with a sample control file containing the respondents' E-mail address in addition to the normal respondent contact information of name, address, and phone number. The collection form is a standard "Web page" containing the questionnaire, survey instructions and hypertext links to definitions. An E-mail address is provided for reporting problems and inquiries. As the collection cycle approaches, the respondent opens their E-mail to find a reminder, connects to the CES homepage, accesses the data collection screen, and fills in the requested data. The moment the respondent clicks the "submit data" icon, the data are transferred to the BLS. Schedules are checked-in and, at predetermined time periods, E-mail nonresponse prompting messages are sent.

Our current labor-intensive editing and reconciliation operations can be vastly streamlined under Web collection. The respondent will address all edit failure questions through on-line edits generated immediately after data entry. This change will allow the elimination of the large semi-clerical operations of reviewing edit failures.

We can implement both longitudinal and data integrity edits in the Web environment. Integrity edits are based exclusively on rules, while longitudinal edits require immediate access to several months of previously reported data. Security considerations become important if historical data are located behind a firewall. We have developed prototype integrity edits and are developing longitudinal edits.

3.2 Total Design Method On-line

The eventual replacement of traditional methods with the Web will require a careful review of all mail-based research. The results serve as reasonable starting points for Web methodology. TDE has attained high response rates using a combination of advance notices, easy to use data entry interfaces, and carefully-timed nonresponse prompts. Will Web methodology work the same? Also, the Total Design Method (TDM) offers a rigorous approach to maximizing response rates [5]. Under the TDM, each survey feature (prenotification message, the survey instrument, reminders and the timing of each) carries potential for improving response rates. Will Web collection behave similarly to mail with regard to these? Will the response rate increases seen be commensurate under Web? How does forms design research carry over into research on screen design and human-computer interface? These, among other questions will be evaluated in the CES Web pilot tests.

3.3 Web Versatility

Unlike telephone collection methods, Web collection can accommodate a wide range of surveys and survey operations. The use of telephone collection is often limited by the length and complexity of the questionnaire, the frequency of the collection cycle, and the need to immediately respond to an interview question:

- ***Length:*** The Web has the ability to accommodate structured questionnaires of any form or length including "form-layout" or "question-by-question" designs. The respondent has the ability to refer to records as needed or to partially complete the questionnaire and return to it at a later time with no noticeable effect on costs.

- ***Frequency:*** Ongoing Web surveys will be easy to maintain if a file of contact information, including the E-mail address is used. One-time multi-mode surveys are more difficult to implement because accurate and up-to-date contact information may be difficult to obtain.

- ***Altering Content:*** The Web system can be modified and loaded at a single point. Once loaded, all respondents have immediate access to the modified software.

3.4 Product and Customer Service Improvements

The improvements offered by automation will ultimately lead to more accurate microdata, more timely responses, and improved customer access to our survey products:

- ***Accuracy:*** The respondent is able to see all data displayed prior to submission, providing an additional opportunity to review the data for any errors. Response rates should also increase since nonresponse prompting can be handled on a more timely and controlled basis.

- ***Timeliness:*** Our customers will benefit from more timely data. For some surveys this will mean "final" estimates will be quickly available and thus will improve the accuracy of "preliminary" estimate surveys.

- ***Customer Service:*** We will be able to provide our respondents a profile of their firm's information against national (or State) industry averages derived from the survey results.

3.5 CES Model for Web Collection

The entire Web environment is rapidly changing. Features which were not available even a year ago are entering the marketplace on a daily basis. Each new advance in hardware, software, and communications represents new opportunities and challenges. In 1995, the CES program developed a “proof of concept” model of a Web data collection system. This prototype was implemented on a Sun Sparc 10 workstation, using Solaris 2.4 UNIX. The server software we chose was the National Center for Supercomputing Applications HTTPD Version 1.4. The site was developed for Mosaic browsers. In 1996, we moved to a new configuration. The CES Web prototype system uses a Windows NT server, Netscape Secure Commerce Server, the Netscape browser, and a digital ID from Verisign, Inc. The respondent needs to have a browser that supports Hypertext Markup Language (HTML) tables and the Secure Sockets Layer (SSL) protocol. Browsers, such as Netscape’s, can be obtained free. HTML tables are required for forms-based data entry while SSL provides security.

3.6 Security

Perhaps the single most critical feature of the Internet infrastructure is the security of the transmitted information. This limitation is repeated by every student of the Web and is drawing the attention of much of the computer community. The CES has met its goal of a C2 level of security. C2 refers to a

security standard developed by the National Security Agency. Some characteristics of the Web security profile are:

- Authentication of the respondent
- Protection against snooping during transmission (Packet data security)
- Protection of the session (hijacking)
- Protection of confidential data once it has arrived at the server
- Prevent non-BLS user access to the BLS LAN

3.7 The CES Web Pilot

In April 1996, we began collecting CES data from 7 firms. These respondents are in the Services industries and were already reporting by TDE. By November 1996, the sample had expanded to 55 firms. Respondents were contacted to determine whether they have access to the Web, their E-mail address and their willingness to participate. Eligible units received a specially developed package describing Web reporting. They were asked to try the system within the next 2 days and a CES interviewer would contact them to assess their reactions. Periodic follow-up will provide a source of feedback for ongoing improvements in the user interface.

3.8 Results

As of November 1996, we had converted 14% of the units we contacted in the hi-tech industry. It is too early to provide any estimate of an upper bound to the number of Web-eligible respondents. Respondent comments have been favorable. Our results are summarized in Table 2. Firms in Computer and Data Processing Services were selected first for solicitation, as we believed they would be more familiar with the Internet. Other services industries and local government reporters were then contacted to see if characteristics were similar.

Our results lead to some interesting observations. While many companies have Web access, not all staff have access to the Web. Our solicitation criteria for Web reporting are relatively strict. We only enroll respondents, in our case, payroll clerks who 1) have Web access from their desktop PC, and 2) have at least Netscape 2.0 or Microsoft Internet Explorer 3.0 browsers. Web access is not yet universal for the CES respondents. In fact, it is very limited. Those with Web access can receive E-mail, respond to E-mail advance notice prompts, and respond to non-response prompting. Based on our small sample, we have found that respondents take action based on the E-mail messages in the proportion as FAX or postcards messages.

There is anecdotal evidence of a rapid growth in Web access among payroll clerks. In 1995, the BLS asked a non-scientific panel of firms a series of questions regarding E-mail availability and use. Comparing these results to our current results shows that there has been an increase in the availability of E-mail access. In 1995, 7% of firms could send E-mail outside of their company[4]. In 1996, the number is 34%.

Table 2. Web Sample Solicitation Results: November 1996

	Current Totals	Computer and Data Processing Services, n=313	Other Services n=121	Local Government n=264
E-mail only	4%	7%	0%	2%
E-mail and Web, not on desktop	2%	4%	0%	1%
Compatible browser, E-mail/Web	**11%**	**14%**	**0%**	**10%**
Prefer other mode	2%	3%	0%	1%
No E-mail/Web, out of business	66%	53%	87%	76%
Incompatible browser	5%	6%	6%	1%
Firm has capability, respondent doesn't	10%	13%	7%	9%
Total	**100%**	**100%**	**100%**	**100%**

3.9 Future Plans

Over the next year, we plan to implement many system enhancements to our Web site. We will expand

the industry coverage to allow all industries to use Web reporting. Client side editing will be introduced. Our current edit prototype uses JavaScript to test data integrity. Our preference is to have a Java applet that would do client side editing. In addition, we would like to integrate automated generation of nonresponse prompting and advance notice messages with the collection system. We are researching approaches that would allow us to conduct longitudinal editing of data. Enhancements to the user interface are also needed.

The use of video offers a broad area for research, drawing on knowledge about respondent-interviewer interaction. A quality enhancing activity that we are studying is the use of streaming video for special surveys and general information. Streaming video technology allows video to be transmitted across the Internet at relatively slow speeds, 28.8kb per second. If the respondent has a sound card in their PC, they can also receive audio. Special surveys could be introduced and explained with a video clip.

As was discussed earlier, the BLS is now offering EDI collection. The standard transmission approach to EDI is via a private network or a BLS bulletin board. While there is a cost for this type of transmission, Internet is relatively free. The economic benefit of using the Internet will eventually cause EDI and Web reporting techniques to merge. We envision EDI respondents linking to a Web server and securely transmitting their datafiles.

The essential production activities supporting Web and TDE/VR will be integrated into a single system. A single sample control file will record the type of messaging required for each respondent. A standard data record will be produced and uploaded to the estimation system.

4. EDI AND WEB COSTS

Costs for ongoing collection activities can be reduced through electronic reporting. First, normally labor-intensive activities such as mailout, mailback, and entry costs are eliminated. Secondly, consistently rising postage rates are replaced with less expensive and declining costs of telephone access and usage. Lastly, the remaining workload can be averaged over the number of individual reports received. Looking only at the variable costs of data transmission, Table 3 compares EDI and Web unit transmission costs with three other collection methods. The unit transmission cost under EDI is about 1/3 of mail while under the Web there is virtually no transmission cost. Note the dramatic impact on costs obtained from moving to TDE and Web collection [3].

Table 3. Monthly Unit Cost of Data Transmission

Item	Mail	CATI	TDE/ FAX[1]	EDI	Web
Phone Charges	$0.00	$0.88	$0.49	$0.28	$0.00
Postage	0.79	0.23	0.00	0.00	0.00
Labor	0.29	1.10	0.04	0.01	0.00
Total	**1.08**	**2.21**	**0.53**	**0.29**	**0.00**

[1] FAX message is used to prompt delinquent respondents.

4.1 EDI Costs

EDI costs are not easily estimated. For the BLS, EDI will incur new costs for hardware, software, systems maintenance, labor, organizational changes, and high value-added technical tasks. Currently, the largest cost component of EDI is labor. Unlike the other automated collection methods where mass respondent conversion is utilized, EDI requires that each individual respondent be converted one at the time. The EDI methodology demands more from the respondents in the initial stages. The respondents must first develop a data extraction program for the data items as defined by the BLS, create a file in the specified format, and transmit the file to the BLS. This cost can be reduced by clear and well-documented data definitions, file format and telecommunications specifications and should be amortized over the life of the reporting relationship. Before the respondents begin this work, the BLS must first convince them that once the programming is done, the reporting burden will be significantly reduced. To minimize the burden on the respondent and to achieve the BLS quality goals, it is critical that this programming be done correctly the first time.

4.2 Web Costs

Over the decades we have invested large sums of money to develop and refine the labor-intensive operations which help ensure the quality of our estimates. These operations include: collection and collection control, multiple modes of nonresponse follow-up, key entry with verification, and editing with reconciliation of edit failures. However, under Web reporting, all collection activities can be fully automated and centralized using a dedicated LAN system. Messages are electronically sent at

predetermined dates and information checked-in on a flow basis. On-line edits are implemented as part of the Web data collection session.

The cost-effectiveness of Web collection is difficult to fully measure at this time; however, enough is understood about the economics of software to come to some general conclusions. Most analysts point to the fact that software has a high fixed cost for development and very low marginal costs for adding a new user[2]. In contrast, conventional production assumes that producers face decreasing returns to scale. That is, it costs money to add new users of a product, and that these costs will increase to the point where it is no longer is profitable to produce output. Software's increasing returns to scale can be applied to applications using the Web.

For organizations purchasing unlimited Web access, the average cost of a session should approach zero, as the constraint on Web usage would be the capacity of a telephone line, renting for a fixed charge. Telephone line rentals could be spread over a large number of users, minimizing data transmission unit costs. Under other collection methods, efforts are always made to keep respondents' costs to a minimum by providing pre-paid postage, or toll-free telephone service. Using a TDE system, the respondents call a toll-free number to gain access to the system. The technology also exists for providing "800" number access to the Internet.

5. CONCLUSIONS

Both EDI and the Web are promising collection vehicles for establishment surveys. For large, multi-establishment firms, EDI is a viable data collection method. The current trends of businesses seeking streamlined operations, lower costs and increase competitiveness coupled with the survey agencies' efforts to reduce collection costs will steer both parties towards less labor intensive and burdensome approaches such as EDI.

Although still in the test stages, preliminary observations from the Web collection indicate that the combination of the TDE self reporting messaging methodology with the Web graphical interface offers a powerful, promising tool for high quality, low cost data collection. New technology is constantly driving methodological research to improve the timeliness, accuracy and relevance of our survey products. These efforts continue to push surveys toward greater reliance on technology as a key factor in improving quality and reducing costs and respondent burden.

REFERENCES

[1] Ambler, Carole A., S. M. Hyman, T. L. Mesenbourg. Electronic Data Interchange, International Conference on Establishment Surveys, Buffalo, New York, 1993.

[2] Anderson, Christopher. A World Gone Soft: A Survey of the Software Industry, *The Economist*, May 25, 1996.

[3] Clayton, Richard L., Louis J. Harrell Jr. Developing a Cost Model of Alternative Data Collection Methods: MAIL, CATI and TDE, *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 1989, pp. 264-269.

[4] Clayton, R, George S. Werking. Using E-Mail/World Wide Web For Establishment Survey Data Collection, *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 1995.

[5] Dillman, D. A. Mail and Telephone Surveys: The Total Design Method, New York, Wiley-Interscience, 1978.

[6] Werking, George S., R. L. Clayton. Enhancing Data Quality Through the Use of Mixed Mode Collection, *Survey Methodology*, 17, No. 1, June 1991, pp. 3-14.

[7] Werking, George S. Establishment Surveys: Designing the Survey Operations of the Future, *Proceedings of the Section on Survey Research Methods,* Invited Panel on the Future of Establishment Surveys, American Statistical Association, 1994, pp. 163-169.

QUALITY OF OPTICAL READING THE CENSUS '91 IN CROATIA

By Srdan Dumičić, Central Bureau of Statistics, Croatia
and Ksenija Dumičić, Faculty of Economics, Zagreb, Croatia

ABSTRACT

This paper describes the work done by optical readers during the Census '91 data entry. The reading of approximately 7.5 million questionnaires about persons, households and farms collected for the Census '91 in Croatia was performed by 14 optical readers. For the reading quality estimation, systematic sampling with a sampling fraction of about f=0.006 was used. The estimates indicated that the optical reading quality was very high, especially for numerical characters. However, during the 6 months' reading period, the reading quality and speed were decreased by about 20 percent.

Keywords: Census data; optical reading; systematic sampling.

1. INTRODUCTION

In 1991 the population census was carried out to collect data about persons, households and farms in the Republic of Croatia as of 31 March 1991.

For this purpose, the territory of Croatia, which already consisted of municipalities and settlements, was divided into about 23 000 enumeration areas. These enumeration areas were adjusted to comprise groups of around 100 households and 200 to 300 persons. After all necessary arrangements were made, approximately 7.5 million hand-written questionnaires about persons, households and farms were collected by the Central Bureau of Statistics (CBS).

The phases of the Census '91 data processing after the collection of questionnaires were as follows: manual preparation; optical reading of data; automatic coding of 14 textual answers; coverage control; data consistency checking; automatic correction; and, finally, production of about 300 tables as basic Census results.

Some monitoring systems were defined as well, with the purpose of collecting various information on the processes as well as to produce necessary reports for the management. Some of these control functions were done automatically.

A special part of the management system was based on the sample of census data. This sample was used in optical reading and automatic coding control, see [3]. The sampling method was used for the following purposes:

- monitoring and managing the optical reading process; and
- analysing the results of the optical reading and automatic coding, comparing them with the results of the manual data entry and manual coding.

2. OPTICAL READING

To illustrate the number of questionnaires collected, if one could pile up all these questionnaires, the heap would be about 1000 meters high. The transfer of such a huge quantity of data to magnetic media is always a complex problem.

Following is some basic information on questionnaires used. There are four different types of questionnaires:

- P-0, leading questionnaires (45 000), with 10 numeric questions about enumeration area, totaling 41 characters.
- P-1, person questionnaires (4.8 million), with 49 questions. For 15 of them the answers were textual. A maximum of 682 characters per questionnaire.
- P-2, household questionnaires (1.5 million), with 55 questions. The answers were numerical only. A maximum of 146 characters per questionnaire.
- P-P, questionnaires about farms (0.7 million), with 83 questions. The answers were numerical only. A maximum of 242 characters per questionnaire.

In general, the main characteristics of optical readers are the working speed and the reliability level. The speed depends directly on the speed of drawing the paper into the machine and on the average character reading speed. There are some parameters we can influence:

- the number of questions and characters that have to be read;
- the reliability level required for every particular character;
- the quality of the hand-written characters according to desired standards;
- the number and complexity of desired checks; and,
- the operator's skill in serving the optical reader (OCR).

It is clear that the demand for a higher reliability level to be read properly ensures a better quality of reading but also slows down the reading. Those

unrecognized characters which did not reach the defined reliability level are corrected later by the operator, and these corrections can seriously slow down the whole process. It is very important to maintain the same speed of optical reading and corresponding manual correcting.

The reliability level was determined very high for all the identification data and for the first letter of each word. A lower level was required for all other numerical answers. Finally, the lowest level was required for all remaining characters in textual answers. It turned out that consistency checks slowed the process down and that neither operators who served the readers nor those who corrected unrecognized characters had sufficient knowledge to correct the errors flagged earlier. Therefore, the number of checks was reduced to a minimum.

Along with such defined conditions, the average reading speed for an average completed questionnaire of any type was evaluated. In this manner, and because of the desired time-limit, the 14 optical readers were used. Each of them was connected with a PC XT with reading software, one console terminal and one terminal for the corrections. All the data were transferred through two master PCs and via telephone lines to the IBM 4381 mainframe in the Bureau.

At the beginning, the average reading speed was about 350 questionnaires per hour. It was slowly decreased so that at the end it was about 300 questionnaires per hour. This work lasted for 6 months, in two or even three shifts, partly in war conditions, and was finished in December 1991.

3. SAMPLE

The validation of the optical reading process as well as the validation of the common effect of optical reading and automatic coding were performed on the sample of the census data containing some of P-1, P-2 and P-P questionnaires.

For data entry control, systematic sampling with a sampling fraction of about f=0.006 of questionnaires was used. The sampling rates were as follows: for the farms 1/f=55 (P-P); for the households 1/f=80 (P-2); and for the persons 1/f=150 (P-1). The various sampling rates ensure that adequate proportions of each kind of sampling units in the population to be sampled are kept.

Sampling was planned to be used to validate two things:

- deviation of the census material read by optical readers towards the manually entered material; and,
- deviation of census material read by optical readers towards the properly entered material.

Using the first type of evaluation mentioned above, the validation of optical reading against the present technology of data entry could be obtained. With the second, the real reliability of optical readers with built-in reliability levels and consistency checking could be measured.

We decided to carry out the second type of validation. Since it was based on a comparison of optically read data to those properly entered, let us describe the production of this "authentic" set of sample data. We named the set of optically read questionnaires marked for being in the sample as "reading". These questionnaires were, after optical reading, collected (daily, weekly, etc.), and manually entered. This set of data was named "entry". After saving data from the same questionnaire into both sets, the corresponding records were joined by their identification and then compared position by position.

Every dissimilarity between the characters from "entry" and corresponding characters from "reading" was signaled as "suspect" on the special report. This means that there could be some erroneous manually entered characters and thus the special group of people working on producing an "authentic" set of sample data was searching for the questionnaires reported as "suspect". It is important to note that they did not have any possibility of seeing the optically read sample. They examined (established as correct) every suspect character and, if it was necessary, corrected it to read the same as that on the paper questionnaire.

This repetitive process, comparing and correcting, was carried out until all reported "suspect" characters were the same as those on the paper questionnaire. The result of this process was that the set of properly entered data could be named an "authentic" sample.

After voluminous work on its production, different kinds of dissimilarities between "authentic" and "reading" samples were counted, for example:

- "position-level" dissimilarities (simple counting of cases where the optical reader had read one character, but at this position on the questionnaire was an other character);

- "field-level" (where the field is the set of characters comprising the answer to a certain question).

In the following section the first results of the "position-level" dissimilarity analysis are described.

4. REVIEW OF THE RESULTS

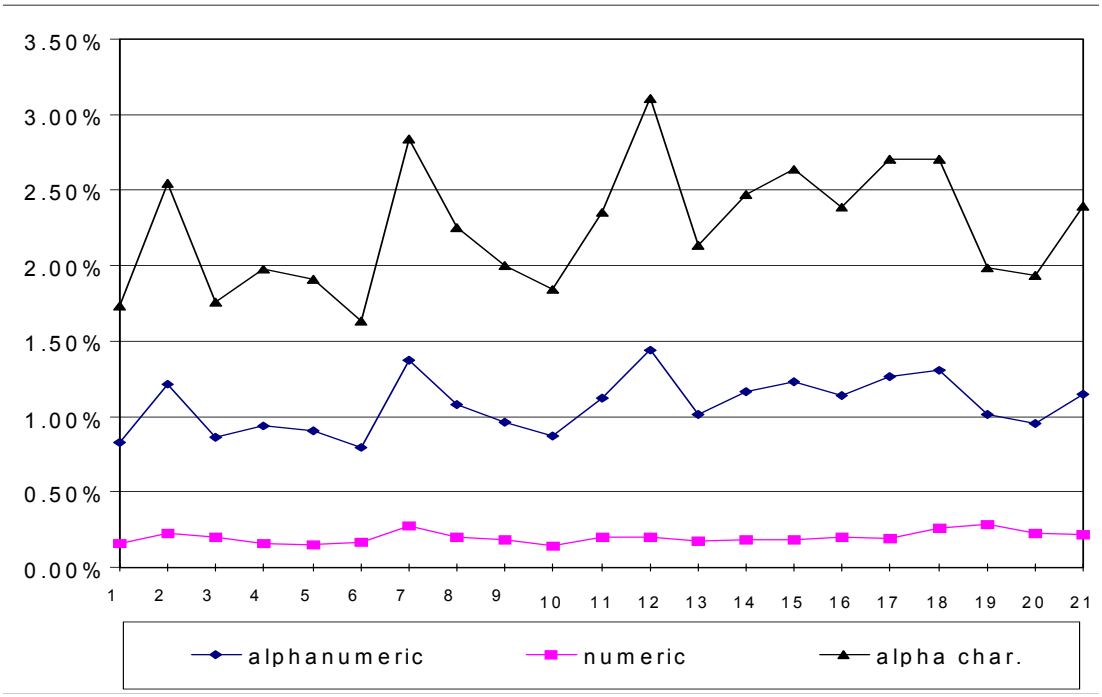
Unfortunately, we did not count the number of characters unrecognized by optical readers and recognition software. They were detected and corrected during the optical reading. The only measure we have was the number of operators which were responsible for those corrections. There was one operator for two optical readers, in other words, one operator for 700 questionnaires per hour.

The review of time-measured substitution (which is, in fact, proportional to the number of characters read out) of optical reading quality is presented in Figure 1. For this purpose the unique measure of optical reading quality was introduced. This is the total number of reading mistakes which would have been made if the whole census material had been read out with the quality from the observed period of time.

The relative participation of every character in total number of mistakes had been calculated and, afterwards, applied to the entire sample. Then the mistakes were summed up for all characters. This way, we obtained the same validation of optical reading on the mixed-character set as the one we had in our census. To clarify the view of the results, these sums were finally divided by the total number of characters, thus giving us the relative measure for the entire set.

Figure 1 displays the relative number of errors in optical reading in relation to time expressed in weeks.

Figure 1



We applied a linear regression model separately for all three groups of characters to assess the association between the relative number of errors and time. Although the curves show the trend of slow growth over time (expressed more clearly by alphanumeric and alpha characters), the linear terms in regression equations were not significant.

Furthermore, we tested whether the mean over the first 10 weeks was significantly different from the mean over the second period of time, i.e. the last 11 weeks, again separately for each group of characters. We were using two-sample test and rank-sum test. No significant differences were detected in numerical characters, while the mean for the second period of time was significantly higher in both alphanumeric (1.17 vs. 1.00) and alpha (2.04 vs. 2.44) characters, with p-values from rank-sum tests of 0.0182 and

0.0219 respectively.

This fact and the decrease of optical reading speed were probably caused by some mechanical parts wearing out. However, considering that at each of 14 optical readers more than 500 000 questionnaires had been read out, the decreasing reading quality level was not considered to be significant, and was therefore acceptable.

It could also be seen that the numerical characters were read out substantially better. This is to be expected in terms of required higher reliability level for all numerical fields. Also, since the possible values were reduced to 10 numerical characters, the substitution possibility was greatly reduced.

Blanks were read with the lowest level of errors, practically without errors. There were about 75% blanks in our sample and we excluded them from our estimates as well as from the previous graph and from table 1.

During the working period it was also observed that the optical reading quality and speed of reading were very closely connected with the quality of written characters and, of course, with operator's concentration, the latter being a bit lower during the work in the third shift.

In the table all characters classified into different groups considering optical-reading quality could be seen. All digits, except the digit 3, are in the first two groups, where the most accurately read characters are. It is significant that the characters A, E, F, L and T, which participate in the alpha part of the census material at the rate of 75%, are in the second group, with a very good reading quality. It is also remarkable that nearly 75% of all characters are read out with the substitution errors less than 2% and that in the two groups of the worst read out characters are the characters Ć, Č, Q, W, X and Y which participate at the rate of less than 1,5%.

It was to be expected that characters C, Č & Ć, D & Đ, S & Š and Z & Ž had often been substituted and that automatic coding was made to decrease this deficiency.

Table 1

% of errors	characters	total number of characters	%	cumulative
0.00 - 0.50	9,0,7,2,1,5	903 348	27.3	27.3
0.51 - 1.00	4,T,6,8,F,P,L,E,A	768 391	23.2	50.5
1.01 - 2.00	C,R,I,O,N	769 448	23.2	73.7
2.01 - 3.00	3,S,G,H,M	273 165	8.2	81.9
3.01 - 4.00	K,U,Z,Đ,V	350 333	10.6	92.5
4.01 - 5.00		0		92.5
5.01 - 6.00	Ž,Š,B,D,J	200 924	6.0	98.5
6.01 - 10.00	Ć,Č,Q	46 230	1.4	99.9
> 10.00	Y,W,X	109	0.1	100.0

5. CONCLUSION

The first estimates of the implementation of sample method on optical reading control showed that data processed by optical readers are of better quality than data entered manually. For example, only 0,25% numerical characters were read incorrectly, compared the usual 0,5% in classical data entry technology used in our Bureau.

It should be noted that the CBS used the same OCRs to read 6 million questionnaires prior to the job described above - to obtain preliminary results of the Population Census '91.

The need for improvement of optical reading has also been pointed out, e.g. the digit 3 is consistently badly read, more so than the other numerical characters. Worse reading of alpha characters was mitigated by automatic coding of textual answers because the program for automatic coding recognized

the words regardless of small differences in some letters.

The time needed to complete the entire Census data entry was reduced. Considering the lower price of optical reading and its higher speed, it could be said that the decision to use OCRs in Population Census '91 in Croatia was justifiable.

REFERENCES

[1] Dumičić, S., Kecman, Nataša, Dumičić, Ksenija. An Implementation of Sampling Method on Optical Reading Control in the Census 1991, *Proceedings of the 14th International Conference "Information Technology Interfaces" (ITI'92)*. Pula, 1992.

[2] Granquist, L. A Review of Studies on Impact of Data Editing on Estimates and Quality. *UN-ECE Joint Group on Data Editing*. Washington, 1992.

[3] Kovašević, M. Kontrola kvaliteta obuhvata popisnog materijala optićkim ćitaćem. Radni materijal, Republićki zavod za statistiku Republike Hrvatske, Zagreb,1991.

[4] Perron, S., Berthelot, J.-M., Blakeney, R.D. New Technologies in Data Collection for Business Surveys, *UN-ECE Joint Group on Data Editing,* Washington, 1992.

Chapter 6

AUTOMATED CODING

FOREWORD

by Pascal Rivière, Institute Nationale de la Statistique et des Etudes Economiques, France

If free answering is a part of statistical survey then the coding step in the survey processing is unavoidable. Coding in this connection means transforming a textual answer into a code, which belongs to a nomenclature. Unfortunately, the coding procedure is one of the most expensive phases of survey processing. That is why so many statistical institutes are attempting to introduce automatic coding tools for this purpose.

Speaking about automated coding, it is important to distinguish between the concepts of *fully automatic coding* and *computer-assisted coding*. In the statistical application very often both methods are used.

Automatic coding refers to a coding algorithm requiring no human intervention. It is implemented automatically, usually in the batch processing mode. When an algorithm of this kind operates successfully, the result is one single code.

Computer-Assisted Coding (CAC), on the other hand, comprises all techniques facilitating the coding task which is carried out case by case by a professional "coder". In particular, it provides lists of possible answers for the coder to choose from. This approach gives a better chance of finding an appropriate code than in a fully automatic coding system. CAC can also include automatic coding programmes, applied interactively.

The quality of automated coding can be expressed by the following parameters: (i) efficiency - percentage of textual answers that have been automatically coded, (ii) accuracy - calculated as percentage of coded texts where an "automatic" code is considered to be the right one, and (iii) speed - time required to code one textual answer.

The aim of this chapter is to give a necessary theoretical background to the automated coding followed by some examples of automatic coding software tools and CAC techniques implemented in the National Statistical Offices.

The Dutch contribution compares the three basic CAC methods. All of them have been developed and implemented by Statistics Netherlands. The hierarchical and the alphabetical coding techniques have already been introduced some time ago, while the trigram coding is a more recent one. The comparison is conducted on the Family Expenditure Survey.

The Swedish coding system MIKADO shows how to code multiple causes of death. This variable is well-known to be very difficult to code, because of the complexity of the nomenclature and the high variety of spellings. The paper underlines that in many cases, the answer is ambiguous and that the automatic coding system should be able to cope with this.

The contributions prepared by Austria, Canada, Croatia and France report about general software systems for automatic coding which are applied to code different variables (such as places of residence, countries, products, occupations, etc.).

The Austrian method is based on a similarity measure using overlapping N-grams. The same method is used for CAC in the preceding Dutch paper. The coding software was applied as both the tool for automatic coding and CAC. The variables on occupation, economic activity, type of education and municipality of work have been coded. It is pointed out that automatic data capture should be used before an automatic coding is applied.

The Canadian article reports about the use of the system for automatic coding by text recognition (ACTR) for the 1991 Canadian census. It shows numerous parsing methods which can be applied to standardize the text. The system permits the coding of many variables such as religion, place of residence, country, ethnic origin, major field of study, etc.

The Croatian paper concerns the application of optical reading and automatic coding for the 1991 census in Croatia. It has been used for a large scale of variables such as economic activity, nationality, occupation, etc.. The results show differences in the

difficulty to code individual variables.

The French contribution about the automatic coding software SICORE describes the method used to code numerous variables such as occupations, company names, places of vacation, cities, day-to-day activities, financial products, etc.. The clear message of this article is that, although it has a good coding algorithm, the related environment is not static at all. For example, new expressions appear, classifications change, etc.. It is therefore vitally important to update permanently a knowledge base related to the coding algorithm.

In all the presented papers it can be seen that automated methods, at the beginning, follow a similar approach. They all need a reference file and relevant parsing rules (such as, for example, replacement strings, deletion of strings, empty words, etc.). After parsing, however, the techniques differ. Some are based on similarity measures using weights that are computed for each word. Others use word recognition throughout the text, in order to simplify the following steps. Another technique uses an entropy-based tree structure which summarizes the information contained in the reference file. Many methods work with overlapping trigrams. Some are based on separate bigrams. Others do not use N-grams. Some tools examine many algorithms to increase the coding efficiency.

To sum up, there is no unique answer to the coding problem. It could be expected, however, that the examples presented in this chapter would give some ideas on how to approach different statistical applications.

OUTLINE OF A THEORY OF AUTOMATED CODING

By P. Rivière, Institute Nationale de la Statistique et des Etudes Economiques (INSEE), France

ABSTRACT

The article aims to provide a theoretical framework for automated coding, defining the coding process through its various elements. The formal definition of the coding process is given, and its component functions are explained in more detail. Three quality criteria to compare automatic coding programs: efficiency, reliability and speed, are analysed. The article also gives an overview of three different methods for the recognition of verbal responses: the simplistic, general, and specific methods.

Keywords:automatic coding, coding quality, parsing, recognition of verbal responses, coding software.

1. INTRODUCTION

In the statistical survey processing the coding phase is becoming more and more computerized. However, no generally valid methodological approach to automated coding is available. The article attempts to define the coding process through its various elements, e.g. verbal responses, classifications, supplementary variables, handling of errors, etc. It allows to introduce a more formal description of this process and to identify more systematically its components. Clearly, the most problematic component is the recognition of verbal responses. Therefore, this article is touching this issue in more detail.

2. CODING

2.1 General principles

As a first approximation, *coding* consists of the transformation of a *verbal response* or *description* into a *code*, i.e., into a pre-defined category.

This broad definition raises several questions:

1. What is the source and form of the verbal responses?
2. What is the source and the structure of the codes?
3. How to establish the correspondence between the verbal responses and codes?
4. Is additional information available that could be used in coding?

2.2 Verbal responses

In statistics, verbal responses are only needed when there are too many possible answers (i.e. codes). In

other words, when the classification is too complex.

When data is collected by the interviewer, verbal responses can be avoided wherever the classification is simple enough to be handled by the interviewer. In this case, the interviewer has a short *codebook* describing the classification (a list of correspondences between verbal responses and codes). Nonetheless, many variables such as profession (a detailed classification), economic activity, and cause of death are too complex to be handled by a codebook. The values for these variables collected in a survey therefore have to be verbal responses.

Prevailingly, the verbal responses are recorded in natural language. In some cases, the 'in-house' codes of responding units are used, which makes coding even more difficult. Sometimes, the verbal responses are subject to certain restrictions regarding the number of characters or words.

2.3 The codes

The set of codes forms a reference framework for the statistician. For certain variables, this may be a first approximation of what could be called a **classification**. There are quite a number of these: professions, activities, training, patents, products, illnesses, etc.

A classification is associated with a certain concept and thus with a *specific semantic field*. The classification's elements have to partition this field, i.e. any two elements must have an empty intersection and a complete set of the elements of the same level must cover the entire field considered.

A classification is by its very nature *structured* hierarchically in the form of a tree. The coding process should follow this hierarchical structure. Individual classification elements generally have names that correspond to a specific level in this hierarchy.

The distinction between levels is essential for coding: coding is possible even when the information supplied by a verbal response is imprecise (for example, when a profession is given as "TEACHER"). In this case the code will not be complete as it is not possible to reach the final level of the tree.

A classification is managed by a *classifier*, who is responsible for its systematic maintenance. Such changes are not made easily and usually require some form of (legal) approval. The person in charge of the classification needs to pay attention to changes in this complex area to ensure consistency as well as equivalence with other classifications in the same field.

2.4 Correspondence between verbal responses and codes

Coding requires that a verbal response be associated with a code. Therefore it is necessary for automated coding to have, in some form or other, a list of matches between verbal responses and codes.

Correspondences of this kind are normally contained in the classification, wherein, each class and each code is associated with one or more possible verbal responses, perhaps accompanied by certain comments. All this appears explicitly in an official reference document. The problem is that there are sometimes several such documents, not necessarily compatible with each other.

Assuming that one and only one reference has been validated, it is tempting to say that the correspondence between verbal response and code exists, because it is explicitly contained in the classification. However, experience has shown that this kind of reference is not enough on its own. Individuals express themselves in their own words and do not necessarily use the official jargon developed by the experts. Coding has to adjust to the material available: the replies provided in practice.

The basis for automatic coding is a large number of manually coded descriptions taken from the survey or administrative source used. The whole forms a reference file for automatic coding.

2.5 Additional information used

For many variables, the verbal response alone is not always sufficient for coding.

At most, it suffices for coding products and activities. However, it is not even sufficient for towns or villages, as knowledge of the region is needed to distinguish between towns or villages with the same name. One of the most extreme cases is 'profession', where many supplementary variables are used: the enterprise's activity, the person's qualifications, status, etc.

Consequently, the general coding framework does not consist of simply transforming a description into a code, but transforming a description *and the values of supplementary variables* into a code.

These supplementary variables resolve any ambiguity concerning the code. All such variables and their possible values must be fixed and known in advance. It is preferable that the supplementary

variables be already coded before the coding process starts. Moreover, these variables have to be capable of taking the value "missing" (not available, not known) or "N/A" (not applicable).

3. AUTOMATIC CODING

Before looking at the automatic coding process in detail, it is important to know that the choice of automatic coding for survey processing profoundly influences the way the survey is organized.

First of all, automatic coding involves **data entry**. However, coding is often done manually, directly on the survey file. Thus, the use of automatic coding involves the additional cost of entering all the verbal responses.

Moreover, automatic coding is never 100% efficient. This means that the problem **of processing uncoded descriptions** has to be solved, e.g. by using a computer assisted coding program. The absolute minimum would be a computer data entry grid containing the description to be coded, as well as any necessary supplementary variables and a location for the code.

3.1 The Principle Underlying Automatic Coding Programmes

An automatic coding algorithm has a very simple structure. The input is a verbal response (plus any supplementary variables). The output consists of a *coding response* (stating whether the coding has succeeded or failed, and specifying the type of success or failure), a *coding result* (either a code, no code or several codes), plus any other information the user would like to keep.

This can be broken down into two main steps: the **standardization** of the description and the **recognition** of the description, which is the coding itself.

Standardization is a pre-processing step for obtaining a "clean" description that is easier to understand.

This procedure uses different programs: synonymization, elimination of empty words, processing of prefixes and suffixes, phonetization, spellchecks, etc. Statistics Canada has developed an automatic coding program ACTR, which is particularly sophisticated in handling this "standardization" process.

The verbal response recognition programs generally use the classification (or any correspondence between descriptions and codes) as data. The simplest example of this is a program that only codes descriptions matching exactly what is found in the classification: this is called the *rough method*. For the remaining descriptions it has to be decided when they can be considered "close" enough to a category in the classification.

One possible technique is to code the description according to the presence of a given *key word*. A list of key words is provided in addition to the classification, and each key word is associated with a code. As soon as the program finds one of these key words, it codes the entire description regardless of any other words it may contain.

It is also possible to define *measurements of distance between texts*: any description within a sufficiently small distance from a reference description is coded. This is called the *score method*. The measurement of distance could be defined, for example, by the number of divergent characters between the two descriptions. The difficult part of this approach is how to define the thresholds: what is the maximum distance for an acceptable description? These thresholds depend on the variable to be coded and require meticulous attention. The MCA program, developed by INSEE (France), codes enterprises using this technique based on their names and addresses..

Automatic coding programs often involve several phases: first, the rough method is used first, followed by the keyword method, then a score method, and finally the keyword method using another keyword file. In this way, a quality indicator can be assigned to the result (the earlier the coding takes place in this procedure, the better the indicator). This type of multi-phase procedure was used by Croatian statisticians for their last census.

3.2 Quality Criteria

Quality criteria are needed to compare automatic coding programs. Such criteria are fairly easy to define: automatic coding can be considered "good" if it codes a large amount of descriptions accurately and quickly.

Thus the following indicators are used:

- **efficiency** = percentage of descriptions coded automatically from a given set of descriptions,
- **reliability** = percentage of correctly coded descriptions,
- **speed** = average time needed to code a description.

Obviously, these criteria are not absolute: they depend on the complexity of the source data. *For a given program,* the third criteria is probably the most stable and also the most difficult to influence.

The first two criteria can be more easily improved but they have the disadvantage of being antagonistic. Efficiency can be increased by altering the algorithm parameters (e.g. changing the thresholds), but the consequence is decreasing reliability. On the other hand, we can be more demanding regarding the closeness of the match between descriptions to be coded and official descriptions, which results in increasing reliability and decreasing efficiency.

Considering that the knowledge required for coding is fixed (e.g. classification), one of the major problems involved in automatic coding is the **trade-off between efficiency and reliability**. In practice, this means setting certain thresholds (e.g. minimum reliability) and finding the appropriate instruments to make this trade-off.

4. FORMAL DEFINITION OF THE CODING PROCESS

4.1 The domains

a) Notations

The set of possible verbal responses: L
The set of possible words: M
The set of words after parsing: M_0
The maximum number of words: K
The set of parsed (structured) verbal responses: L_0
Supplementary variables: $x_1, ..., x_n$
Set of possible values of variable x_i: P_i
Set of admissible values of variable x_i: A_i
Set of intermediate codes (before using supplementary variables): C_0
Set of codes: C

Strictly speaking, we have a vector X of n supplementary variables, belonging to a set A included in the Cartesian product $A_1 \times A_2 \times ... \times A_n$.

Design of this coding phase requires cooperation between the statistician and the data processing specialist. The statistician's knowledge of the variable context and the way the respondents answer a question are highly useful at this stage. The data processing specialist, for his part, makes known his constraints, notably as regards limitations of space.

The same applies to the ***structuring of the verbal responses***, i.e. distinguishing individual elements in the text flow. For example, in a field containing a postal address, distinction has to be made between the house number, the type of the street, the name of the street, the postal code, the region and, above all, *this structure has to be recognized in entering the code*.

The domains A_i of the supplementary variables are known to the statistician. In addition to that, they have to include a special value "*missing*" which corresponds to the value of the variable.

The set of codes C must also contain special values:

1. The value **"not known"**, corresponding to the lack of information.

   This can arise in two cases:

   - when the provided supplementary information is not sufficient for coding (e.g., "secretary", without any indication of the activity of the enterprise, or of function).
   - in specific cases, where additional variables are needed for the coding.

   "*Not known*" is not the same thing as "*missing*". The value "*not known*" concerns the codes, "*missing*" concerns the supplementary variables. If it is possible to make a partial coding (see 1.5), the value "*not known*" is not needed.

2. The value "**error**", which applies to impossible combinations of values of supplementary variables. Inconsistencies of this kind can occur between two supplementary variables, regardless of the verbal response.

4. 2 The functions

a) Notations

Decomposition into words:	$D : L \to M^K$ (purely technical)
Parsing individual words:	$N_m : M \to M_0$
Structuring a word string:	$N_s : M_0^K \to L_0$
Transformation of the supplementary variable:	$t_i : P_i \to A_i$ (generally an identity)
Verbal response recognition function:	$r : L_0 \to C_0$
Coding function:	$h : C_0 \times A_1 \times A_2 \times ... \times A_n \to C$

The ***decomposition into words*** lies in breaking down the chain of input characters into distinct parts (words). In data processing, this operation is known as *lexical analysis*. Obviously, the words concerned are not necessarily those used in normal language; they are

strings of characters that do not contain the separating character, in practice - the space.

Parsing individual words lies in "normalizing" words, for example, by eliminating superfluous characters (mainly special characters such as full stops, colons, etc.), by reducing the characters to a predetermined number, by replacing the word by a standard synonym found in the table of synonyms, by correcting the spelling if possible, or by taking the phonetic equivalent.

Transformation of the supplementary variable is indispensable when the possible values of this variable are different from those needed for the coding function. In many cases, the set of all possible values is much larger than is required for the coding function.

A particular case of this transformation is the one in which the supplementary variable *is itself a verbal response that should be coded*: for example, the verbal response to a question about economic activity is a supplementary variable for coding occupation.

Recognition of the verbal response is the most delicate phase, and is usually meant, when talking about coding. However, it is necessary to distinguish this *recognition*, in the sense of the discipline known as "pattern recognition", from coding process. In practice, as we have seen, it is often impossible to code without having the supplementary variables.

In fact, *recognition* means finding the matching category. In practice, the verbal response *and* the supplementary variables are often processed in a single stage.

The ***coding function*** produces the final code using all the available information. The input for this function consists entirely of codes: the intermediate code of the verbal response and the values of the supplementary variables.

The function can be represented by *coding rules*, taking the form:

If intermediate Code (l) = y and $x_{i1} = v_1, \ldots, x_{ip} = v_p$, *then Code (l) = c*

When such rules exist, the coding function can be formally defined, unlike the recognition of the verbal response.

In fact, it is impossible to represent the set of all the possible verbal responses: neither in intension (there being no mathematical definition of "possible verbal response"), nor in extension (the set being much too large). On the contrary, the coding rules constitute a formal description based on the fact that the sets A_i of the values of the supplementary variables are sufficiently small to be defined in extension.

4.3 Methods for recognition of the verbal response

A good coding algorithm must have three characteristics:

1) its quality must improve as the number of "learned" examples increases;
2) it must be a fast performer;
3) it must have a certain inductive capacity, since it should be able to code cases which resemble those already coded.

A. The simplistic method

This is a rough method which lies in *determining, for each verbal response, whether the same case has already been processed*, and using the same code if there is a strictly identical case.

Besides slowness, such a method is considerably rigid: if there is a slightest difference from a known case the algorithm does not code. It has no inductive capacity, nor any possibility to recognize when two strings for practical purposes identical (similar names, typing or spelling mistakes).

Of the three qualities listed above, the *simplistic* approach has only the first: the greater the number of examples, the better it performs.

B. General algorithms

A general algorithm for recognition of verbal responses is by definition an algorithm that does not depend on the variable to be coded. The variable is a parameter of the general verbal response recognition function.

Algorithms of this kind involve a two stage process: ***learning*** and ***application***.

Automated recognition of a verbal response essentially signifies *learning* the recognition function *r*, or *estimating* it.

Once the estimation has been made with the help of an estimating function $n : L_0 \; 6 \; C_0$, the coding itself is immediate, involving simply the *application* of the

function n.

As we have seen, in some cases the automated coding can fail or return an error message.

The difficulty then lies in the estimation of n, in other words the determination of the coding function. It will nevertheless be noted that the sets L_0 and C_0 are finite.

n can be constructed by using a data base of examples (*learning data base*), in other words, a set of elements of $L_0 \times C_0$.

The learning process of the automated coding function will therefore be carried out by means of:

A function G : $\mathbf{P}(L_0 \times C_0)$ 6 $\mathbf{F}(L_0 \times C_0)$

which associates a coding function with every set of examples, i.e. a function of L_0 6 C_0.

The simplistic method, though crude, is a general method of coding.

Its learning process can be described by a function G_0 applicable to any example data base:

Let X = $\{(l_1,c_1), \ldots ,(l_p,c_p)\}$ be the example base concerned.

The coding function $F_X = G_0(X)$ is defined by

$F_X(l)$ = *"not known"*, if for all (l_i , c_i)0 X , l_i ...l

$F_X(l) = c$ if there is c 0 C, such that (l , c)0 X

$F_X(l)$ = *"error"*, if there are c, c' 0 C, $c \ldots c'$, such that (l , c) 0 X and (l , c') 0 X

The "*error*" value indicates an inconsistency *in the example data base*, i.e. the example base contains two identical cases coded differently.

From a practical standpoint, *it is preferable that such errors be identified in advance,* so that the example data base can be modified and made internally consistent prior to any automated coding.

In general, drawing an analogy with the theory of statistical tests, it can be noted that automated coding is subject to two opposite kind of risks:

The first kind of risk is **over-induction**, i.e., the algorithm codes "*not known*" too infrequently. This is the risk incurred through attempts to code absolutely everything, thus over-interpreting the available information .

In this case, the proportion of coded cases will be high, but the coding quality will be low.

On the other hand,. there is the risk of **under-induction** (over-timidity), i.e. only cases identical to existing examples will be coded.

Consequently, the quality will be high, but the efficiency limited.

As in mathematical statistics, these two risks cannot be simultaneously reduced.

The described *simplistic* method corresponds to:

- a nil risk of the first kind; induction never occurs, since anything which is not strictly similar to an example already encountered is considered as "not known".
- 100% quality for those cases actually coded (assuming that there are no inconsistencies in the set of examples). Obviously, the proportion of cases coded will be small compared with a system capable of induction.

C. Specific methods

In contrast to a general algorithm, a specific algorithm operates in one stage: there is no learning phase. Instead, it is specifically associated with a certain variable and incorporates knowledge about this variable *within its programmes*.

As those methods do not use an example data base, they are not flexible and not capable to adapt.

On the other hand, the specific methods can have a fairly good second characteristic (speed) and the third (capacity for induction): working on imperfect verbal responses merely increases processing time.

INSEE uses specific coding algorithms for coding socio-professional category (the first two digits of the occupation code). These algorithms give good results, but their lack of generality is a great limitation.

REFERENCES

[1] Bruneau E. SYNAPSE, serveur de nomenclatures ,*Le courrier des statistiques* n 61-62, June 1992.

[2] Lorigny J. Questionnaire theory applied to

wording recognition", *IEEE Congress at Les Arcs*, Ed. CNRS GR23, Paris VI, 1982.

[3] Lorigny J. QUID, une méthode générale de chiffrement automatique, *Survey methodology*, vol.14, n 2, December 1988, pp. 289-298

[4] Lyberg L., Dean P. Automated coding of survey responses: an international review, *Work Session on Data Editing*, Washington, March 1992.

[5] Riviere P. The SICORE automatic coding system, Working Paper, Conference of European Statisticians, ISIS 94 Seminar, Bratislava, May 1994.

[6] Wenzowski M.J. ACTR - a generalized automatic coding system, *Survey methodology*, vol.14, n 2, December 1988, pp. 299-308.

TRIGRAM CODING IN THE FAMILY EXPENDITURE SURVEY IN STATISTICS NETHERLANDS

By Martje Roessingh , Jelke Bethlehem, Statistics Netherlands

ABSTRACT

This paper reports the use of three CAC methods: alphabetical coding, hierarchical coding and trigram coding. The last method is particularly interesting as it is a new one. The authors compare these methos by analysing the behaviour of the coders of the expenditure descrptions. The coders are free in choosing the appropriate method for each description. The article provides statistics on the chosen methods, the time needed and the success rate. It shows that the hierarchical coding is the best way to code. But trigram coding seems to be a good compromise and can be combined with hierarchical coding when the latter fails.

Keywords: CAC, hierarchical coding, alphabetical coding, trigram coding.

1. INTRODUCTION

This paper compares three computer assisted coding methods used in the Netherlands Family Expenditure Survey.

The software for computer assisted coding is Blaise. Its coding module can be used in two different ways that can be denoted by hierarchical coding and alphabetical coding.

Hierarchical coding starts by entering the first digit of the code by selecting the proper category from a menu. After entering a digit, a submenu is presented containing a subdivision of the previously selected category. So, the description becomes more and more detailed until the final digit is obtained. In the case of **alphabetical coding** a verbal description is entered, and the computer tries to locate it in an alphabetically ordered list. If the description is not found, the list is displayed, starting at the point as close as possible to the entered description. The list should be made in such a way that it contains almost all possible descriptions, including synonyms and alternative spellings.

The Blaise team felt that the usefulness and efficiency of the coding module could be improved. Research led to the development of **trigram coding**. This paper describes a test that was carried out with a special version of Blaise that contained a prototype of **trigram coding**. It was used in processing the Family Expenditure Survey. Since this special version recorded a lot of extra information, more insight could be obtained in the way trigram coding was used.

Section 2 describes the framework in which the test was carried out. It gives a short overview of trigram coding, and also presents some information about the Family Expenditure Survey. Section 3 contains an overview of the results of the analysis of the collected data. The subsequent sections go into more detail. The final section gives some conclusions.

2. TRIGRAM CODING IN THE FAMILY EXPENDITURE SURVEY

The CBS has carried out a Family Expenditure Survey since 1978. The survey collects data on income

and expenditure of households. The sample consists of approximately 2000 households. They report on income and expenditure habits by means of questionnaires and diaries. The processing of the diaries with detailed daily expenditures is a particularly costly and time-consuming activity. Since 1988, the CBS uses a Blaise CADI program to process the diaries. It uses the coding module to classify the expenditures. The coders first try the hierarchical approach to coding, and only if they do not succeed they will switch to alphabetical coding.

Trigram coding is a new approach. Trigrams are three-letter combinations. If trigram coding is applied, an entered description is split into all its subsequent three letter substrings. For example, the trigram set of the text 'bread' is {' br', 'bre', 'rea', 'ead', 'ad '}. Note that also the leading and trailing space are included. For trigram coding, Blaise splits all descriptions in the dictionary into trigram sets, and creates a special trigram index file for these sets of trigrams. After entering a description, it is split into trigrams, and then the program locates those descriptions in the dictionary that have a high percentage of trigrams in common with the entered description. Only descriptions with a fit percentage above a certain threshold value will be displayed on the screen. The coder can pick the proper description and the attached code from that list.

Trigram coding has a number of advantages:

1. It is able to cope with spelling errors. For example, if 'brown bread' is in the dictionary, and the entered description is 'bron bread', there will still be a high trigram match. So, the item 'brown bread' will be in the list on the screen.

2. Permutations in the wording of the description are not a problem. The entered text 'bread, brown' would still be linked to the dictionary item 'brown bread'.

3. If the entered text is a substring of a dictionary description, the program will also present the complete text as a possible candidate for classification. So entering 'bread' would lead to the suggestion 'brown bread', but maybe also to 'white bread'.

For more details on trigram coding that are not explained in this paper, we refer to [1].

Coding with trigrams was first implemented in a special test version of Blaise 2.4. From March 1992 through to March 1993, this version was used for processing the diaries of the Family Expenditure Survey. Each diary contains all expenditures in one week, expenditures of more than 50 guilders during holidays, the total amount spent per holiday, and expenditures of more than 25 guilders during a full year.

The article classification consists of 2289 items and the dictionary (with descriptions, different spellings, and synonyms) contains 11,221 items. Every code consists of three one digit levels and one two digits level. For example, the code for brown bread is 111.01.

The special test version of Blaise recorded data on the use of the different coding methods. Among the recorded information were: the coding method used, the keys pressed during coding, intermediate and final codes, and time needed to code one article. This way information was collected on the coding process of 146,171 articles. The data is analysed in the subsequent sections. There were 150 records not included in the analysis, because for these records coding took more than 10 minutes. This implies that the coding process was interrupted for some reason like drinking coffee or taking a small break.

3. OVERVIEW OF THE RESULTS

The coders of the expenditure descriptions had three coding methods available during the test period: hierarchical coding, alphabetical coding and trigram coding. They were completely free to choose the appropriate method for each description. Most coders had a lot of experience with hierarchical and alphabetical coding for the Family Expenditure Survey. Trigram coding was new to them, and they first needed some instructions on how to use this method. Their usual strategy is to start with hierarchical coding. Only if they do not succeed in finding the complete code that way do they change to either alphabetical or trigram coding.

A coding attempt is classified by the final coding method used. So, if a coder starts with hierarchical coding and then changes to trigram coding, it is classified as trigram coding.

A coding attempt can lead to a success or a failure. An attempt is classified as a success if a complete code is obtained, and it is classified as a failure if no code or only a partial code is obtained.

In this respect, the hierarchical coding is the absolute favourite. There are a two reasons for this. In the first place, some descriptions occur very frequently. Many coders know the corresponding codes by heart.

For example, the description 'brown bread' appeared 2,767 times during the test period. That is 3% of the cases. The corresponding code 111.01 is well known, and also easy to remember. In the second place, the coders do not need to type in the description for hierarchical coding, whereas alphabetical coding and trigram coding can only be carried out after the description has been entered. So hierarchical coding needs much less work. In a very small percentage of cases both trigram coding and alphabetical coding are used as alternatives to hierarchical coding.

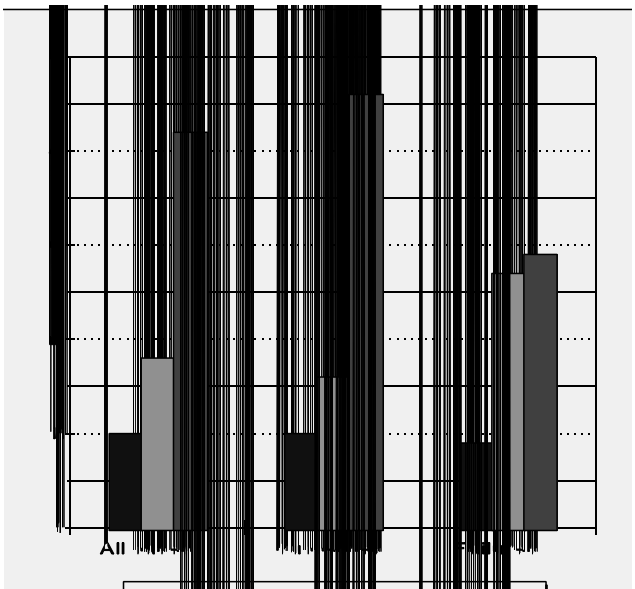
Table 1 compares the success rates of the three coding methods. Clearly, hierarchical coding is the most successful approach. This is not very surprising since this method is generally used for the easiest cases. Trigram coding is more successful than alphabetical coding. It is a much more powerful search method than alphabetical search. Again, this is no surprise as trigram coding has been designed to work in situations where simple alphabetical coding fails.

Table 1. Success rates of the three coding methods

Coding method	Percentage of attempts	Percentage of successes
Hierarchical	80 %	94 %
Trigram	18 %	85 %
Alphabetical	2 %	78 %

Success rate is only one aspect of the usefulness of a coding method. Another aspect is the time needed to find a code. Figure 1 contains a bar chart with the average time used for coding one article with one of the different methods.

Figure 1. Average time needed for coding one article



The first three bars, labelled 'All cases', relate to all cases (successes and failures). The second set of three bars, labelled 'Successes', relates to successful coding attempts only, and the last set of three bars, labelled 'Failures', denote the cases where no final code was determined.

Clearly, hierarchical coding is the fastest coding method. This can be explained by the fact that no text has to be entered, and moreover this method is used for the easy cases. Trigram coding takes more time, and also alphabetical coding is very time-consuming. However, in case of failure the time spent is approximately the same for trigram coding and alphabetical coding.

From this general overview we can draw the conclusion that hierarchical coding is the preferred method for coding expenditures. In case hierarchical coding fails, one should turn to trigram coding and not to alphabetical coding. Indeed, trigram coding is a valuable improvement in the Blaise coding module.

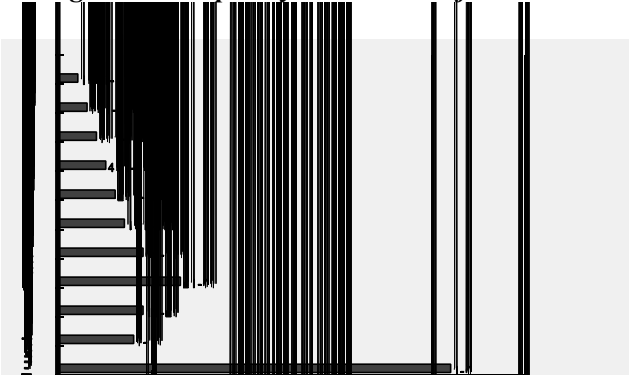
In the next sections the data about the three coding methods are analysed in more detail.

4. HIERARCHICAL CODING

Coders prefer hierarchical coding because they do not need to enter descriptions. In more than 65% of the cases (95,439 cases) they determine the final codes this way. These cases relate to 1126 different articles. When a coder does not know a code by heart and still wants to use only hierarchical coding, he has to make three choices from a list with at most 9 items and one choice from a list with at most 90 (but usually about 15) items.

Figure 2 contains a graph of the frequency distribution of the codes. The frequencies are divided into a number of classes, and for each class the length of the bar denotes the number of different codes in that category. On an average, each article appears 85 times in the file, but the frequency distribution is very skew. For example, ‘brown bread’ appears 2767 times (3% of the cases).

Figure 2. Frequency distribution of the codes



Coders learn from experience. The higher the frequency of an article, the faster they code it. This is illustrated in Figure 3. Here the length of the bar denotes the average time needed to code items in that category.

Figure 3. Average time needed for hierarchical coding

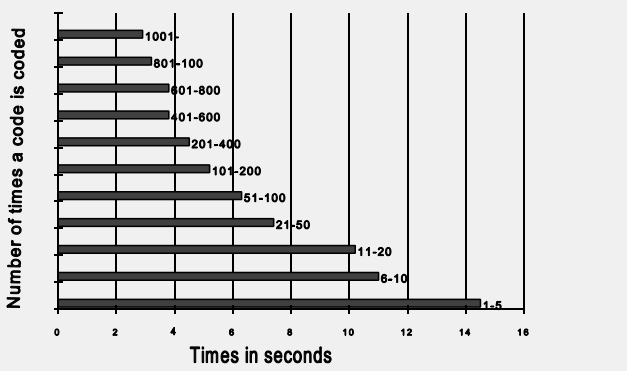


Figure 3 shows a clear correlation: the higher the frequency of an item, the less time is needed to code it hierarchically. On an average, it takes a little more than 10 seconds to code an item hierarchically. Items with a frequency of less than 5 require approximately 15 seconds, whereas items with a frequency of more than 1000 need less than 5 seconds. 'Brown bread' is coded in the shortest time: in 2 seconds. This code is also rather easy to type: 111.01.

To reduce errors, most codes are checked by a second coder. This check is carried out much faster than assignment of the code by the first coder. The reason is that the assigned code and the corresponding description are already displayed on the screen. In 98% of the cases the second coder agrees with the result, and simply presses the enter key to confirm this. This requires on average 3 seconds. In cases where only the final level was wrong, it took 8 seconds to change that. All other cases relate to more serious problems, and there the average time was 15 seconds.

In 8% of the hierarchical coding cases a text was also entered. It is not clear why the coders did that. Maybe they thought they would not find the code hierarchically and would eventually have to change to trigram or alphabetical coding. This theory is not very likely, because the average coding time (7 seconds) is shorter than in the cases where no text is entered. Perhaps some coders always enter text.

In 4% of the cases the coders try to find a code by hierarchical coding, but do not succeed in obtaining a final code. In approximately 78% of these failures there is no code at all, and in 22% there is only a partial code.

The quality of the hierarchical coding is quite high: 94% of the cases result in a final code on an average of 10 seconds. The cases in which the coder starts with hierarchical coding and switches to trigram or alphabetical coding are handled in the next sections.

5. ALPHABETICAL CODING

Alphabetical coding is the least used coding method. In Table 1 it was already mentioned that this type of coding was used in only 2% of the cases. Moreover, the success rate of alphabetical coding is relatively low (78%).

Table 2 presents a further subdivision of the cases in which alphabetical coding was used in the final stage.

Table 2. Use of alphabetical coding

Direct alphabetical coding	55 %
First hierarchical, then alphabetical	10 %
First hierarchical, then alphabetical and then hierarchical	12 %
First hierarchical, then alphabetical, no text entered	9 %
No code	12 %
Other	2 %
Total	100 %

In the majority of the alphabetically coded cases (55%), the direct approach is followed: first, the coder enters a text and then he tries to locate it in the alphabetically ordered list. In 31% of the cases the coders try hierarchical coding first, and only if they are unsuccessful do they change to alphabetical coding.

In 9% of the cases the coder starts hierarchical coding and changes to alphabetical coding without entering the text. This is not a very efficient method, because the coder has to look through a long list to find the right code. In the worst case this list contains 3,252 articles and on average more than 1,000 articles (100 screens) .

Table 3 contains the average number of seconds required to code items in the various categories of alphabetical coding.

Table 3. Average time required for alphabetical coding (in seconds)

Direct alphabetical coding	22
First hierarchical, then alphabetical	49
First hierarchical, then alphabetical, and then hierarchical	51
First hierarchical, then alphabetical, no text entered	70
No final code	12
Total	42

Direct alphabetical coding requires twice as much time as hierarchical coding. Starting with hierarchical coding and then switching to alphabetical coding at least doubles the time spent on coding an item. It also becomes clear that alphabetical coding without entering text is very inefficient.

6. CODING WITH TRIGRAMS

In 18% of the cases the coder used trigram coding to obtain a final code. The success rate of trigram coding was 85%. Trigram coding can only be activated after the descriptive text of the article has been entered. In our test the lengths of the larger part of these texts varied between 3 and 20 characters. The modal length was 10 characters.

The time needed to code these descriptions varied between 9 and 15 seconds. The average time was 11 seconds, making it slightly longer than hierarchical coding (10 seconds). Very short texts (3 characters) and very long texts (15 characters or more) required most time (14 seconds or more). Texts of 6, 7 or 8 characters required the least amount of time (less than 10 seconds).

Table 4. Trigram coding after a hierarchical coding attempt

	Percentage of cases	Average time in seconds
After coding 1 level	12 %	36
After coding 2 levels	26 %	27
After coding 3 levels	57 %	26
After coding 4 levels	5%	29
Total	100 %	29

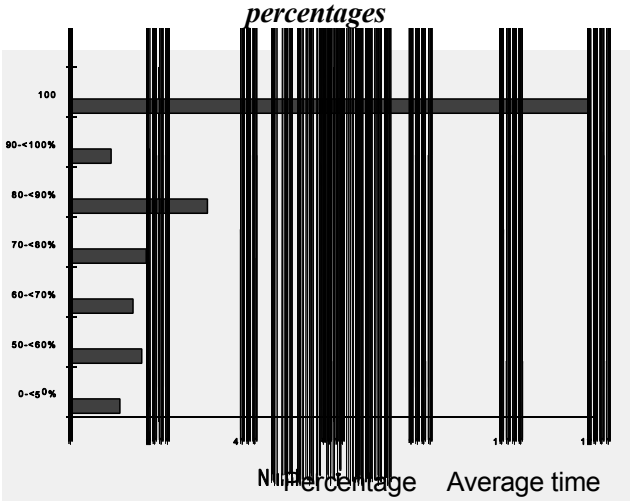
In 2% of the cases, the coder first attempts hierarchical coding: if he does not succeed in finding a code, he switches to trigram coding. Table 4 gives an overview of these cases.

The trigram search is also used after a partially hierarchical search (2,546 cases, 1.7% of the total). In total the coding takes then about half a minute. These must be some of the more difficult cases. Probably the coder thinks he can do it hierarchically, but at a certain point he cannot continue so he switches to trigrams. On average, this type of trigram coding takes about half a minute.

In 1% of the cases the coder used trigrams, but did not succeed in finding the right code. In a small portion of these cases the length of the descriptive text was less than three characters, in which case the trigram method does not work.

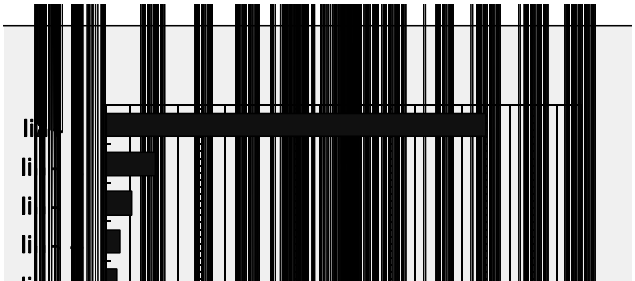
The trigram algorithm works in such a way that only descriptions and codes which have a high matching percentage compared with the entered description are displayed on the screen. Furthermore, the candidates are ordered in decreasing order of matching percentage. To see how well this works we recorded two quantities for items that were successfully coded with trigrams: the matching percentage for the selected description and the line number of the proper code. Figure 4 contains the frequency distribution of the matching percentages.

Figure 4. Frequency distribution of the matching percentages



It turns out that the matching percentages are rather high. For more than 70% of the cases the matching percentage lies between 80% and 100%. Figure 5 gives the distribution of the line numbers on which the correct item appeared.

Figure 5. Frequency distribution of the line numbers



In almost 97% of the cases the correct item was among the first ten lines, and in 75% of the cases it appeared on the first line. This is an indication that trigram coding works well, and is also very simple to use. In more than 75% of the cases the coder only has to read the first line and press enter to select the right code. Still, it takes on average 13 seconds to make this selection on the first line. Of course, this includes typing in the text.

7. CONCLUSIONS

Blaise version 2.5 offers three ways to code verbal responses: hierarchical coding, trigram coding and alphabetical coding. In our test with coding articles for the expenditure survey, hierarchical coding turned out to be the best method: it is easy to carry out and also requires relatively little time. However, one should observe that the coders are very experienced and know a lot of codes by heart. So this result is not very surprising. It should also be taken into account that hierarchical coding can only be successful if a good classification of items is available. It often happens that a coder gets stuck in the middle of the process and has to switch to a different coding method.

Alphabetical coding is not much used in this experiment. It has the disadvantage that the descriptive text has to be entered, making it more time-consuming. Also there is no guarantee that this text will be encountered in the alphabetically ordered list. To make this method useful, the list has to contain alternative spellings, permutations of words and synonyms. This will make the list very long and thus difficult to maintain. It is also possible to use alphabetical coding without entering the text. This means that the coder must page through a very long list, which takes a lot of time, and reduces the chance of locating the required item.

Trigram coding turns out to be a very attractive compromise. Although text also has to be entered, the method seems to lead to results in time that is not much longer than that of hierarchical coding. Of course, the success of trigram coding also depends on the quality of the list with descriptions. Fortunately, this list need not be as long as the alphabetically ordered list, and therefore makes maintenance a lot easier. A final point in favour of trigram coding is that it clearly turned out to be useful although the coders had no experience with it at all.

We may conclude that trigram coding can be a valuable tool for coding verbal responses. However, more research may be necessary to see how this method works in other situations.

REFERENCE

[1] Lina, M. Blaise 2.5 / Interactive Coding. Netherlands Central Bureau of Statistics, Voorburg, The Netherlands, 1993.

THE 1991 CANADIAN CENSUS OF POPULATION EXPERIENCE WITH AUTOMATED CODING

By Jocelyn Y. Tourigny and Joanne Moloney, Statistics Canada

ABSTRACT

Automation can improve the quality of coding and save resources. The paper details the 1991 Canadian Census of Population experience with a coding software developed by Statistics Canada (ACTR). The census automated coding is justified and the benefits are explored.

Keywords: automated coding; coding software; computer assisted manual coding; parsing; matching; census.

1. INTRODUCTION

The 1991 Canadian Census of Population completed the automated coding of 10 questions with write-in responses using a software called ACTR (Automated Coding by Text Recognition). In this paper we discuss the problems of coding in a census environment and the advantages of an automated coding system. We review the development of the Census coding application and ACTR.

2. JUSTIFICATION OF AUTOMATED CODING

In the context of a survey, questions requiring written responses are useful when the studied characteristic has a large set of possible response categories or when some of the outcomes cannot be predicted. Written responses allow the survey taker:

- to simplify the formulation of the question by offering the respondent fewer multiple choice questions;
- to be more objective by reducing or eliminating the artificial structure of the multiple choices proposed (and the order of the choices) thereby countering the respondents' tendency to check only the first relevant choice;
- to obtain a variety of responses that can lead to a re-examination of the classification structure and, when necessary, its modification; and
- to simplify the respondents' task because their responses are in the same medium as the question.

In order to facilitate statistical tabulations and analysis, it is necessary to group the written responses semantically using a structured classification system. This operation is called coding.

Traditionally, coding is a manual operation. Using the written response (and possibly other information provided by the respondent), and coding instructions, a coder searches for the response or an approximate alternative in the corresponding classification manual or reference material. The associated code is entered on the questionnaire. This code is then captured and used for subsequent tabulation and analysis.

Organizing manual coding of census results always rises many problems related to the specific requirements for personnel, difficulties to ensure quality and timeliness, and to integrate the coding operation into census process.

Because of that, alternatives to manual coding were sought. Automated coding was selected because of its potential to reduce the dependency on coding staff and reduce overall cost to some extent. Improvements in the quality of results arise from the predictability and consistency of computer systems.

Statistics Canada has developed an automated coding system that can meet the needs of various surveys. This generalized system, known as ACTR, is used in several surveys, the largest of which is the 1991 Census of Population.

3. AUTOMATED CODING METHODOLOGY (ACTR VERSION 1.06)

3.1 General

The methods used by the ACTR system are based in part on methods that were originally developed at the US Bureau of the Census [4] and in part on the experience of Statistics Canada in developing matching algorithms and systems for administrative files processing. The response to be coded is compared to a series of pre-coded responses, called a reference file. If a match is detected the corresponding code is recorded and the operation is complete. If not, the search continues, and an algorithm is introduced to locate the most comparable response. Once this operation is completed, the system attributes the corresponding code.

This search is made more complex because of the fact that the human language has several ways to express the same notion. Words are not always in the right order, an important word may be missing, an extraneous word may be present, a word may be a synonym or abbreviation of an expression, or the rules of punctuation and syntax may not have been respected. ACTR addresses these problems through prior processing (called parsing) of responses as well as through its two matching techniques.

Figure 1 depicts the various modules of the ACTR system that we shall describe.

3.2 Reference file

For each question to be coded, it is necessary to create a reference file consisting of typical written responses (called **phrases**) for that question and the associated numeric code. Ideally the phrases chosen are representative of the phrases most frequently observed in a matching operation. It is recommended that the phrases be retained in their original form, with errors in spelling, grammar and syntax. This file of phrases and numeric codes is integrated into a data base serving to facilitate matching operations. The reference file is constructed using entries from standard classification manuals, phrases coded by experts from a similar survey conducted previously, or a combination of these two sources as in the case of the 1991 Census of Population.

Figure 1. ACTR system

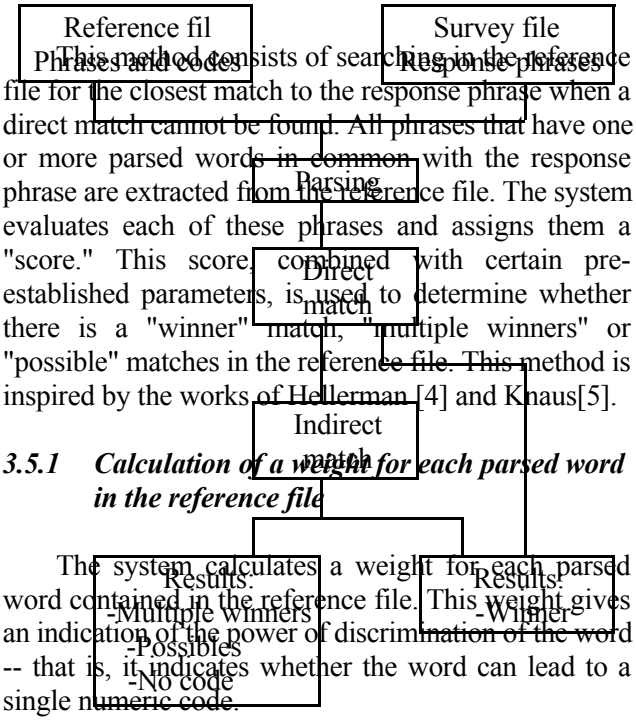
3.3 Parsing

The phrases in the reference file and those to be coded are converted in standardized form, or "parsed," in order to enable the computer to recognize, as identical, responses that are semantically equivalent. ACTR provides the user with a highly flexible parsing module. First, the phrase is considered as a continuous string of characters; it is not recognized as containing words, spaces and punctuation marks. This string of characters is analysed by the system to identify separate words. The separate words are then scrutinized and parsed; the latter stage reduces the problem of synonyms, double words, trivial words, different suffixes, etc. Annex 1 provides a list of the parsing functions offered by ACTR.

3.4 Direct matching

The parsed words of the response phrase are put in alphabetical order and the phrase is condensed to a length which averages 35% of the initial length of the phrase; the result is called the CPK (Compressed Phrase Key). This key is constructed by eliminating spaces between the parsed words and by converting individual characters (letters and numerals) and frequent combinations of characters to bit code representation. The key is then used to search for an "exact" match in the reference file, where each phrase already has its own key.

3.5 Indirect matching



This method consists of searching in the reference file for the closest match to the response phrase when a direct match cannot be found. All phrases that have one or more parsed words in common with the response phrase are extracted from the reference file. The system evaluates each of these phrases and assigns them a "score." This score, combined with certain pre-established parameters, is used to determine whether there is a "winner" match, "multiple winners" or "possible" matches in the reference file. This method is inspired by the works of Hellerman [4] and Knaus[5].

3.5.1 Calculation of a weight for each parsed word in the reference file

The system calculates a weight for each parsed word contained in the reference file. This weight gives an indication of the power of discrimination of the word -- that is, it indicates whether the word can lead to a single numeric code.

The heuristic weight of the word is constructed in such a way that it decreases as the number of codes with which it is associated increases. The weight H of a word has the following form:

$$H = \frac{E_U \& E_M \% \varepsilon}{E_M \% \varepsilon}$$

where:

$$E_M = \& \sum_{i=1}^{n} (p_i (\log_2 p_i) \quad , \quad E_U = \& \sum_{i=1}^{k} \frac{1}{k} (\log_2 (\frac{1}{k})$$

E_M is the entropy of the word. Entropy is a measure of the uniformity of a distribution. When a word is specific to a single code, the entropy is nil; it reaches its maximum when the word is associated with all items (that is the *n* codes) in the classification system.

p_i is the proportion of occurrences of the word in the files for the i^{th} code; this quantity is therefore a measure of the probability that given the word, the appropriate code is code *i*.

$$k = \sum_{i=1}^{n} x_i \quad , \quad p_i = \frac{x_i}{k} \quad , \quad \sum_{i=1}^{n} p_i = 1$$

x_i is the number of occurrences of the word in question in the phrases that have code i.

ε is an arbitrary small constant to avoid division by zero in the event that $E_M = 0$ (which corresponds to the situation where a word is specific to a single code).

$$\varepsilon ' \frac{k}{k\%1}\log_2\frac{k}{k\%1}$$

3.5.2 Calculating a score for each matched phrase

Each reference file phrase that contains at least one parsed word in common with the response phrase is considered a potential match. A scoring method was developed in order to determine the closest phrase; this score is based on the number of words contained in the response phrase that are "valid" (ie. present) in the reference file, the number of words in the reference file phrase, and the weight of the words common to the two phrases. The formula used is as follows:

$$P = \frac{(\text{number of words in common})^3 * (\Sigma \text{ weights of words in common})}{(\text{number of valid words in the response phrase}) * (\text{number of words in reference file phrase})}$$

When a response phrase matches exactly a phrase from the reference file, the formula becomes:

$$P = (\text{number of words in common}) * (\Sigma \text{ weights of words in common})$$

3.5.3 Evaluation of matches and selection of a winner

To resolve indirect matches, the user assigns values to the following three parameters:

1. MIN: lower limit of score
2. MAX: upper limit of score
3. PCNT: percentage difference

Let us assume that there are m possible matches in the reference file. The scores obtained by these phrases are arranged in descending order:

$$P_1 > P_2 > \ ... \ > P_m$$

As a result, four situations may arise:

(i) If $P_1 >= MAX$ and $(P_1 - P_2) / P_1 >= PCNT$

then the phrase that obtained score P_1 is the **winner** and its numeric code is assigned to the response phrase.

(ii) If $P_1 >= MAX$ and $(P_1 - P_2) / P_1 < PCNT$

then all phrases i such that $P_i >= MAX$ are considered as being **multiple winners**.

(iii) If $MIN <= P_1 < MAX$

then all phrases i such that $MIN <= P_i < MAX$ are considered as **possible** matches.

(iv) If $P_1 < MIN$

then no match qualifies.

All response phrases in situations (ii), (iii) or (iv), as well as those with no potential match in the reference file must be coded manually. During the tests prior to production, all such response phrases available are studied in order to improve the reference file, the parsing rules and the matching evaluation parameters.

3.6 ACTR performance

Owing to its use of the compressed phrase key, the direct matching technique is highly efficient, even when the reference file is very large.

To make indirect matching more effective, ACTR extracts from the reference file all the phrases that contain the word in the response phrase with the highest heuristic weight H, and determines their scores. Next, the word in the response phrase with the second highest weight is identified and, using this weight, a "maximum possible" score is estimated. If this score is lower than the MIN parameter (the score for a valid match) the process is halted. Otherwise extraction of reference file phrases and calculation of their scores continue.

4. 1991 CENSUS CODING APPLICATION

4.1 General

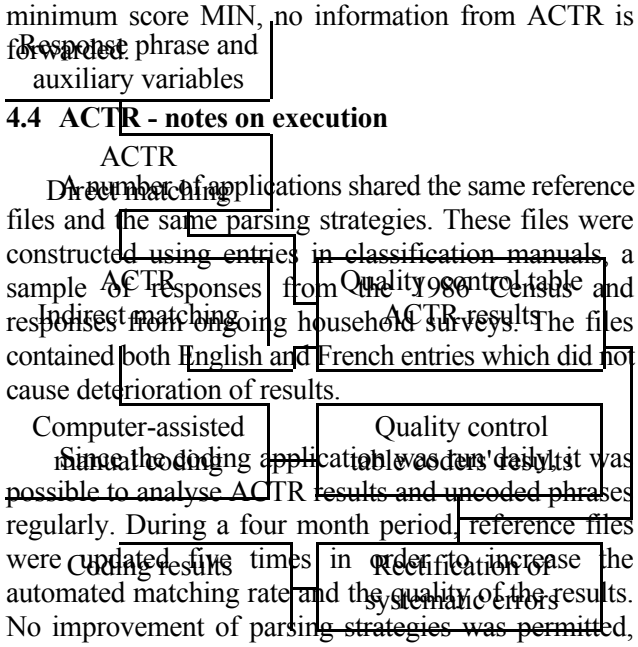
The Canadian Census of Population and Housing uses two types of self-administered questionnaires to canvass more than 10 million dwellings. When establishing the list of dwellings in his or her enumeration area, a census representative distributes a short questionnaire to 80% of the dwellings and a long questionnaire to 20% of the dwellings, following a systematic sampling. The respondent returns the completed questionnaire by mail.

The long questionnaire serves to collect information on the characteristics of individuals. The short questionnaire is an abridged version of the long questionnaire; it includes only basic questions on housing and individuals (e.g., type of dwelling, owner- or tenant-occupied dwelling, relationship to Person 1, sex, date of birth, legal marital status, common-law status and first language learned). To respond to a question, the respondent must mark a circle, write a number or write in a response.

Some write-in responses are coded manually in preparation for data capture. All the information on short and long questionnaires, except for write-in responses already coded, is captured in a single operation over a four-month period. For each variable subject to automated coding, the write-in response as well as auxiliary variables relating to the person and other occupants of the dwelling are transferred to a data base to facilitate the coding operation.

The 1991 Census automated coding application is illustrated in Figure 2. The application is highly integrated. It encompasses automated coding by ACTR, computer-assisted manual coding, quality control of the two types of coding, and rectification of systematic errors. No return to the actual questionnaire is necessary.

Figure 2. Census coding application modules



The 10 questions subjected to automated coding are shown in Appendix B. Of these, 12 similar but customized applications were established.

4.2 ACTR - direct matching

Only the response phrase without its auxiliary variables is used for automated coding. The system identifies unique response phrases. It is this unique phrase that is parsed and matched with the parsed phrases in the reference file. If there is a match, all the response phrases corresponding to the unique phrase receive the same code and the result is entered in the quality control table for ACTR results.

For the Census, the automated coding of 9 of the 10 questions was done solely by way of this matching method. Only the **Place of Residence 5 years ago (write-in Canadian cities, towns and municipalities)** question also used indirect matching to increase its automated coding matching rate.

4.3 ACTR - indirect matching

All unique phrases that are not coded via direct matching are then subjected to the indirect matching method. Information concerning the "multiple winners" and "possibles" (the matched reference file phrase, the corresponding code and the score) is forwarded to the computer-assisted manual coding. If there is no match, or if there are only matches with scores below the minimum score MIN, no information from ACTR is forwarded.

4.4 ACTR - notes on execution

A number of applications shared the same reference files and the same parsing strategies. These files were constructed using entries in classification manuals, a sample of responses from the 1980 Census and responses from ongoing household surveys. The files contained both English and French entries which did not cause deterioration of results.

Since the coding application was run daily, it was possible to analyse ACTR results and uncoded phrases regularly. During a four month period, reference files were updated five times in order to increase the automated matching rate and the quality of the results. No improvement of parsing strategies was permitted, because the impact on the quality of the results was unforeseeable.

4.5 Computer-assisted manual coding

For phrases failing automated code assignment, the computer searches the original file of response phrases (ordered alphabetically) and prepares batches of 200 uncoded phrases. Coders do not have access to the original questionnaire, but the following information appears on two screens (see Figures 3 and 4). On the first screen the coder sees the phrase to be coded,

Figure 3. FIRST SCREEN

MANUAL CODING - MAJOR FIELD OF STUDY

RESPONSE PHRASE	Type	Code
RENAISSANCE ARCHITECHTURE	_____	_____
Phrases provided by ACTR	Codes	Choice
ARCHITECHTURE	267	_____
ARCHITECTURE D'ART	048	_____
BOAT ARCHITECHTURE	308	_____

Data of each household member for the same question

Check boxes : . . . . . .
Write-in phrases : . . . . . .

PF1 = Help . . . . . PF9 = Referral

Figure 4. SECOND SCREEN

MANUAL CODING - MAJOR FIELD OF STUDY

Number of years

Elementary and secondary school :	12
University :	4
Other schools :	NONE
Education during the past 9 months :	NO
Diploma :	SEC / CERT BACC MASTER
Economic activity :	8531 TEACHING / UNIVERSITY
Occupation :	2711 TEACHER / UNIVERSITY
Major field of study :	RENAISSANCE ARCHITECHTURE
Relationship to person 1 :	PERSON 1
Date of birth :	30 / 01 / 1927
Sex :	M

the ACTR results (matched phrases and associated codes), and lastly the responses of other members of the household to the same question. On a second screen the coder can obtain the person's responses to other questions. The coder may either select one of the ACTR results, enter a code based on a classification manual or refer the case to an expert. Each time the coder selects a code, the system shows at the bottom of the screen the official rendering from the classification manual; the coder must read and confirm the code. The result of the coding is entered in the quality control table for the coder's results.

The computer electronically transfers the phrases referred to the expert on duty. The expert has on-screen access to additional information, such as the ACTR scores and auxiliary information for all other members of the household. In addition, he or she can consult more specialized reference manuals.

4.6 Quality control table for ACTR results

Quality control for automated coding has the same objectives as those of traditional coding. However, it differs in scope, since much more information on the operation is available and this information may easily be

altered.

Every aspect of quality control exploits the systematic nature of automated coding, since a phrase always receives the same code if there is no human intervention. Thus, the examination of a single occurrence of a phrase is sufficient to determine its quality. The conclusions as to its quality extend to all replicates of this phrase.

The quality control (QC) table contains one entry for each phrase-code pair. A status indicator is associated with the pair. Its value is 1 for a pre-approved pair, 2 for a pair that has been verified and found valid, 3 for a pair that has been verified and found invalid and 4 for a not verified pair. During production, each new automatically coded phrase-code pair is added to the table, and the frequency of occurrence is increased with each repeated pair.

Since the initial entries in the reference file have been intensively tested, all pairs included in this file are entered on the QC table with pre-approved status, and they are not verified. This makes the quality control more efficient.

The other pairs are sampled on a priority basis. As soon as a phrase-code pair has a frequency of three or more, one of the replicates is selected and coded by a coding clerk.

The system compares the code assigned by ACTR with the one supplied by the coding clerk. If the codes correspond, the pair is considered valid. Otherwise the case is submitted to another coding clerk. If the new code corresponds to the ACTR code, the pair is considered valid. If it corresponds to the one assigned by the first coder, the pair is considered invalid. Finally, if it does not correspond to either of the codes, the case is sent to a referral coder.

This type of quality control identifies the differences between the manually established code and the ACTR code, and it assists in detecting operational problems in the two types of coding.

In addition to facilitating sampling for quality control, the QC table serves to regularly calculate error rates. The subject matter specialist may also scrutinize phrase-code pairs not verified and determine the coding quality.

4.7 QC table for coders' results

The QC table for coders' results contains one entry for each response phrase processed. This phrase is accompanied by the code assigned by the coder, a batch number, the coder's identification number and the final code assigned after quality control.

The objectives of the quality control are to determine coder performance, identify problem areas, ensure that quality objectives are met, provide feedback on the process and prevent the recurrence of error.

The quality control method used is that of sampling by attributes, with 100% rectification of rejected batches. In practice, 5 phrases from a batch of 200 are verified by a coding clerk. As in the case of quality control of ACTR results, there is no further inspection if the codes correspond. Otherwise, it is necessary to bring in a second coding clerk and lastly a referral coder to determine the correct code.

A batch is rejected and recoded if only one phrase has an erroneous code.

The code that appears in the census file is either the code established during inspection or the original code if it has not been inspected. Error rates are calculated regularly.

4.8 Rectification of systematic errors

The two QC tables contain the history of the automated coding and the manual coding. When analysing these tables, the subject matter specialist identifies the errors to be corrected. The analysis may also lead to a change in the classification system to reflect a new reality. The census application includes a rectification module that is used at the end of production, immediately before the results are integrated into the main census processing data base.

The systematic error rectification module acts globally on erroneous phrase-code pairs and extends its action to all replicates of each pair. Detailed reports of the actions taken are produced in order to ensure proper control of this operation.

4.9 Results and observations

4.9.1 Coding volume and match rate

For the presentation of results, the responses to the 10 questions subject to automated coding were grouped under 7 variables which corresponded to separate reference files and parsing strategies. Table 1 presents these variables and some processing statistics relating to them.

Table 1. Automated Coding - variables and statistics

Variable	Processed	Matched by ACTR	ACTR rate	*CAC* Manually coded
Ethnic origin	1,160,491	1,062,015	91.51%	98,476
Language	5,998,021	5,741,294	95.72%	256,727
Registered Indian	236,501	169,675	71.74%	66,826
Place of residence 5 years ago (city/town/muni.)	1,042,951	793,425	76.08%	249,526
Major field of study	1,905,959	1,485,196	77.92%	420,763
Province - Country - Territory	880,077	821,510	93.35%	58,576
Religion	4,859,569	4,752,021	97.79%	107,548
Total	16,083,569	14,825,136	92.18%	1,258,433

Of the 16 million responses sent for automated coding, the ACTR system coded 14.8 million or 92.18% (the match rate). The remaining 1.2 million cases were resolved through computer-assisted manual coding.

The match rates fall into two main clusters: in the 71% to 78% range and in the 91% to 98% range. The disparity in match rates by variable may be explained by the volume processed, the response variation, the length of the responses, respondents' use of abbreviations, changes in national boundaries owing to the collapse of the Communist bloc and the fact that certain variables (for example, a municipality name associated with several codes) were deliberately sent for manual coding where auxiliary information could be used to obtain the correct code.

The **Registered Indian** question was new, and it was difficult to anticipate the responses, particularly since various names have recently undergone numerous changes. The **Place of residence five years ago** variable avoided the use of duplicate place names by not including them in the reference file. Duplicate place names include geographic locations that have the same name either within a province or, if the province is not identified, in more than one province. In addition, a name such as "Québec" was excluded since it could refer to either the province or the city. The **Major field of study** variable had a number of quite varied responses, diverse nomenclature, the use of abbreviation and wordy responses. The problem with wordy responses is that errors in any one word may prevent direct matching, the only type of matching allowed for this variable. In addition, it was not possible to list all possible spelling variations and abbreviations for these responses. Lastly, lengthy responses are more prone to keying error in the data capture operation.

4.9.2 Update of reference files

During production there were five reference file updates. It is estimated that they raised the match rate by 2 percentage points, which reduced the manual coding volume by approximately 25%. In some cases reference file phrases were deleted because they were found to generate errors.

4.9.3 Analysis of the QC table for ACTR results

As noted above, all unique phrase-code pairs have one of the following statuses: pre-approved, verified and found valid, verified and found invalid, and not verified.

The term "invalid" as used here indicates that there is a difference between the ACTR code and the code identified during quality control. Differences may be due to various factors: erroneous codes in the reference file, overly parsed phrases, or coders not having the latest instructions or making errors of judgment or oversight. Another cause of differences is that the response may be associated with several codes. Thus what we are measuring here is a gross difference that must be analysed before a rectification is initiated. The analyst also has the task of detecting any errors that have been missed during quality control.

Table 2 presents the volume of phrases by status. More than 87% of the phrases coded by ACTR were pre-approved. Fewer than 1% of the phrases were identified as having an invalid code.

4.9.4 Quality control resources

The resources allotted for quality control provided for verification of 3.0% of the ACTR-coded responses and 10.0% of manually coded responses. The final rates were 0.251% (Table 2: [2,705 + 34,499]/ 14,825,136) for automated coding and 10.02% for manual coding.

The rate of 0.251% can be attributed to the high frequency with which pre-approved phrase-code pairs occur and the fact that each unique pair was selected and verified only once. Such an inspection strategy is impossible in a traditional quality control operation. This rate thus indicates that using all the information produced by the system can increase the efficiency of the inspection without compromising quality.

Table 2. Results of quality control - all variables

Status	Unique Pairs	Frequency	Freq. (%)	Control (%) total
Pre-approved	14,787	12,898,773	87.01	
Verified and invalid	2,705	89,743	0.61	0.018
Verified and valid	34,499	1,735,931	11.71	0.233
Not verified	82,128	100,689	1.67	
Total coded by ACTR		14,825,136	100.00	

Table 3. Average frequency of phrase-code pairs by variable and by status

Variable / status	Pre-approved	Verified and invalid	Verified and valid	Not verified
Ethnic origin	528	12	27	1
Language	1,906	167	128	1
Registered Indian	103	13	37	1
Place of residence 5 years ago (cities/towns)	---	19	44	1
Major field of study	180	16	29	1
Province - Country - Territory	588	393	38	1
Religion	4,252	25	105	1
All variables	872	33	50	1

Table 3 illustrates, for each variable, the average frequency of occurrences of unique phrase-code pairs coded by ACTR.

The average frequency of pre-approved phrase-code pairs is 872. The most interesting frequency is that of pairs that were verified and found invalid, with an average of 33. This means that correction of one of these pairs rectifies an average of 33 errors.

The average frequency of pre-approved phrase-code pairs is 872. The most interesting frequency is that of pairs that were verified and found invalid, with an average of 33. This means that correction of one of these pairs rectifies an average of 33 errors.

For the next Census, the goal will be to pre-approve as many pairs as possible in order to minimize the resources allocated to quality control. The resources thus freed up can be used to better analyse the two quality control tables.

4.9.5 Rectification of systematic errors

Approximately 94,000 codes were corrected by the rectification module. The codes were obtained from the two types of coding (automated and manual). Most of the rectifications resulted in an improvement in quality. For the **Ethnic origin, Language** and **Province - Country - Territory** variables, several codes were changed to reflect the new world reality, which changed considerably between the production of the questionnaire and the end of processing of the Census data.

Our estimate of the final quality for the two types of coding is a combined error rate of less than 1%; manual coding is the main source of errors. However, the rate achieved is remarkably low, since in earlier censuses the error rate was in the range of 4% to 8%.

5. BENEFITS OF THIS NEW CODING PROCESS

In its first large scale use, automated coding successfully met its objective of reduced coding staff and costs, with improved data quality. Within 4 months of processing, the Automated Coding project was able to reduce the number of coders from 600 to 25 and save over $3.5 million on an estimated budget of $5.9 million, while achieving an outgoing error rate of less than 1.0% - a fraction of the error rate associated with manual coding. The dollar savings for coding takes into account the development of the census application systems, the development of reference files and parsing strategies for the automated portion and coding and training material for the computer-assisted portion; these savings were offset by a charge of $900,000 for the data capture of the write-ins.

Additional benefits are grouped under 4 headings; these are:

- Maximum control retained by the subject matter (SM) specialists

SM specialists developed the reference files and the parsing strategies using specialized tools provided by ACTR. During production, they regularly monitored the write-in responses that were not coded by the ACTR system and were able to add entries to the reference files to improve the match rate, and to remove or modify entries in the reference files to improve the quality of the results. The quality control processes provided additional feedback to the SM specialists leading to immediate corrective actions.

SM specialists directly trained the manual coders for CAC because their number was small and the operation was centralized. The computer-assisted portion of the training ensured consistency across coders. SM specialists were present during processing to

solve difficult or unanticipated cases and to quickly update coding instructions when required. This would have been impossible if manual coding was used and the SM expertise would have been lost.

The CAC system was programmed to implement certain complex procedures such as enforcing the use of expert referral for multiple responses (e.g. "Polish French" language spoken at home).

The more controlled environment permitted the SM specialist to design and implement partial action at the coding stage that could be completed at the imputation stage. An example related to the coding of the mobility variable is described in [8]. If inadequate information in the response leads to a match with duplicate place name (ie. the write-in is associated with more than one place name), a pseudo-code is assigned. The census imputation process attributes a final code at random based on the frequencies of the correctly coded places related to the duplicates.

At the end of the process, after a review of the files produced by the quality control systems and the analysis of other system-generated files, the SM specialists were able to make global enhancements to the results. This procedure can consistently correct frequent errors or implement a modification in the classification.

- More efficient quality control and certification

The fact that the automated coding system repeats its coding actions can be exploited. Only one replicate of a unique write-in phrase/code combination need be subject to quality control verification. Moreover, combinations of write-in/code found in the reference file before the production start are exempted from quality control (pre-approved) because they have been reviewed and certified accurate by SM specialists. These two measures allow savings that can be used to verify less frequent combinations.

For the CAC operation, the quality control plan is completely automated; the computer selects the sample and sends it electronically to another coder for verification. For each write-in in the sample, the computer compares the codes and determines if the original code is correct. Depending upon the number of original codes in error, the computer determines if a work load is accepted or rejected and in the latter case sends the work load to be recoded. Feedback to the coder can be provided as needed throughout processing in a timely manner.

The computer allows a quality control sampling plan for each coder and more elaborate selection scheme (such as sampling at a higher rate codes known to be prone to error). These are possibilities for our next census.

For both the automated and the CAC systems, all the quality control results are on files that can be reviewed by SM specialists. Quality control statistics can easily be tabulated and forwarded to management, other subject matter specialists and coders. Quality problems identified can be corrected globally.

At the end of the process, when reviewing the code distribution against another source of statistics, if the frequency of a code assignment is "suspect", it is easy to review all the write-in entries associated with it; the entries can be printed and analysed, diminishing the need to return to the original questionnaires.

- Better management of the coding process

Because the information is in electronic format, it is easier to predict the volume of work at each step. Regular monitoring reports can be generated from the files produced by the systems. Ad hoc analyses and reports can be easily generated.

- Potential for improvement in the next application

All the write-ins and their corresponding codes are available for other uses - other surveys or future censuses. Frequently occurring write-ins can be added to the reference file and/or the coding manual in order to ensure better coverage and coding of these responses; unused reference file entries can be removed. New parsing strategies and matching techniques can be developed and refined using as test data all or a sample of the write-ins.

Edit rules could be devised to identify write-in responses that are provided in error and should therefore not be coded, responses for which a code should be imputed, as well as responses that should be forced to the CAC operation.

Realistic 'test decks' could easily be created to train the CAC coders.

Future recoding of the data when a new classification system is introduced is a possibility. To answer a specific request, it is feasible to recode in more detail write-in responses corresponding to a certain code.

Finally, someone can exploit the capacity of the

computer to track and record paths taken by the computer or the coder for determining the code. This audit trail information can be useful for streamlining complex coding applications.

6. CONCLUSION

The use of automated coding for the 1991 Census was an outright success on which to capitalize for the 1996 Census.

Our intentions for the 1996 Census are as follows:

The ACTR software will again be used, but it will undergo certain changes to increase its versatility. It will have the option to specify the order of functions in the response parsing process, to retain the original word order when creating the compressed phrase key used in direct matching, and to retain duplicate words in parsing.

The 1991 coding applications will be moderately enhanced to make them more effective. The reference files and parsing strategies will be updated. A new module to be situated at the beginning of the application is being considered; it will decide whether a response should be subjected to automated coding, or be assigned a provisional code indicating that there is insufficient information to code it. Lastly a classification manual will be available on screen so as to facilitate manual coding.

Two new questions will be coded: Relationship to Person 1 and Place of Work (coded at the block level). For these questions, the coding application will be more complex and will utilize ACTR and other softwares to match files or edit codes (see [9]).

The challenge for the 2001 Census will be to provide automated coding for the last two questions that have write-in responses: Industry and Occupation. Ironically, the original intention when ACTR was developed was to code these two questions.

ANNEX 1

PARSING OF PHRASES

The ACTR automated coding software contains a module that allows for the parsing of phrases from the reference file and the survey file. It offers a fixed sequence of fourteen functions which, depending on the coding application, may or may not be used. The first four functions identify the words of the phrase; the other ten functions parse these words. For each function used, the subject matter specialist must provide a list of valid characters, words, replacement words or suffixes.

Processing of text:

The phrase is treated as a continuous string of characters, so as to be able to eventually identify separate words.

Function 1: exclusion clauses - For the phrases in the reference file, the text that indicates an exclusion clause (for example, "clerk (except in the armed forces)") must be excluded, since respondents do not express themselves in this manner. The result will be identical parsed phrases in the reference file that will lead to "multiple winners" matches. ACTR will not assign a code; rather, these matches will be routed to a coder who will have to decide on the appropriate code.

Function 2: deletion strings - Serves to eliminate extraneous characters, such as apostrophes, which would be interpreted as word delimiters by function 4.

Function 3: replacement strings - Serves to replace an abbreviation by one or more words, since otherwise the meaning of the abbreviation will be destroyed by function 4. For example, "T.V." is replaced by "television."

Function 4: word delineation - If a character is not in the list of valid characters for a word, this function indicates the beginning or end of a word. For example, if only numerals, letters and the hyphen are valid, the following two phrases will be divided into two words: "T.V." = T V, "English/French" = English French; the phrase "Electrician's Apprentice" will be divided into three words.

Processing of words:

The phrase is treated as a collection of words. Consequently the following functions apply to the words considered individually.

Function 5: hyphenated words - Serves to retain as a single word two words that together have a specific meaning, such as "post-secondary." If the hyphenated word is not in the list, it is split into two words; otherwise it is replaced by a new word.

Function 6: invalid word characters - If a word is made up of a character string that makes it unintelligible, it is deleted without further consideration. In some applications, this function is used to delete words

containing numeric characters.

Function 7: replacement words - This function operates in the same way as function 3. The main difference is that the search is restricted to whole words, as opposed to word parts. This function ensures that two synonymous words are recognized as being the same for matching purposes. It can also be useful for correcting common spelling errors.

Function 8: double words - If two words, when taken together in a certain order, have a particular meaning, this function serves to replace them by a single word. For example, the two words "radio" and "active" are replaced by "radioactive," and the French "garde" and "malade" are replaced by "infirmier." This function can resolve spelling inconsistencies and prevent the word order from being altered as would occur in the construction of the "compressed phrase key" in the case of a direct match.

Function 9: trivial words - an extraneous word such as an article or a pronoun does not contribute to the semantic content of the phrase and can be deleted without further consideration.

Function 10: root words - Functions 11, 12 and 13 may operate in such a way as to reduce two semantically different words to the same root. This function examines words to identify root words. If it finds one, the whole word is replaced by a substitute word, and the following three functions are not activated.

Function 11: replacement suffixes - A word is scrutinized from right to left to find the longest form of suffix listed. If such a suffix is detected, it is replaced by the approved substitute. For example, the plural marker may be eliminated so that the suffix is recognized by function 12. Thus the ending "ies" is replaced by "y".

Function 12: suffixes - Usually a suffix does not change the semantic content of a word. This function scrutinizes a word from right to left to find the longest form of suffix listed, such that once the suffix is removed, the word contains at least five characters. If a defined form of suffix is detected, it is deleted. Examples of suffixes are "able", "alist", "ian" and "er".

Function 13 - duplicate letters - the deletion of double consonants or vowels does not usually change the semantic content of the word. This deletion can eliminate spelling mistakes or data capture errors.

Function 14 - duplicate words - Only one occurrence of each parsed word is retained in the parsed phrase.

ANNEX 2

1991 Census Questions Subject to Automated Coding

First language learned

What is the language that this person **first learned** at home **in childhood** and **still understands**?

Response: if the language is other than English or French, the person specifies which one.

Note: This question appears on both the short and the long questionnaires.

Home language

What language does this person speak **most often** at home?

Response: if the language is other than English or French, the person specifies which one.

Non-official languages

What language(s), **other than English or French**, can this person speak well enough to conduct a conversation?

Response: the person may specify up to three languages.

Place of birth

Where was this person born?

Response: if born in a country other than the six countries mentioned, the person must state which country.

Ethnic origin - Ancestry

To which ethnic or cultural group(s) did this person's ancestors belong?

Response: if the person belongs to a group other than the 15 groups mentioned, he or she may specify up to two other groups.

Registered Indian

Is this person a **registered Indian** as defined by the Indian Act of Canada?

Response: if the "yes" box is marked, the person specifies the Indian band or First Nation.

Religion

What is this person's religion?

Response: the person specifies a denomination or religion or marks the "No religion" box.

Place of residence one year ago

Where did this person live **1 year ago**, that is, on June 4, 1990?

Response: if the person did not reside at an address in the same province/territory, he or she must specify the other province/territory or the name of another country.

Place of residence five years ago

Where did this person live **5 years ago**, that is, on June 4, 1986?

Response: if the person did not reside at an address in the same city or town, he or she must specify the name of the other city or town or the name of another country.

Major field of study

What was the major field of study or training of this person's **highest** degree, certificate or diploma (**excluding** secondary or high school graduation certificates)?

Response: the person indicates that the highest diploma is a secondary/ high school certificate or specifies a major field of study or training.

REFERENCES

[1] ACTR (Automated Coding by Text Recognition) Version 1.06 - User Manuals

[2] Ciok, R. The Use of Automated Coding in the 1991 Canadian Census of Population, *Paper presented* at *the 1991 Annual Meeting of the American Statistical Association*, Atlanta, Georgia, 1991.

[3] Ciok, R. The results of automated coding in the 1991 Canadian Census of Population. *Paper presented at the 1993 Annual Research Conference, organized by the US Bureau of the Census,* 1993.

[4] Hellerman, E. Overview of the Hellerman I&O Coding System. Internal document, US Bureau of the Census.

[5] Knaus, R. Pattern-based Semantic Decision Making, *in Empirical Semantics*, edited by B Rieger, Bochum, West Germany, 1981.

[6} Knaus, R. Methods and Problems in Coding Natural Language Survey Data, *Journal of Official Statistics*, Statistics Sweden, Vol 3, No. 1, 1987, pp. 45-67.

[7] Lyberg L. and Dean P. International review of approaches to automated coding. *Work session on Statistical Data Editing*. Geneva, 28-31 October 1991.

[8] Norris M.J. and Coyne S. Automated coding of Mobility Place Name data for the 1991 Census. *Symposium 91 -Spatial Issues in Statistics*, 1991, pp. 83-94.

[9] Tourigny, J., Moloney J. and Miller D. The 1991 Canadian Census of Population experience with automated coding. *Work session on statistical data editing,*, Stockholm, Sweden, 1993.

[10] Wenzowski, M.J. ACTR - A Generalized Automated Coding System. *Survey Methodology*, 14, 1988, pp. 299-307.

AUTOMATIC CODING AND TEXT PROCESSING USING N-GRAMS

By A. Haslinger, Central Statistical Office, Austria

ABSTRACT

This paper presents the first experiences made in Austrian Central Statistical Office (ÖSTAT) in developing software for automatic and computer-assisted coding of verbal responses. The method of using N-grams turned out to be quite satisfactory. Apart from the quality aspect, another goal of our experiments was to find out whether the gains in time when using automatic coding offset the time necessary for data capture, which is a prerequisite for automatic coding. The results confirm that we should use form scanners and recognition software instead of manual data capture. The article also reports about another application of the N-gram method, that is, the matching of the Austrian business register with a register of the Austrian Federal Economic Chamber.

Keywords: automatic coding; CAC; N-grams technique

1. INTRODUCTION

One of the biggest projects where ÖSTAT undertook basic research with the objective of computerizing the coding process was the Population Census.

After studying the coding projects of various Statistical Offices we arrived at the conclusion that the method of N-grams, which is used in the BLAISE coding module, is the most promising one for automatic coding as well as for interactive coding. The main advantages of this method seem to be its independency of language - it can be used in German as well as in any other language - and the fact that is does not require a sophisticated parsing strategy.

This paper describes the first steps taken in developing mainframe software for automatic coding using N-grams. The first tests were carried out with verbal responses of the 1981 and 1991 Censuses. The aim of these tests was to find out how far the coding operation in the coming 2001 Census could be accelerated and/or improved by using the computer.

Section 2 gives an overview about the method of N-grams. The tests were carried out with data from the Census. Therefore, section 3 presents some information about the Austrian Population Census. In section 4, some preliminary results of our tests are presented. Whereas all information included in the first 4 sections is the result of a research project, section 5 presents the results of a real application of the N-gram-method. The final section contains some conclusions.

2. THE METHOD OF N-GRAMS

Essentially, there are two basic approaches to the computerization of the coding process [2]. One is a dictionary method matching new responses with dictionary entries, each of which is selected from a set of actual responses. The other method uses an expert system to build a dictionary based on phrases that appear in coding manuals. ÖSTAT is using the first method. The basic features of this method are:

- a dictionary of actual responses is stored on the computer, each response equipped with the assigned code number of a classification system,
- text phrases are entered into the computer,
- text phrases are matched with dictionary descriptions and based on this procedure code numbers are assigned.

Since automated coding is a process where an input phrase is compared with the phrases which are contained in a dictionary, the question of how to measure the similarity between two phrases is of high importance. We decided to base our similarity measure on overlapping N-grams, a procedure which is implemented in the BLAISE-system in the form of trigrams [3]. An N-gram is a sequence of N successive characters of a text string. For instance, the word 'VIENNA' can be split into 5 overlapping Bigrams:

'VI', 'IE', 'EN', 'NN', 'NA'

or 4 overlapping Trigrams:

'VIE', 'IEN', 'ENN', 'NNA'.

The set containing all N-grams which occur at least once in a phrase q is called the information trace of q and will be named t(q) [4]. For the word VIENNA the information trace is {'VI', 'IE', 'EN', 'NN', 'NA'} when using bigrams, or {'VIE', 'IEN', 'ENN', 'NNA'} for trigrams. When comparing two phrases, a similarity

measure has to take into account the number of N-grams the two phrases have in common. Secondly, the length of the compared phrases has to be considered. Measuring the length |t(q)| of a phrase q by the number of elements in the information trace t(q), the similarity S between two phrases p and q can be defined as

$$S(p,q) = 1000\frac{|t(p) \cap t(q)|}{\sqrt{|t(p)| \times |t(q)|}}$$

The figure is multiplied by 1000 for the convenience of receiving a result in the interval [0, 1000]. A value of 0 stands for no common N-grams and a value of 1000 signifies that the two compared phrases are identical. We have also experimented with other similarity measures [1] but found that the above measure delivers the most satisfactory results as pairs of similar phrases have a high value and pairs of different phrases have low values.

During the process of automated coding a given phrase q is compared with all phrases p of a dictionary (or with all phrases of similar length) and the code of the dictionary phrase with the highest similarity value is assigned to q, provided that the similarity measure is above a certain threshold value. Otherwise, the input phrase has to be put in a separate file for interactive coding at a later stage. There is no generally applicable solution to the problem of a threshold value for transferring a phrase to interactive coding. The threshold depends on the item to be coded and must be determined by the subject-matter coders.

During interactive coding of not automatically codable descriptions, the information trace of an input phrase is used to find all the descriptions in the dictionary with a high similarity. These descriptions, together with their codes, are presented on the screen and the coder can select the proper description and code from the displayed list.

The advantage of the N-grams-method is the ability to cope with spelling errors. For example, if 'VIENNA' was erroneously written as 'VIENA' then 4 of the 5 bigrams, but only 2 of the 4 trigrams, would still be in agreement. This example demonstrates that bigrams are better proof against spelling errors than trigrams.

The N-gram-method is also resistant against permutations of the words in a description. The two phrases 'Technical advisor' and 'Advisor, technical' are very similar. The exact value of the similarity measure depends on the character set which is permitted for N-gram construction. If for the information trace only capital letter N-grams are used (and no special characters like 'blank' or 'comma'), the above-mentioned phrases bear a similarity of 1000.

Fortunately, the N-gram-method for N>1 is not robust against permutations of the characters of a word because the meaning of a word is influenced essentially by the sequence of the characters. Usually two words consisting of the same characters but differently ordered also have a different meaning and should not be given the same code (e.g. 'OTTO' and 'TOTO'). This characteristic can raise problems of misspellings of short words, e.g. one wrong character in a 3-digit word brings the similarity measure down to zero. This weakness can be tempered by using blanks at the beginning and at the end of the phrase and by admitting a blank as a valid character for N-grams (as in BLAISE).

In practice, the comparison of an input phrase with all or a part of the dictionary phrases can only be done with bigrams or trigrams. The advantage of bigrams is that they are more robust against errors in the input phrases at the expense of computer time. To find out the pros and cons of bigrams and trigrams was one of the reasons for developing a test version for automatic and interactive coding on IBM mainframe system. Other reasons were:

- to find out if it is possible to integrate the automatic coding of a large survey like a census into the other processing steps;
- to obtain estimates of the computer time necessary for the automatic coding of a census;
- to see how the size of the dictionary influences the time estimates; and
- to see how the efficiency and quality of automatic coding depends on errors in input data.

3. EXPERIMENTS WITH THE COMPUTER

For the experiments IBM 9121/742 mainframe system and MVS/ESA operating system were used. The interactive version of the test program runs on TSO/E using some of the facilities of ISPF/PDF. The programs were written in PLI , as this is the most frequently used programming language in ÖSTAT.

For automatic as well as for interactive coding the program starts with reading dictionary phrases into suitable structures. Then every input phrase to be coded will be compared with all, or with a subset of, the

dictionary phrases, and the similarity measure is calculated. The code of the dictionary phrase with the highest similarity will be assigned as code to the input phrase. To compare each input phrase with each dictionary phrase ('full search') would be too time-consuming for large dictionaries and/or large surveys.

Therefore it is necessary to store information in the data structures to reach areas in the dictionary where the probability of finding phrases with high similarity is high. The advantage of PLI is that it offers possibilities (e.g. pointer variables) which can be used for the restriction of the search area in the dictionary. Information about some of the tested algorithms and data structures for bigrams is given in [1]. This paper concentrates on the necessary characteristics of the dictionary to obtain high efficiency in automated coding with N-grams. Since the reported experiences were gained during the coding of some questions of a sample of the Census of 1991, a short overview of this survey is also given.

4. MANUAL CODING IN THE AUSTRIAN CENSUS OF 1991

In Austria, a census takes place every ten years; the last one was in 1991. Simultaneously with the population census, a housing census and a census of local units of employment was carried out. For each enumeration element, a separate questionnaire was used, that is, one for buildings, one for housing units, one for persons and one for local units of employment. The questionnaires were sheets in the DIN A4 format and designed to be readable by an optical reader.

The enumeration was performed by the municipalities. They were free to choose among three methods of enumeration: (1) distribution and collection of questionnaires by enumerators, (2) employment of interviewers or (3) invitation of the respondents to the town hall and subsequent completion of the questionnaires by census-clerks according to the answers of the respondents.

The filled-in questionnaires were sent to ÖSTAT where they were processed centrally. Most of the information in the questionnaires, i.e. the marked boxes and numerical fields (e.g. date of birth), was captured by an optical reading device (IBM 1288). Verbal responses were not read by the IBM 1288, the only exceptions being three questions about nationality, colloquial language and former place of residence. As for the first two questions, the most frequently occuring items were preprinted in the form of mark boxes and only the residual answers had to be written as a text. The question on the former place of residence had to be answered only by a small fraction of the population who migrated during the last 5 years between two different municipalities. This small amount of text was also read by the reading device.

The optical reader was combined with an optical video image digitizing system (OVID), which presented an image of unrecognized numbers or characters on a screen for manual correction via keyboard. This system was also used for the coding of the above mentioned 3 questions with partially textual answers.

To avoid a too long delay in publishing the census results, the processing was split into two stages. In stage 1, all the variables of the housing census (where most questions could be answered by marking the appropriate boxes) and the demographic data from the questionnaire for persons were processed. In stage 2 (1993/1994) the verbal responses of the population census (type of education, occupation, place of work, etc.) were coded manually, and the results were published.

For the manual coding of all verbal responses, about 100 person years were necessary (400 records per day and per one person, or about 50 records per hour). This productivity measure is calculated without any loss of time due to illness or holidays of the coders. So the actual number of person years of the permanent staff is much higher.

5. TESTS WITH THE AUTOMATIC CODING OF CENSUS DATA

When starting our first tests with automatic coding in the beginning of 1994, the following text files were at our disposal:

(a) the verbal responses of 500 000 persons - a sample of the Census of 1981;

(b) the coding manuals which had been used in the manual coding of the Census of 1991. Such manuals existed for the following variables: locality and municipality names (6000 phrases), education (3 000 phrases for subject of studies, type of intermediate or higher vocational schools, types of apprenticeship), occupation (short version with 18 000 descriptions and long version with 31 000 descriptions), economic activity (11 000 descriptions). Additionally there were coding manuals for some important state or municipality-run enterprises and for the municipal authorities of Vienna. All these dictionaries have been

built up by experts from the subject-matter divisions.

Because each of these coding manuals consisted of a great number of phrases and the appropriate codes of the classification system, they could be used as dictionaries for automatic coding with bigrams and trigrams. It soon became apparent, however, that the results of automatic coding with the dictionaries developed by experts in manual coding - let us call them systematic dictionaries - were not satisfactory as far as efficiency and quality are concerned. The coding procedure can be considered efficient when the percentage of input phrases automatically coded is high. In other terms, when a high percentage of the input phrases have a high similarity measure. A procedure is of high quality when the percentage of wrongly assigned codes is low.

The main reason for the low efficiency of automatic coding with regard to systematic dictionaries was that the phrases built up by an expert are often very different from the answers of the respondents of a survey. Therefore, a good dictionary should be composed of data originating from the actual language responses (natural language dictionary), e.g. it may suffice to find the term 'trade with motor vehicles' for the corresponding economic activity in a dictionary for manual coding. But the dictionary for automatic coding with N-grams should contain all synonyms, subclasses, abbreviations, etc. (such as 'car dealer',...). In the near future, all the knowledge of an experienced manual coder should be represented in the dictionary for N-gram coding.

Another disadvantage of the systematic dictionaries was that they contained some, or all, of the permutations of a phrase consisting of two or more words, e.g. 'technical advisor' and 'advisor, technical'. This is an advantage for the manual coding but not for automatic coding, as the similarity measure is nearly the same for the two terms mentioned above. Two phrases which differ only in the order of the words but have the same meaning and the same code should be included in the dictionary only once. If the used dictionary becomes too big the automatic coding is slow.

5.1 Data capture of a 1991 census subsample

In addition to the above-mentioned problems, that the variables 'occupation' and 'economic activity' (the only two open questions from the Census of 1981 which were captured) are among the variables which are difficult to code automatically. If the aim is to prove that automatic coding could be a solution to some of the problems with manual coding, it is better to start with simpler variables.

In this connection, simple means that the answers to a question can be coded independently of the answers to other questions. A typical example of such variables is a topographical description, such as place of birth or municipality. Geographical names have the advantage that the item list is well-known and standardized by convenience. It is therefore rather easy to develop a dictionary. Other examples of simple characteristics are nationality or colloquial language.

All the above-mentioned examples were included in the 1991 census, but because of manual coding the answers were not captured and stored. To obtain test data it was decided to capture a sample of approximately 150 000 person sheets of 1991. The sample was selected in 3 provinces (Styria, Tyrol and Vienna). In each province between 40 and 50 enumeration districts were selected at random, and all the verbal responses on the person sheets of the selected districts were captured (plus the district number).

For the capture of the text phrases of the sample, the boxes containing the sheets of the selected enumeration districts had to be picked out of the census archives. The data capturing staff had to search for the person sheets in the boxes and enter in the computer the verbal responses for the following variables: nationality, colloquial language, former place of residence, apprenticeship, intermediate and higher vocational schools, faculty, subject of studies, occupation, economic activity, company name and address. In terms of person sheets the capturing capacity was about 70 person sheets per working hour (manual coding 50 person sheets per hour). One of the conclusions from this project is that manual data capture and subsequent automatic coding should be performed in parallel.

An automated alternative to typing the textual answers is scanning and character recognition. Using TOSHIBA EASYREADER 1720 machines and recognition by a microprocessor driven recognition box from AEG, type 6160, between 1200 and 1400 A4-forms per hour can be scanned. To test this technology, about 100 persons were asked to fill in 5 forms of natural language phrases in handwriting (for municipality, occupation and branch of economic activity). These texts were captured manually as well as by scanning and recognition. The error rate of recognition was 17% of all characters (2/3 not recognized characters and 1/3 substitutes) but it is remarkable that 50% of all errors were concentrated in 15% of the persons who filled in the forms. At the moment, the conclusion is that the recognition of

handwritten letters has not yet been sufficiently developed. However, we hope that this technology will be improved within the next few years.

5.2 Automatic coding of the subsample

Table 1 below shows the efficiency of the automatic coding of three variables of the census file (commune of work, type of education, branch of economic activity). For the coding of the text, phrases were split into trigrams. Each of the three variables stands for a class of variables with similar difficulties with regard to automatic coding. For each variable a new dictionary was built up. To obtain a natural language dictionary for 'branch of economic activity', the 500 000 input phrases of 1981 were used. The phrases were alphabetically sorted and only those with a minimum frequency were accepted. The selected phrases were coded interactively, and the result was the new dictionary. A similar procedure is planned for 'occupation'. For other variables for which no texts of 1981 were available, e.g. 'type of education', the captured Viennese responses of 1991 were used for dictionary construction whereas the responses of the other two provinces (Tyrol and Styria) were used or will be used for testing and validation of the N-gram automatic coding procedure.

'**Place of work**' stands for geographical descriptions with a great number of well known phrases from administrative and every day life. Problems in automatic coding arise because there are some municipalities with the same or nearly the same name but different codes. Fortunately, the postcode of the commune was also asked for in the census. With this additional information a decision can be made in most cases. The number of phrases in the dictionary is much greater than the number of Austrian communes for the following reasons: firstly, the dictionary contains about 500 names of states and foreign communes (places of work for Austrians abroad). Secondly, the dictionary contains not only the official municipality names but also earlier versions of municipality names, names of dissolved communes and names of places and villages with important employers. Thirdly and most importantly, the dictionary covers abbreviations of official names.

'**Type of education**' is an example of a variable with a medium degree of difficulty for automatic coding. A series of Census questions were asked for this variable, one on every level of education: type of apprenticeship (with 100 different classification codes), type of intermediate vocational school (95 codes), type of higher vocational school (50 codes), faculty and subjects of studies (157 codes). The phrases for all 4 educational levels have been combined in one dictionary.

Both for 'place of work' and for 'education' the efficiency of automatic coding turned out to be above 90% when measured as the percentage of input phrases for which a dictionary phrase with a similarity above 800 was found. Though the value of 800 is arbitrary it proved to be a useful threshold. The exact distribution of the similarity measure for the Tyrolean sample can be seen in Figures 1 and 2. The efficiency of the newly built dictionaries is much higher than that of the coding manuals (96% compared to 85% for topographical names and 92% compared to 70% for education).

Table 1

Variable	Number of different codes	Size of dictionary	Efficiency of automatic coding
Place of work	2 400	8 000	96
Type of education	402	3 800	92
Branch of economic activity	80	7 200	50 - 80

Figure 1. Similarity for municipality of work

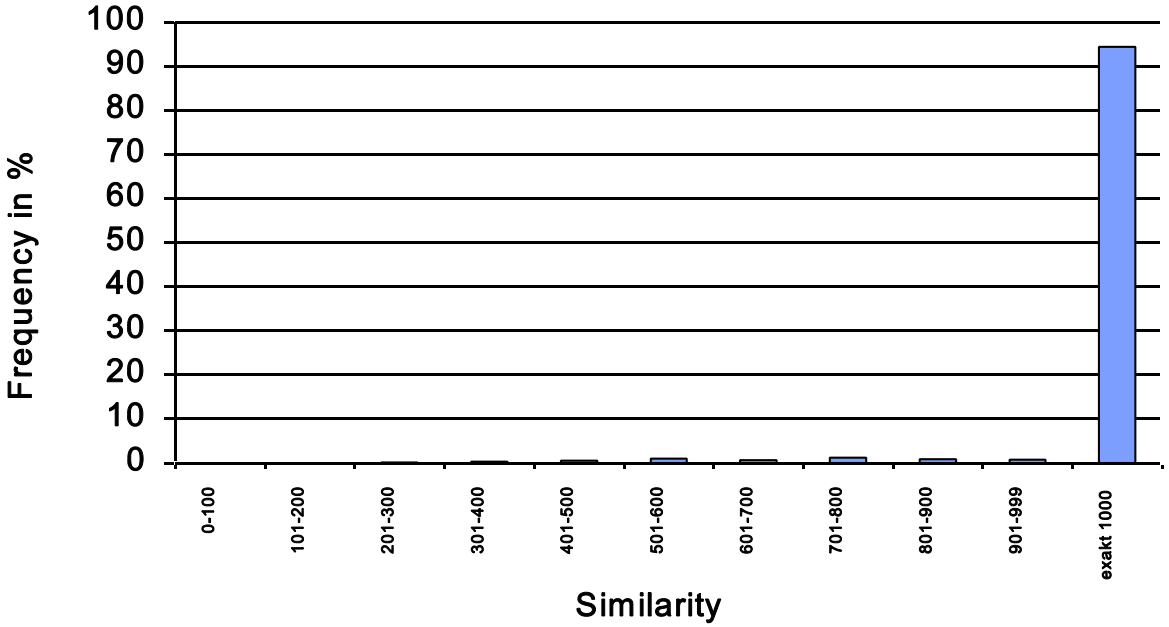


Figure 2. Similarity for type of education

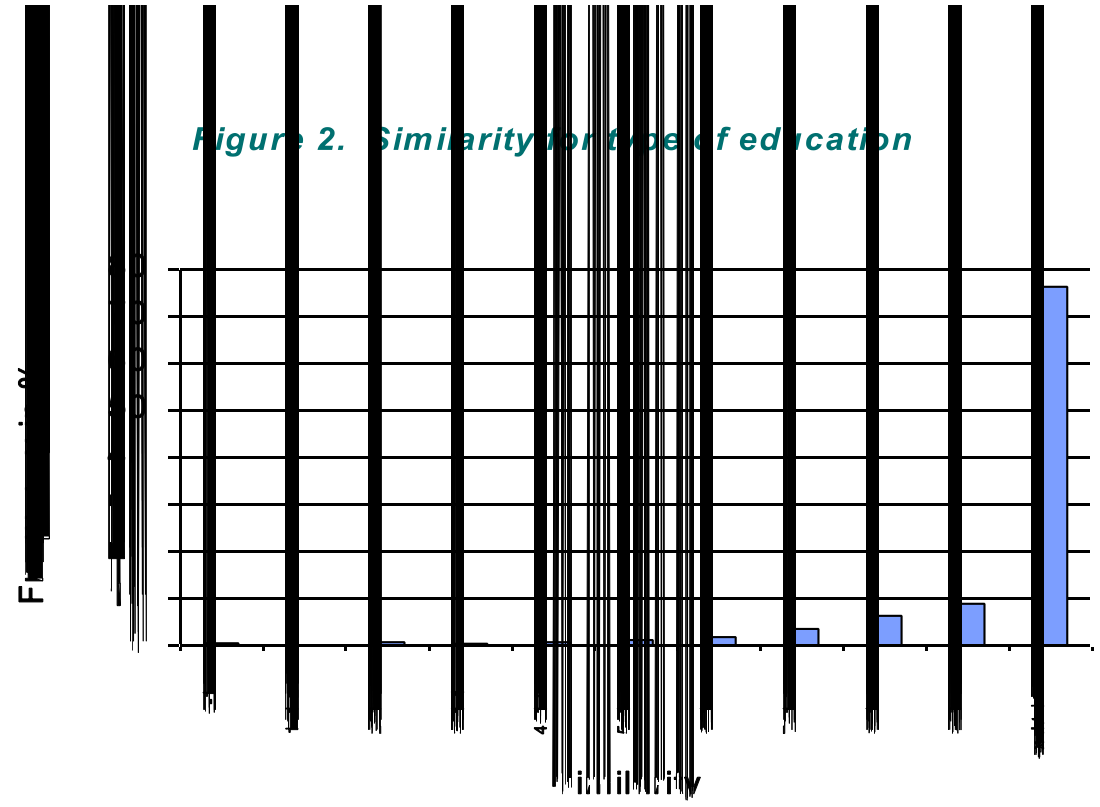
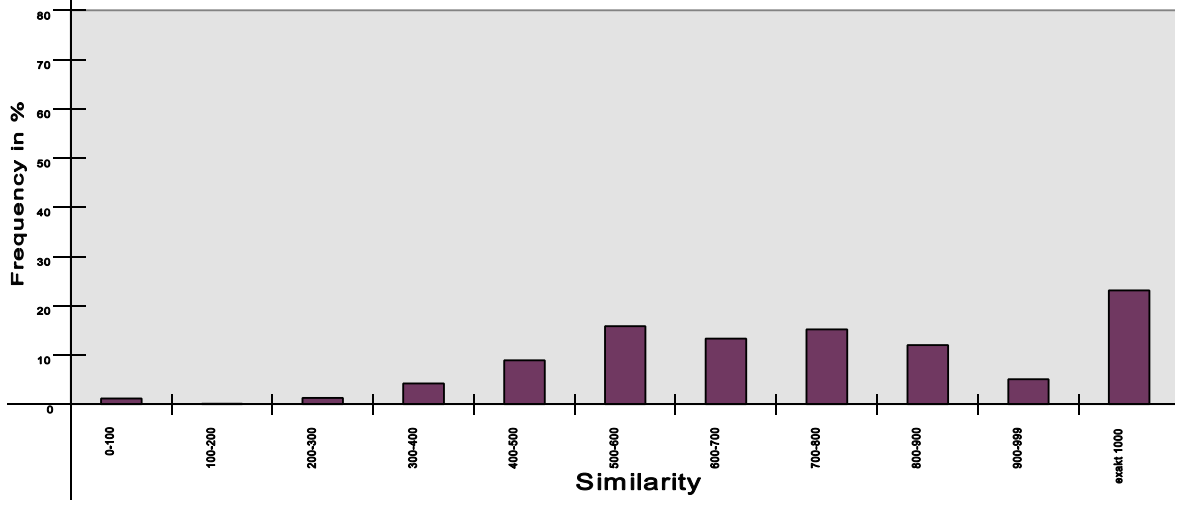
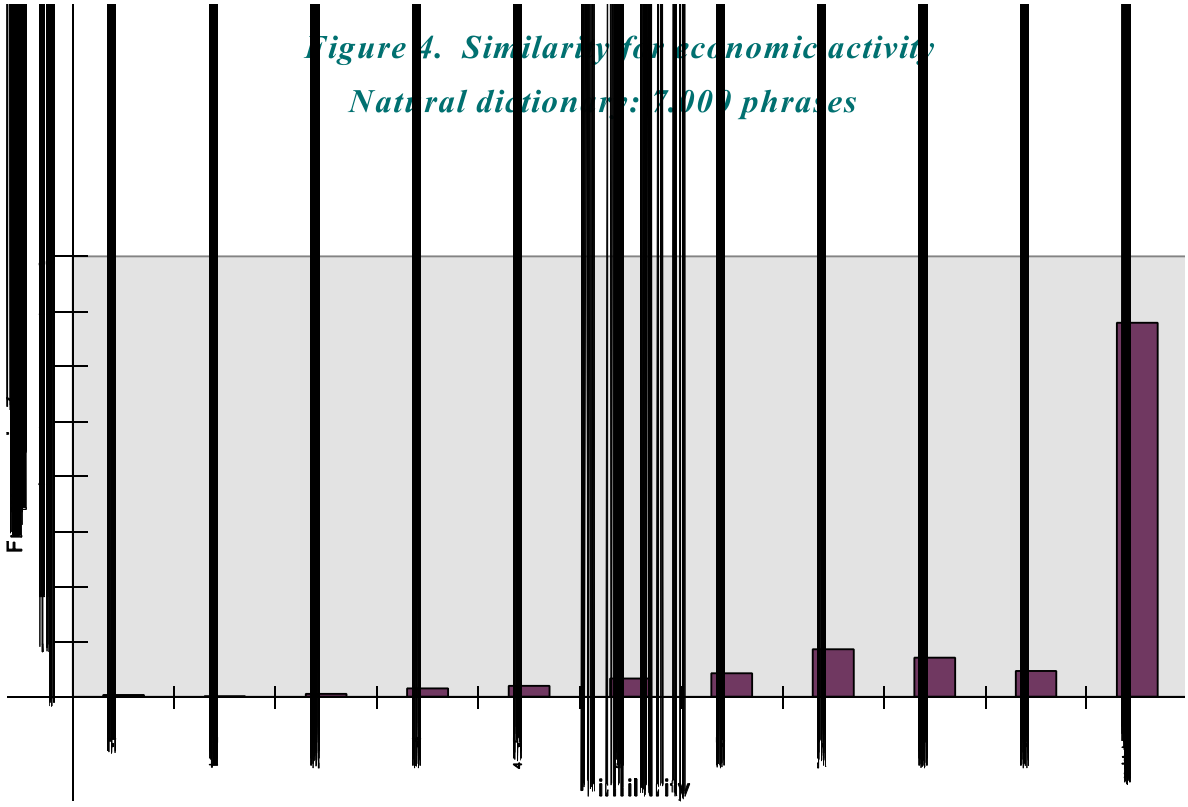


Figure 3. Similarity for economic activity
Systematic dictionary: 10.000 phrases



Variables like '**occupation**' or '**branch of economic activity**' are the most difficult ones for manual and automatic coding. Nevertheless, we have already made great progress which becomes evident when comparing Figures 3 and 4. Figure 3 shows the distribution of the similarity measure for 'branch of economic activity' using the coding manual of the 1991 Census as a dictionary for automatic coding (systematic dictionary). This manual consisted of about 10 000 phrases and produced only for about 40% of the input phrases a similarity measure above 800. In the new natural language dictionary with its mere 7 000 entries (at the moment; it is planned to complete this dictionary during the next few months), about 80% of the input phrases with a similarity above 800 could be found (see Figure 4).

Figure 4. Similarity for economic activity
Natural dictionary: 7.000 phrases



This high degree of congruity between input phrases and descriptions of the dictionary could only be achieved by the inclusion of ambiguous descriptions in the dictionary. Approximately 2000 of the 7000 entries of the dictionary are ambiguous, which means that the corresponding descriptions are associated with two or more code numbers of the classification system. In the case of 'branch of economic activity' the classification system consists of 80 different classes. 2000 descriptions of the dictionary are not specific enough for a clear-cut relation to one and only one of the 80 classes. To give an example, the classification system distinguishes between wholesale and retail trade but many colloquial responses to economic activity do not make a distinction.

To test the coding program with the census sample, about one third of the responses were attributed to an ambiguous dictionary phrase, which means not 80% but only about 50% of the responses for 'economic branch' can be automatically coded at the moment. As a next step we will attempt to supplement our programs with a module which tries to select the fitting code from a set of possible ones for ambiguous responses. This could be managed through decision rules and information from responses to other questions which are interdependent with economic activity.

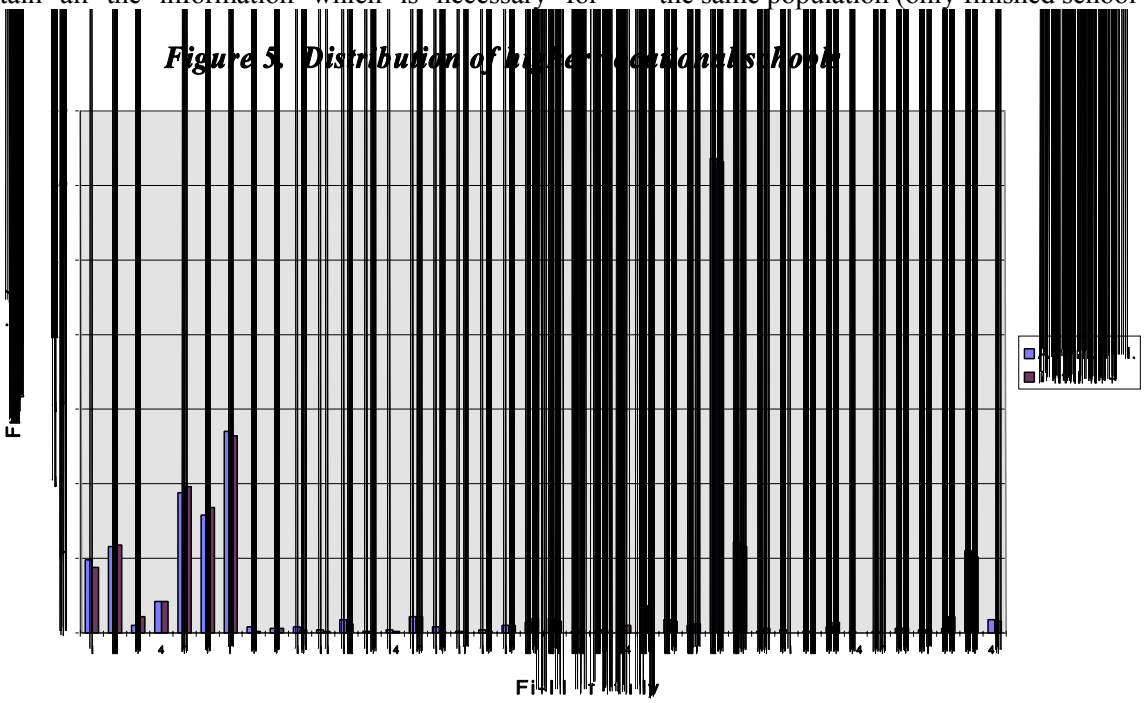
An important experience gained in the tests concerns the question of the wording of a future census. For variables like economic branch it is unlikely to obtain all the information which is necessary for (manual and automatic) coding based on one question only. Since the classification system is a multidimensional one, the question has to be split into more questions. Otherwise, the respondent would reply only to that aspect of the question which seems important to him. If the dimensions and questions are too numerous, an alternative could be to present all categories of the classification system as mark boxes on the questionnaire.

5.3 Quality of automatic coding

High quality is defined as a high proportion of the automatically coded answers bearing the correct code. The simplest way to check the quality is to link the records of the sample with the corresponding records of the authentic census file for each person and to compare the codes resulting from manual coding with those from automatic coding. If the corresponding records cannot be identified, an alternative would be the repeated manual coding of the sample. To date, only a sample of 10 000 phrases for the variable 'occupation' have been coded twice (automatically and manually).

For the variable 'type of education' we compared the frequency distributions of the sample (manually coded) with the authentic census file restricted to the same enumeration districts as in the sample. On an aggregated level there was no great difference (see Figure 5). When looking at Figure 5 one has to bear in mind that the two distributions are not based on exactly the same population (only finished school types in the

Figure 5. Distribution of highest educational school



1	Higher vocational school n.e.c.	21	Reproduction and printing
2	Construction	22	Silicate engineering
3	Wood industry	23	Land surveying
4	Chemistry	24	Other industrial and trade programmes
5	Electrical engineering	25	Tourist trade
6	Telecommunication	26	Clothing trade
7	Mechanical engineering n.e.c.	27	Arts and crafts
8	Industrial engineering	28	Higher commercial school
9	Aeronautical engineering	29	Higher school for home ecconomics
10	Metal casting	30	Higher agric. & forestry school
11	Plumbing, heating & air-conditioning	31	Agriculture
12	Motor-vehicle engineering	32	Alpine agriculture
13	Welding	33	Horticulture
14	Tools making	34	Agricultural engineering
15	Textiles engineering n.e.c.	35	Agricultural home ecconomics
16	Design, fashion design	36	Fruit & wine growing
17	Textile - commercial programme	37	Forestry
18	Knitting	38	Other higher vocational schools
19	Precision instruments engineering	39	School for teacher training
20	Plastics engineering	40	Other high school for teaching and education

census, all phrases in the sample) and that 8% of input phrases were automatically assigned to a phrase with a similarity below 800. Unlike in real application, no interactive editing of these 8% took place.

6. MATCHING TWO REGISTERS

Apart from coding, the method of N-grams can be used for other objectives too, e.g. for the matching of two registers. Until now in Austria, bigram- and trigram-coding has been used only in the above-mentioned research project but not in a real application. In contrast to coding, trigrams were used for a real production process, that is, the matching of the Austrian business register (BKT) with a register of the Austrian Federal Economic Chamber (UEV).

Each record of the BKT represents a non-agricultural Austrian establishment. The BKT served as a computerized mailing database for all surveys conducted on economic statistics from 1980 to 1994. From 1995 onwards, it will be replaced by a new register (UBR) which has been arranged according to the Council Regulation No. 2186/93 on business registers for statistical purposes within the EU. The BKT contained only enterprises (separate economic units in the cost account), whereas the new UBR is supplemented by legal and local units and the scope is enlarged by free-lance professions, e.g. physicians, lawyers, civil engineers.

The most important source of information for the updating of the BKT-register was data on new members provided by the Austrian Federal Economic Chamber every month. Every person who wants to pursue a business has to apply for a trade licence. A separate licence is necessary for each different trade in which a person wishes to engage. The UEV contains all these licences. There may be several entries in the UEV for one person. All in all, the UEV has about 350 000 records compared to 250 000 records in the BKT.

On the occasion of the switch from the old register to the new EU-conform register, it seemed useful to match the business register of ÖSTAT with the total UEV-register and to store the identification number of the UEV on the corresponding records within the BKT file. With this key, it should be possible to take over the NACE-Code or other information from the business register of ÖSTAT to the UEV.

The two files were matched by means of the establishments' name and address. For each record of the BKT, an attempt was made to find at least one corresponding record in the UEV or, in the coding terminology, with a high similarity of name and address. The UEV served as a dictionary, the name and address of the UEV-records were the dictionary descriptions. The search process for a given record of the BKT was restricted to all UEV-entries belonging to the same municipality as the BKT-record.

With the help of a batch program the BKT-register was divided into the following:

- File 1: All establishments of the BKT for which exactly one record with a similarity measure of at least 800 was found in the UEV (approximately 50% of the BKT register establishments);

- File 2: All establishments of the BKT with at least two records with a similarity above 800 in the UEV or with at least one record with a similarity between 670 and 800 in the UEV;

- File 3: All establishments of the BKT with at least one record with a similarity between 500 and 670 in the UEV.

- File 4: All remaining establishments of the BKT.

The corresponding records of file 1 were referred to as identical without further manual control. The BKT records of file 2 were stored together with all the corresponding records in the UEV. Specialists had to check these cases interactively with the help of an online application. On the screen, all the records at issue were presented with the suggested candidates for a match sorted by name and address together with additional characteristics from both files (BKT and UEV). The person in front of the screen had to mark the records belonging together. File 3 was checked manually. In the few cases where identical records were found the identification numbers of the BKT and the UEV registers were written on a sheet and captured. File 4 was regarded as containing all cases for which no counterpart could be found in the UEV.

For the interactive matching of establishments of File 2, the performance of one person was 2 200 cases per day compared to 150 cases attained by the traditional approach (looking for a certain enterprise within the total UEV with the help of a text-editor but without the use of trigrams). The use of the N-gram method has raised the performance of interactive matching by a factor of 15. Together with the fact that half of the enterprises could be matched automatically, the use of the new procedure has accelerated the whole process by a factor of 30.

7. CONCLUSIONS

The results of the experiments described above, confirm that the method of N-grams can be used successfully for automatic and computer assisted coding, provided that the following conditions are fulfilled:

- the input data are transferred to the statistical office electronically. If data are collected using paper and pencil and are captured manually before or simultaneously with automatic coding, no or only small gains in time can be expected;

- an automated alternative to the manual input of verbal responses is scanning and character recognition. Our tests showed that the recognition of handwritten letters is not yet sufficiently developed. However, we hope that this technology will be improved in the course of the next few years;

- automatic coding with N-grams requires a dictionary which contains all the descriptions or phrases in natural language. This dictionary should be based on the requests from a former survey;

- an automatic coding system cannot be expected to assign codes to 100% of the input phrases. Therefore it is necessary to have an interactive version of the N-gram coding system as well. Such a system could also be used for CAPI or CATI collection systems. According to the results of the tests of our interactive N-gram version, dictionaries with up to 50 000 phrases can be used. Such an upper limit ensures a response time of the system tolerable to the coding staff.

The savings in time and resources are only one reason for computer assisted coding. Another one is the quality aspect. The manual coding by hundreds of persons of a large survey like a census implies that the quality of the coding staff will vary from person to person. In the case of very complex variables, each coder has a slightly different classification system. Therefore, the code assigned to a given input phrase depends on the coder. This is not possible when N-gram coding is used with a dictionary. As soon as the know-how of an excellent coder is stored within the dictionary the system will always code in the same manner.

The method of N-grams can also be successfully used for construction of dictionaries, for matching of files and for searching for doubles in a file. The computer assisted match of the Austrian business register with a register of the Austrian Federal Economic Chamber saved 97% of the time a traditional procedure would have taken.

REFERENCES:

[1] Burg T., Current Status of Automated Coding in the Austrian Central Statistical Office, *EUROSTAT Workshop on Census Processing,* Fareham 1995

[2] Lyberg L., Dean P., Automatic Coding of Survey Responses: An International Review, *Work Session on Statistical Data Editing,* Washington, March 1992.

[3] Roessingh M., Bethlehem J., Trigram Coding in the Family Expenditure Survey of the CBS, *Work Session on Statistical Data Editing,* Stockholm, October 1993.

[4] Teufel T., Informationsspuren zum numerischen und graphischen Vergleich von reduzierten natürlichsprachlichen Texten, Dissertation der ETH Zürich, 1989.

AUTOMATED CODING IN THE CENSUS '91 IN CROATIA

By Srdan Dumičić, Central Bureau of Statistics, Ksenija Dumičić, Faculty of Economics, Damir Kalpić, Vedran Mornar, Faculty of Electrical Engineering and Computing, Croatia

ABSTRACT

Introducing automatic coding techniques requires writing very large software programs, and can rarely be undertaken. In the case of a census, the cost of such a project is generally lower than the cost of manual coding. Moreover, automatic coding can be easily implemented for some simple variables. Statistics Croatia introduced these techniques in the last census for numerous variables, even complex ones, like occupation. In this article, the authors describe the used algorithm, which consists of several steps. The paper reports the results of the application of this algorithm to the Census ´91 in terms of the quality and efficiency of each step.

Keywords: census data; automated coding; systematic sampling.

1. INTRODUCTION

In 1991 the population census was carried out in the Republic of Croatia to collect data about persons, households and farms. These data represent the actual situation on March 31, 1991. Approximately 7.5 million hand-written questionnaires were collected by the Central Bureau of Statistics.

After the questionnaires were collected, the following phases of the Census '91 were carried out: manual preparation; optical reading of data; automated coding of 14 textual answers; coverage control; data consistency checking; automatic correction; and, finally, production of about 300 tables as basic census' results.

Since managing such a complex process was a rather complicated task, monitoring systems were developed, aimed at collecting different information on the processes, as well as producing necessary reports to support management decisions. Some of these control functions were done automatically.

A sample of census data was used in optical reading and automated coding control (see [3]). The sampling method was used to monitor and analyse the results of both the optical reading and automated coding. These results were compared with the results of manual data entry and manual coding.

2. BASIC PRINCIPLES

There were 14 verbal responses on the census questionnaire, and these texts were later to be translated into code equivalents. Therefore, a complex application was developed (see [2]). Only a brief overview of the algorithm can be presented here.

The automated coding procedure consists of two principal steps:

- single word recognition,
- phrase recognition.

The creation of a thesaurus for each of the domains is a prerequisite for application of automated coding. Each single word appearing there becomes a candidate to be matched with the input word. The same word can be stored several times in different forms (cases, genders, tenses) due to the highly inflected Croatian language. Multiple synonymous phrases define the same code in a domain. For each thesaurus the relative word weights were calculated to represent the discrimination

power of every word. A word that appears at least once in every set of synonymous phrases would have the weight 0 because this word contributes no relevant information for deciding the code. A word that appears only in phrases related to a single code has the maximum weight. An input word is expected to match exactly a word from the thesaurus or to be similar to some of them. The reason for the absence of exact match may be that the thesaurus does not contain the word in the same case, count or gender, or that the input word was erroneously hand-written or optically read. In such a case the similarity between the input and the candidate word from thesaurus is based upon:

- difference in words' lengths,
- matching equal or similar characters,
- matching after shift,
- length of continuous matching strings.

Definition of similar characters was established by monitoring the errors due to both hand-writing and optical reading. A table of empirical a-posteriori probabilities was produced. It estimates the probability of an unexpected character in the input word being the result of an error. These probabilities are based not only on character reading but also on the language morphology and habitual syntax errors done by the considered population. For example, if in a certain word the letter È was expected and Æ appeared instead, the reason is not only the physical similarity of these letters. The a-posteriori probability is increased due to a certain ignorance of the population in proper spelling of the word.

For each input word a set of candidate words is chosen. Each of them has a certain value of its similarity to the input word and its discriminative power. Candidate phrases are introduced starting from those containing the "worthiest" words. Similarity scores for candidate phrases are calculated, also taking into account the missing words. A list of candidate phrases is produced and sorted in descending order according to the similarity score. Arbitration algorithm decides whether the code of the first ranked phrase can be accepted automatically, or a list of candidate phrases should be offered to human arbitration, or no plausible candidate phrases could be found at all. The decision depends upon the absolute value of the similarity score and the distance to the second-ranking phrase.

The whole procedure is controlled by a number of parameters that heavily influence the program performance. Description of these parameters and the applied parameter values, stated in brackets, for automated coding of the most demanding variable 'occupation', may help to obtain an insight into the algorithm.

Value for probability one (1000) is a general parameter which determines the number of significant digits the program is working with. It means that working with 3 significant digits implies the value of this parameter to be 1000. In this way all the arithmetics remains in the integer domain. The similarity measure for a perfect match is normalized to this value. If the value for this parameter is too low, it becomes difficult to resolve slight differences. If the number is too large, unnecessary computing effort is needed, accordingly processing time is wasted. Some of the other parameter values are based on the given **value for probability one**.

The first group of specific parameters determines the procedures regarding word recognition and is used in computing cumulatively the similarity between the input word and the candidate word from the thesaurus. In the short procedure only the words with the same initial character are matched. The complete procedure is invoked when the short one could not yield a satisfactory match. Parameters to calculate the similarity measure for the words and their values are:

- **Weight of an equal character (10)** is added to the similarity measure between the input word and the candidate word for each matched character. For similar characters a-posteriori probabilities multiplied by this parameter are added instead.
- **Weight of an unintelligible character (5)** is added when such a character appears in the input word.
- **Penalty for abbreviation (5)** is subtracted from the measure of similarity when an abbreviation, recognized by the terminal dot, is exploded into the full word, which would inappropriately imply a perfect match. The same parameter is used to penalize the words which did not match perfectly, but due to integer arithmetics normalization the maximum possible similarity, i.e. **value for probability one** was achieved.
- **Bonus for no shift (7)** is added to the measure of similarity if no left or right shifts were applied for word matching.
- **Minimum string length for shift (5)** determines when the attempt of match by shift should stop. It is the minimum length of overlapping strings after

shift.

- **Bonus for continuous string match (3)** is added for each contiguous character match.

- **Minimum character similarity (600 = 0.6 C value for probability one)** determines whether a similar character is similar enough to introduce the bonus for continuous string.

- **Penalty for the difference in word lengths (5)** is subtracted for each difference between word lengths in characters.

Simultaneously with calculation of the similarity measure, the maximum possible similarity measure for the given input word is computed. This maximum value is applied to calculate the normalized similarity measure which, for perfect match, can reach the **value for probability one**.

- **Threshold for entering the complete procedure (749 = 0.749 C value for probability one)** becomes active if any word, with the first character matched, could not achieve at least this similarity. In such a case all the words from the thesaurus are matched with the input word. The complete procedure is as many times slower as many letters appear in the alphabet.

After the normalized similarity measure has been computed, the following parameters determine whether the word from the thesaurus becomes a candidate word:

- **Minimum similarity (500 = 0.5 C value for probability one)** is a threshold value of the measure of similarity for a word to become a candidate. All the words below this value are rejected.

- **Measure of similarity for mandatory choice (900 = 0.9 C value for probability one)** is the threshold value above which every word becomes candidate.

- **Maximum similarity difference (300 = 0.3 C value for probability one)** rejects a candidate word if its measure of similarity is so much worse than the similarity of an already accepted word.

- **Maximum number of candidate words (20)** determines the maximum cardinality of the set of candidate words matched with a single input word. This parameter strongly affects processing speed and accuracy. With the cardinality too high and with small **minimum similarity** value, the number of operations grows heavily for phrases containing multiple words. With the value too low, the proper solution can be lost.

After the sets of candidate words have been formed, the similarity scores for the phrases from the thesaurus are computed. It is the sum of products of thesaurus word weights multiplied by their corresponding similarities to the input words. The result is divided by the sum of word weights. In the case of perfect match, the result is the **value for probability one**. If the match is not perfect, some measures of similarity are smaller than the **value for probability one**. Therefore the similarity score is also below **value for probability one**. However, it happens that the input phrase contains one or more words that cannot be found in candidate phrases selected from the thesaurus. To penalize such an input phrase, for every word lacking the denominator of the previously mentioned ratio is increased by the average thesaurus word weight divided by the parameter **candidate lacks word (6)**. The reverse situation can be simultaneously present: the phrase from the thesaurus contains one or more words that cannot be found among the input words. To penalize such a mismatch, the denominator is increased further on by the weight of every such thesaurus word divided by the parameter **input lacks word (4)**. The division is applied instead of multiplication, to maintain the integer arithmetics.

The search for candidate phrases is not exhaustive. It starts by selecting only those phrases which contain words whose measure of similarity multiplied by the word weight is higher than a parameter value **minimum similarity for quick procedure (950 = 0.9 C value for probability one)**. If in such a way no satisfactory solution can be found, a more extensive procedure is invoked. All the phrases are evaluated which contain a word with weighted similarity higher than the parameter value for the **threshold for mandatory search (900 = 0.9 C value for probability one)**.

After a set of candidate phrases has been formed and sorted in decreasing similarity order, the decision is made whether to use automatic coding by consulting a two-dimensional decision table. One dimension in the table is the similarity score and the other is the distance in similarity from the first rated candidate phrase. The higher the similarity score of the first rated candidate phrase and the more distant the second rated candidate, the higher the expectation that the first rated candidate phrase can be coded automatically. If not, the list of candidate phrases is offered for human arbitration.

Characteristics of the most representative and demanding thesaurus for the domain 'occupation' are:

number of phrases = 21644,

number of distinct codes = 2922,

number of words in phrases = 72148,

number of distinct words = 13109.

The procedure is specially designed bearing in mind a highly inflected language. The prefixes are common and most do not distort the basic word meaning. Therefore, match by shift is applied. The inflection increases towards the end of words. Thus the importance of each subsequent character *i* is lowered through division by **a**/(**b**C*i*). The parameters are **a(1)** and **b(1)**.The authors believe that the procedure could be adapted for most European languages, except for those which have agglutinative characteristics. In the latter case maybe some preprocessing to scan for elementary words could help. The parameter values have been set intuitively and occasionally changed after some testing. Some of them differ in the same language for various domains, so for other languages extensive testing and some linguistic expertise would surely be welcome.

To sum up the application supports:

- automated coding;
- computer assisted coding; and
- monitoring and managing the process.

3. DATA PROCESSING

The text was written by hand in 14 fields of the questionnaire and, once entered, the corresponding codes could belong to 8 domains. For each domain the appropriate thesaurus was made.

Thesauri were made for the following 8 attributes:

- settlement and municipality,
- school,
- nationality,
- country,
- language,
- activity,
- religion,
- occupation.

On the basis of test processing and monitoring the coding procedures, the thesauri were updated during the data processing itself.

The main **coding program** has the following functions:

- analysis and "standardization" of the written text;
- word recognition throughout the text;
- phrase recognition;
- computing the matching level between the written phrases and the thesaurus phrases, as well as computing the "distance" between the potential phrase candidates;
- automated coding of phrases for which the probability was high enough (the level of probability being prescribed), with an additional condition that the distance to the next phrase candidate is large enough;
- for other included phrases the list of potential candidates ordered by probability of being accepted is created (this part served as the entry for the application for computer assisted coding);
- updating and printing all the data that is important for monitoring and managing the processing;
- changing the processing parameters in an easy way. Some of these parameters depended on the special characteristics of Croatian language, what means that the whole application is partly dependent on the language.

The program was written in "C" language.

The application for **computer assisted coding,** to be used by trained computer staff, gives the possibility to choose between several proposed codes. In case none of the proposed choices is acceptable, or there are no proposed choices at all, the application enables to enter a new code. In this case all fields of the inquiry form are displayed on the screen. This facilitates the work of specialists who solve such problems.

The computer-assisted coding is a transaction-processing application, with programs written in PL/I, VSAM datasets, CICS as TP monitor-tailored for the IBM (4381/T91) mainframe.

The main coding program in batch processing mode was carried out on two personal computers with i860 processors. The computer-assisted coding (on-line application) was performed on about a dozen IBM terminals, with trained personnel working in two shifts. The whole procedure of automated and computer-assisted coding was performed successively, following the daily data entry, and adequate interfaces were made

for this part.

4. RESULTS

To monitor the quality of automated coding 5 types of results were analysed, i.e. 5 levels of text recognition were set. These levels are:

- M1 - exact matching ===> automated coding;
- M2 - one candidate has high matching level and great distance from the other candidates ===> automated coding;
- M3 - one or more candidates have too small a matching level or the distance between first and second candidate is too small ===> computer-assisted coding with 1 to 9 choices;
- M4 - there is no candidate with the minimal matching level to be suggested ==> computer-assisted coding with the picture of the whole questionnaire, without any suggestion;
- M5 - there is no need for coding, or, codes were already filled in.

The results are shown in Table 1.

It was agreed that the estimation of the algorithm quality and of the computer coding program should be carried out on the basis of ratio between the number of coded fields of levels M2 and M3. The level M1 was not used because its complexity is negligible. The level M4 was also exempted because it occurs only in the cases of completely absurd entries or of major shortcomings in the thesaurus design. Such mistakes are not caused by the designers of algorithms and automated coding programs.

According to the experience and the estimated reduction of human engagement, the three ratio levels of M2 and M3 were defined: fair, good and excellent. For the different groups of attributes these ratio levels were different, see Kalpic [2]. The financial award to the program designers was proportional to the expected cost reduction. The final result was very close to excellent.

The quality of automated coding program was analysed subsequently. The comparison was made between automatically assigned codes and those assigned in a usual way by the subject-matter people. For automatically assigned codes as well as for those computer-assisted assigned codes in all the four procedures (M1, M2, M3 and M4) the results are given in Tables 2 - 7.

Table 1

level	number of fields	%	cumulative	quality M2:M3
M1	26 183 829	67.44	67.44	
M2	8 454 389	21.78	89.22	67.07 %
M3	4 150 789	10.69	99.91	32.93 %
M4	36 486	0.09	100.00	
M5	29 805 097			
total	68 630 590	100.00		(M2+M3)=100%

Table 2. Procedure M1: automated coding with exact matching (67.4%)

Attribute		Number of fields in sample	Number of equal fields	%
Activity	- 2 digits	1043	1024	98.18
	- 4 digits		1014	97.22
Nationality		15456	15455	99.99
Place of birth		14128	13999	99.09
Country		891	886	99.44

Occupation	- 1 digits	3400	3330	97.94
	- 2 digits		3312	97.41
	- 3 digits		3288	96.71
School	- 1 digits	800	798	99.75
	- 2 digits		690	86.25
	- 3 digits		683	85.38

Table 3. Procedure M2: automated coding using complex algorithm (21.8%)

Attribute		Number of fields in sample	Number of equal fields	%
Activity	- 2 digits	736	627	85.19
	- 4 digits		609	82.74
Nationality		1584	1578	99.62
Place of birth		4878	4716	96.68
Country		301	292	97.01
Occupation	- 1 digits	2658	2443	91.91
	- 2 digits		2330	87.66
	- 3 digits		2204	82.92
School	- 1 digits	2847	2832	99.47
	- 2 digits		1979	69.51
	- 3 digits		1820	63.93

Table 4. Procedures M1 + M2: total automated coding (89,2%)

Attribute		Number of fields in sample	Number of equal fields	%
Activity	- 2 digits	1779	1651	92.80
	- 4 digits		1623	91.23
Nationality		17040	17033	99.96
Place of birth		19006	18715	98.47
Country		1192	1178	98.83
Occupation	- 1 digits	6058	5773	95.30
	- 2 digits		5642	93.13
	- 3 digits		5492	90.66
School	- 1 digits	3647	3630	99.53
	- 2 digits		2669	73.18
	- 3 digits		2503	68.63

Table 5. Procedure M3: computer assisted coding with 1 to 9 choices (10.7 %)

Attribute		Number of fields in sample	Number of equal fields	%
Activity	- 2 digits	1142	848	74.26
	- 4 digits		752	65.85
Nationality		136	125	91.91
Place of birth		1227	964	78.57
Country		16	13	81.25
Occupation	- 1 digits	3125	2406	76.99
	- 2 digits		1959	62.69
	- 3 digits		1711	54.75
School	- 1 digits	4000	3974	99.35
	- 2 digits		2400	60.00
	- 3 digits		2039	50.97

Table 6. Procedure M4: computer assisted coding without suggestions (0,01 %)

Attribute		Number of fields in sample	Number of equal fields	%
Activity	- 2 digits	39	25	64.10
	- 4 digits		24	61.54
Nationality		0	0	0.00
Place of birth		2	1	50.00
Country		16	11	68.75
Occupation	- 1 digits	18	12	66.67
	- 2 digits		11	61.11
	- 3 digits		6	33.33
School	- 1 digits	14	14	100.00
	- 2 digits		8	57.14
	- 3 digits		4	28.57

Table 7. Procedures M1 + M2 + M3 + M4

Attribute		Number of fields in sample	Number of equal fields	%
Activity	- 2 digits	2960	2524	85.27
	- 4 digits		2399	81.05
Nationality		17176	17158	99.90
Place of birth		20235	19680	97.26
Country		1224	1202	98.20
Occupation	- 1 digits	9201	8191	89.02
	- 2 digits		7612	82.73
	- 3 digits		7209	78.35
School	- 1 digits	7661	7618	99.44
	- 2 digits		5077	66.27
	- 3 digits		4546	59.34

It should be pointed out that all the values resulted from joint influences of optical reading and automated

coding, related to the results that the subject-matter specialists obtained after checking and correcting the inquiry forms. It is understandable that if all the material had been coded manually, the quality of the coders would be considerably reduced, and the expected quality would be lower.

It was noticed that the codes in the sample for the level M4 differ by 30 to 50 % from those which were assigned during the regular Census by the staff trained for processing such data sets. The application enables monitoring of the complete inquiry list and it can show the real quality of the manual coding.

For some attributes (activity, occupation, school) the obtained codes have a hierarchical structure. It is obvious that the results of the automated coding and the work of specialists on the sample differ more when analysing the lower level of the codes (**Table 7**.). This could be partly explained by the fact that the data are not unambiguous enough for assignment at the low hierarchical level. Several times it was proved that more answers were possible, the one obtained by computer coding, the one assigned in the sample processing, and sometimes even a third one. A possible conclusion could be that it is not suitable to search for the detailed answers for such complex questions.

5. CONCLUSION

The analysis of the automated coding procedure shows that it was carried out at the expected **high level of efficiency**. Almost 90% of data were automatically coded and 10% were coded in a computer assisted mode. The quantity of data to be additionally analysed was negligible. The coding was done following the daily entry and naturally considerably less time was needed than with manual coding. The average **coding speed** was about 3 000 records per hour.

In addition, the high rate of successfully coded fields indicated the **high quality** of automated coding algorithm and thesauri.

The entire result of automating the census data entry was satisfactory. The time needed for data input and coding was significantly reduced and it could be concluded that introducing new technology was a good decision of the Central Bureau of Statistics.

REFERENCES:

[1] Granquist, L., A Review of Studies on Impact of Data Editing on Estimates and Quality UN-ECE Joint Group on Data Editing. Washington, 1992.

[2] Kalpić, D., Automated Coding of Census Data, *Journal of Official Statistics*, Statistics Sweden. Vol. 10. No. 4, 1994, pp 449-463.

[3] Kovašević, M., Kontrola kvaliteta obuhvata popisnog materijala optickim citacem. Radni materijal, Republièki zavod za statistiku Republike Hrvatske, Zagreb, 1991.

[4] Lyberg, L., and Dean, Patricia, International Review of Approaches to Automated Coding. *Work Session on Statistical Data Editing*, Geneva, 1991.

[5] Perron, S., Berthelot, J.-M., and Blakeney, R.D. New Technologies in Data Collection for Business Surveys. *UN-ECE Joint Group on Data Editing*. Washington, 1992.

AUTOMATIC CODING OF DIAGNOSIS EXPRESSIONS

By Lars Age Johansson, Statistics Sweden

ABSTRACT

When the coding of a variable is complicated, automatic coding is useful for two main reasons: manual coding is time-consuming, and coding errors are common. That is why Statistics Sweden decided to introduce automatic coding for the variable ‘cause-of-death’. The objective was to develop a system that performs automatically multiple ‘cause-of-death’ coding, selects an underlying cause of death, and finally checks the coded records for any inconsistencies or incompletenesses. The paper presents the different modules, the encountered technical and statistical problems and the choices made to solve these problems (e.g. priority rules).

Keywords: cause-of-death coding; near-exact matching; text standardization; code modification.

1. INTRODUCTION

The MIKADO (an acronym for "MultIpelKodning Av DödsOrsaker" - Multipel Coding of Causes of Death) is a PC software for automated coding of multiple causes of death, developed at Statistics Sweden, Stockholm.

Manual 'cause-of-death' coding is, of course, afflicted with the same problems as manual coding in general: it is time-consuming, expensive, and liable to systematic errors. In addition, 'cause-of-death' coding has its own specific problems. By international agreement, causes of death are coded according to The International Classification of Diseases (ICD), which is extremely large, not very well structured, and abounds in arbitrary exceptions to its alleged general principles.

The coding of a certificate may be influenced by medical facts which are not explicitly stated, only implied. This makes the coding dependent on the medical knowledge of the coder. Since the coders' medical knowledge and familiarity with the ICD will inevitably vary, considerable efforts will be required to maintain acceptable coding stability. In our experience, it takes at least two years to train a new coder - if he/she has a basic knowledge of medical terminology and pathology.

On each death certificate, several conditions may be reported. When coding a certificate, the coder will first assign an ICD code to each one of the conditions reported (multiple cause coding), and then go on to select a principle cause of death (named "the underlying cause of death" in ICD terminology) according to selection rules specified by the ICD. Most statistical tabulations and analyses are based on the underlying cause of death.

Since the late sixties, a software has been available which applies the ICD selection rules to multiple cause coded certificates. This software, ACME, is developed and maintained by the National Center for Health Statistics, North Carolina, and is today the de facto international standard in its field.

ACME was introduced at Statistics Sweden in 1987. In 1989, we decided to try to automate the multiple cause coding as well, hoping that the coding would be faster, of higher quality, and less dependent on the individuals who perform it.

2. MULTIPLE 'CAUSE-OF-DEATH' CODING MIKADO

The aim of our project was to develop a module which translates the medical terms reported on the certificate into multiple cause codes. The following criteria were specified for the module:

- it must accept the language actually found on the certificates,
- if several conditions are reported in the same field, the module must be able to code them separately,
- it must allow supplementary and implicit information to influence the coding,
- CAC must be as similar to manual coding as possible, and
- the output must be in ACME compatible format.

Of the coding systems available in 1991, none met all these criteria. By the end of 1991, Statistics Sweden therefore decided to develop a multiple cause coding module of its own. A prototype, called AKK, was available in January 1993. After some modifications to the AKK (including renaming it AMK), a full-scale test was started in July 1993. A year later, the present version (named MIKADO) was introduced.

It was decided to work according to the "prototyping" model, i.e. the project was not aiming to write a complete specification of the coding software. Instead, a primitive prototype was developed very early in the project, and functions and refinements added to it successively. The main part of the work was done by two persons, working part-time on the project (50% for the first two years, 25% for the last year of the project). One was an experienced database programmer, the other a senior coder with previous knowledge of software development. Once the full-scale test was mounted, all coders took part in the evaluation of the software.

3. CONSIDERATIONS ON MATCHING STRATEGY

The first plans were to use near-exact matching, which seemed to be the obvious way to avoid inconveniently large dictionaries. At the beginning, the results were disappointing. A "most discriminating compound" method was tried first and then a strategy based on a computed similarity measure. However, both would yield a large number of theoretically possible, but unfortunately incorrect, dictionary matches. To achieve a reliable match a very high threshold value should be used and it was soon realised that exact matching would give the same results - and much faster.

The explanation of this outcome, which seems to be

at odds with experiences from many other automated coding applications, probably lies in the structure of medical language. Medical terms are often compounds of a comparatively restricted set of basic elements. These elements denote, for example, anatomical site or type of tissue (cervico-, neuro-, myo-, cardio-), or the nature of a disease process (-itis, -oma, -osis, -pathy). Many elements are quite similar, especially in Swedish spelling, which tends to truncate suffixes ("myocardosis" will be "myokardos") and remove letters which are silent in Swedish pronunciation (e.g., "h" in "cirrhosis" or "p" in "symptom"). Sometimes a single letter makes the whole difference between two quite separate entities, e.g. "arter-" and "artr-" ("artery" and "joint" respectively), or the Swedish words "hjõrt-" and "hjõrn-" (heart, brain). Moreover, medical terms are often quite long ("kardioartierionefrocerebroskleros"), and essential information on the nature of the disease is often given by the very last syllable ("myocardit", "-it" denotes "inflammation"). This means that word truncation and weighting methods which give higher weight to the early parts of the word will return many incorrect matches.

It was therefore decided to base the automatic coding proper (the part of the coding which will not be reviewed manually) on exact matching only. In the interactive coding, however, the coder has access to near-exact matching.

Of course, the performance of a system which uses exact matching only will depend largely on the efficiency of the phrase standardization (parsing). Much effort has been spent on the MIKADO parsing procedures, which are described in Annex 1.

Approximately 2% of the responses are written in "ordinary", non-medical language. In such cases (mainly descriptions of accidents and violence) exact matching is clearly not suitable, and the rate of automatically coded responses is low. Our experiences suggest that exact matching is preferable when scientific terminology is concerned, since such terminology consists of a comparatively small number of basic elements and even small variations can be of crucial importance. Exact matching is not, however, appropriate for coding of responses in natural, non-scientific language.

4. SYSTEM OVERVIEW

The MIKADO runs on IBM-compatible PCs with a 486 or higher processor. It uses the Paradox Data Base Manager, version DOS 4.5, and has been developed in the Paradox Application Language (PAL). For the time being, it is designed as a stand-alone application, not for PC network.

About 100 000 ‘cause-of-death’ certificates are sent to Statistics Sweden each year. The certificates are microfilmed and keyed to an ASCII file. A few standard abbreviations are used, otherwise all information on the certificates is entered manually exactly as it appears.

The ASCII files are converted to Paradox format, and then divided into work lots of about 450 records. The work lots are processed by the MIKADO in a batch process, and problematic terms or records will be flagged for manual review.

The work lots, including the code suggested by MIKADO, are then examined interactively by a coder. Any editing is done using a "working copy" of the input text, while the original version of the text is stored separately. To facilitate the work of the coder, we have tried to make the interactive coding as similar to manual coding as possible. Thus, the screen layout imitates the certificate form, the coder always has access to the entire text of the certificate, and the necessary operations can be performed in any order the coder prefers. Before the coder is allowed to return a work lot, MIKADO checks (among other things) that all conditions entered by the certifier have been coded.

Typically, the review may include operations such as correcting misspellings, and supplying codes for expressions not found in the dictionary. The coder can browse the dictionary in alphabetical or code order, and there are several search facilities available. Problematic records can also be referred to a senior coder.

Expressions not previously included in the dictionary will be copied to a provisional dictionary update file. The provisional dictionary update file will be reviewed by a senior coder and only then included in the dictionary. A "cloning" feature is available, by which it is possible to copy the codes and modification variables (see below) of an expression already included in the dictionary to a new expression. Each time the dictionary is updated, a check is run which ascertains that expressions with the same standardized text have been coded in the same way.

5. TEXT STANDARDIZATION AND PHRASE SEPARATION

To keep the dictionary reasonably compact, and to increase the number of matches, the phrases are standardized prior to coding. The standardization

procedure used by MIKADO includes steps such as removal of strings which do not influence the coding, replacement of some strings with synonyms, separation of phrases, alphabetical reordering of words in a phrase, etc.

A special feature of the MIKADO is that some strings will be coded separately when they are removed, e.g., expressions indicating surgery or other forms of treatment, or the duration of a condition. These supplementary codes may be used later to modify the code of the medical condition itself.

For a more detailed description of the standardization procedure, see Annex 1.

6. DICTIONARY OF DIAGNOSTIC EXPRESSIONS

There are two versions of MIKADO's dictionary of diagnostic expressions. One contains the expressions in their original, non-standardized form, whereas in the other, the expressions have been standardized according to the specifications in the current standardisation tables. Thus, an up-to-date version of the standardized dictionary can be prepared whenever the standardization specifications are changed.

6.1 Basic code and modified code

Code modification is a salient feature of ICD coding. Consequently, a medical term may have several different codes, depending on other information on the certificate. Even very common terms, like "heart attack" and "pneumonia", are subject to code modification. Therefore, an important part of MIKADO is the ability to handle such modifications automatically.

For every expression, the dictionary gives a basic code, that is, the ICD code to be used if there is no other information on the certificate that modifies the coding. In many cases there is also a modified code, that is, the ICD code to be used if there is indeed information present that influences the coding.

If an expression can have different ICD codes, the criteria for which code is to be used are specified by the modification variables. There are eight of these:

(1) duration of the condition,
(2) conditions reported elsewhere on the certificate,
(3) recent surgery,
(4) recent injury,
(5) in cases of external violence, possible intent (e.g., suicide, homicide, accident),
(6) age of the deceased,
(7) sex of the deceased and,
(8) specific expressions (text strings) used elsewhere on the certificate.

For modifications depending on the basic codes of other reported conditions or on other specific expressions, MIKADO also recognizes six different relations: (i) the modifying condition/expression immediately precedes the expression to be coded, (ii) immediately follows it, (iii) is reported on the same line, (iv) on a line above, (v) on a line below, or (vi) anywhere on the certificate.

If an expression can have only one ICD code, there will also be only one record in the dictionary, which will contain a basic code only. If an expression can be coded in several ways, there will be one record in the dictionary for each cause to modify the coding. Each record will have both a basic code and a modified code, and a specification of under what circumstances the modified code will be used rather than the basic one. For example, there is only one record in the dictionary for "alcohol-induced cirrhosis of liver", since no other information on the certificate can modify the coding of that expression. On the other hand, there are about 50 records for "diabetes". If there is a complication to the diabetes reported on the certificate, this will modify the coding, and there will be one record in the dictionary for each complication which can affect the coding.

6.2 Code priorities

If an expression can be coded in different ways, and consequently there are several dictionary records containing that expression, MIKADO checks for each case whether the conditions specified by the modification variables are met by the circumstances in the present case. If more than one of the dictionary records meets the criteria, the records are ranked according to a set of priority rules.

If there is more than one dictionary record with the same rank, and the records give different modified codes, the coder has to determine interactively which record to use.

7. RESULTS AND EXPERIENCES

Up to now, we have coded about 270 000 certificates using the AMK and the MIKADO. It has brought undisputable advantages: the coding is more accurate, much faster, and there is less need for

continuous quality checks. In 1992, the coding error (underlying cause, most detailed level) was estimated at 7.2%. In 1993, after the introduction of AMK, the estimated error was 3.1%. Of these, about 0.7% were attributable to the automated coding proper, 1.5% to the interactive coding, and only 0.3% to keying mistakes.

The batch processing of a work lot (450 certificates) will take about 15 minutes. Before standardization of the phrases, a dictionary match will be found for about 40% of the terms. After standardization, the success rate is now more than 90%, as compared to about 70% when the AMK was first put into operation. For about 65% of the certificates, the MIKADO codes every term on the certificate, and no manual review is necessary. With manual coding, it would take an experienced coder almost a full day to code a work lot; with the MIKADO, it will take less than half that time.

It is important to remember, however, that this does not imply that the coding is now 90% (or even 60%) cheaper than before. The MIKADO will take care of the uncomplicated certificates and leave the difficult ones to the coders, who sometimes have the impression that the coding is now slower and more difficult than before. The new technology has also generated several new tasks, such as running the batch jobs and other computer work, and updating dictionaries. Most of this work is done by the coders themselves, and not by IT staff.

The expenses of data entry are, of course, substantially higher, and to some extent use up what is gained at the coding stage. Full phrases are both longer and more difficult to type than digit codes, especially since the typists do not always understand the expressions. Besides, many a doctor's handwriting is quite as bad as is generally reputed. We have tried to scan the certificates, but the character recognition has not been successful enough. Even though about 70% of the characters were correctly interpreted, it took more time to correct the remainder than to key the certificates from scratch.

The introduction of automated coding has made it possible to work off much of the backlog we have had since 1987, even though the coding staff has been reduced from eight coders to six. Due to the backlog, however, no savings in money have been made. Presumably, there will be no substantial savings until the manual entry of the certificates at Statistics Sweden can be replaced by some form of electronic death certificate.

The testing will, however, continue. In 1997, Sweden will implement the tenth revision of the ICD. This is a major operation, which requires, among other things, retraining of the coders and independent recoding of each certificate until acceptable uniformity of coding has been achieved. When the ninth revision was introduced in 1987, the cause-of-death statistics were delayed for almost two years. We hope that MIKADO will make the transition smoother. If so, that will fully justify the resources invested in it.

A great problem with sophisticated automated coding systems is that coding expertise is lost. When the coders learn that MIKADO is usually right, they tend more and more to accept the coding suggested by the software, and gradually lose both their ability to code without computer assistance and to evaluate the performance of MIKADO. To counteract this, and to maintain the coding abilities which are needed to update the software, the coders are therefore required regularly to code training sets of certificates manually.

8. PLANNED MODIFICATIONS TO THE MULTIPLE CAUSE CODING MODULE

The present version of MIKADO was developed in Paradox for DOS, a software which is no longer supported by the manufacturer. Obviously, it has to be switched to another platform.

The tenth revision of the ICD will be introduced in Sweden on 1 January 1996. An ICD-10 version of the MIKADO is planned to be available by then.

For the validity checks, a mainframe system developed in 1986 is still used. These checks will be transferred to the MIKADO.

The ampersands, required in some circumstances by ACME to identify the starting point of a medical sequence, must be supplied manually in the present version of the MIKADO. Depending on the requirements of the ICD-10 version of ACME, we are considering including a feature which will supply these ampersands automatically.

ANNEX 1

MIKADO Parsing Procedures

1) The dictionary is searched for the text string to be coded.

If the string is not found in dictionary:

2) Trim blanks - any blanks first and last in the string are deleted, double blanks in the phrases are replaced by single ones.

3) Exceptions - flagging of strings NOT to be standardized in the usual way. Using this feature, e.g. "left" and "right" can be retained in connection with heart failure, where it influences the coding, but deleted in other cases, where it does not.

4) Hyphens are removed or replaced by other characters.

5) Prefixes and suffixes are removed and replaced.

6) Deletions - words and strings which do not affect the coding are removed, e.g. "the patient had...", "probable".

7) Replacements - spellings and expressions are standardized.

8) Periods - remaining periods are removed or replaced.

9) Standardization of phrase separators - strings indicating the beginning or end of a diagnostic expression are replaced by one of three standard separators (";" for enumeration, "*>>*" for a "giving rise to"-type relationship, "*<<*" for a "caused by"-type relationship).

10) Surgery - expressions indicating surgery or medical treatment are coded separately and then deleted.

11) The exception sign "#" is removed.

12) If an expression has been deleted in its entirety, it is replaced by a "not known" string.

13) The dictionary is searched for the standardized string.

If still not found:

14) Durations - expressions indicating the onset of or the duration of a condition are removed and, if possible, coded separately. If automated coding of the duration is not possible, the expression is marked for manual duration coding.

15) The dictionary is again searched for the standardized string.

If still not found:

16) All remaining blanks are removed from the standardized string, and a corresponding field in the dictionary (containing the standardized diagnostic expressions with all blanks removed) is searched for a match. If no match is found, the blanks are restored.

17) The words of the phrase are sorted in alphabetical order, and the search is repeated, this time in an alphasorted field.

If still not found:

18) Phrase separation - the string is searched for any standard separator (";", "*>>*" or "*<<*"). If a separator is found, each substring will be standardized as described above (1 - 17) and a dictionary search performed.

If still not found, or if no phrase separators are found:

19) Mark the expression for interactive coding.

SICORE - GENERAL AUTOMATIC CODING SYSTEM

By Pascal Rivière, Institute Nationale de la Statistique et des Etudes Economiques, France

ABSTRACT

The article presents the main principles of the automatic coding system SICORE, launched in INSEE in 1993. The coding algorithm is based on the QUID method which has been improved and generalized. Besides being a software tool, SICORE is a general automatic coding system which also handles the pre-coding process, i.e. preparation of knowledge bases, and the post-coding process, i.e. analysis of uncoded or improperly coded verbal responses. The results of this analysis are used for updating the knowledge bases. The article also touches the methodological and organizational aspects of integrating automatic coding into the statistical survey process. The SICORE system has been used in tests and in actual production in a number of different set-ups and on a variety of text descriptions: nationalities, cities, professions, company names, enterprise addresses, countries, human activities, places of residence and financial products.

Keywords: general coding software; automatic coding; text recognition.

1. INTRODUCTION

The SICORE project was launched in spring 1993. Its aim was to give INSEE an easy-to-use general automatic coding tool and to develop the organization and methodology needed to use it efficiently. Three years later, the system is largely complete and has proved to be efficient in operation.

The first two parts of the paper describe how SICORE operates as a coding tool: the distinction between programmes and knowledge, the content of the knowledge bases, and the learning and coding algorithms. However, this presentation should not be limited to the programmes alone. The ?SICORE system" (application of coding skills, improvement of knowledge and the SICORE organization) is described in the third section. The last sections concentrate on practical issues and answers to questions such as: how is SICORE used in practice? How is automatic coding introduced into a survey? What is SICORE used for and with what results?

2. AUTOMATIC CODING USING 'SICORE'

2.1 General Principles

SICORE is a general coding software. As described in the first paper of this chapter, the main principle of the general coding algorithms is keeping the learning file and programs in separate files.

SICORE develops this principle further. All the information needed to code a given variable is found in the *knowledge bases* which are separate from the general programs. The learning file forms a part of the knowledge base. In this way the programs will remain the same regardless of the variable to be coded.

Automatic coding using SICORE therefore involves two operations: loading the knowledge about the variable, and coding. The first step is most important; without knowledge, SICORE is an empty shell incapable of any kind of automatic coding. Thus, knowledge must be provided before anything else can be done.

This observation has an important conclusion: when the automatic coding result is erroneous, the error is mostly caused by the knowledge bases used.

2.2 The Knowledge Bases

A knowledge base is a coherent set of knowledge on a given field *relating to a given variable*.

SICORE contains six knowledge bases on coding, i.e. six fields that summarise the current level of knowledge. These bases describe:

- **standardization rules**

  Definition of transformation rules from raw text to "clean" text suitable for analysis (defining the list of blank characters, empty characters, empty words and synonym pairs, limiting the number and size of words).

- **learning file**

  This is the most important knowledge base. It has a very simple structure: each line contains a description and a code. When the description is not specific enough and supplementary variables are needed, the code is an intermediate code designating a *decision table*. A decision table contains a series of logical rules for the inclusion of

supplementary variables.

- **logical coding rules**

  These are used when the description alone is not sufficient for coding and additional information is needed to resolve any ambiguity. The logical rules' file is the file containing the definition of all the tables of this type.

- **the learning algorithm parameters**

  This is a small knowledge base used to build up the questioning tree, which can substantially improve learning time and coding efficiency.

- **cross-coding of supplementary variables**

  This is specific to the source text that needs to be coded and describes how to ?translate" the supplementary variable values.

- **coding layout**

  This defines where data (e.g. description, supplementary variables) is to be found and how to move these data, if necessary.

Each knowledge base can be loaded and regrouped into a single file, called the **comprehensive field file**. This file contains the decision tree, the standardization rules, etc. in a SICORE internal syntax.

2.3 The SICORE Learning Algorithm

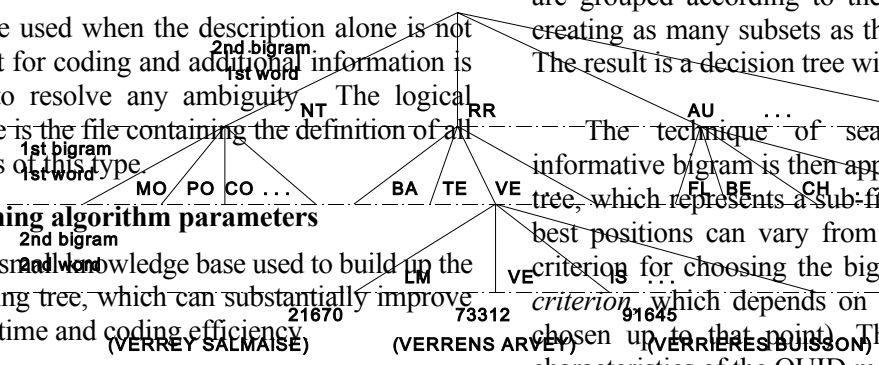
The SICORE learning algorithm is based on a general coding algorithm, known as QUID [1]. It develops the original algorithm further making it more flexible.

Learning is applied to a file *(the learning file*) in which a code is associated with each description for the variable considered. In the learning phase the verbal responses are broken down into sequences of two characters (*bigrams)*. Trigrams, monograms and quadrigrams can also be used, but the bigram has proved to be the best choice.

Each record of the learning file consists of::

- the *text to be coded* in a structured format;
- the *code* associated with the text.

In the original QUID algorithm the entire learning file is searched for the *bigram*[1] that provides the most information for the purposes of coding. To do this, the information gain[2] is calculated for each bigram and the best one is taken. Once this is done, the verbal responses are grouped according to the values of this bigram, creating as many subsets as there are different values. The result is a decision tree with bigrams at the nodes.



The technique of searching for the most informative bigram is then applied to each node on the tree, which represents a sub-file of the initial file. The best positions can vary from one node to next. The criterion for choosing the bigram is therefore a *local criterion*, which depends on the context (the bigram chosen up to that point). This is one of the basic characteristics of the QUID method.

SICORE introduces numerous variations making the algorithm more flexible. It gives the user the possibility:

- to set up a list of priority bigrams, i.e. to set an order for choosing bigrams in the early stages of tree construction;
- to dictate that a bigram is not to be taken into account until a certain level in the tree;
- to dictate that a given bigram is never to be considered before another bigram;
- to make the redundancy check more flexible by "accepting" bigrams made up of two spaces, regardless of the value of the bigram with which it is compared;
- to introduce "joker" words into the learning file.

The most useful of these is the **priority bigram** technique.

Different from QUID algorithm, the building of the tree starts based on a given list of bigram positions. The local information criterion principle is followed further down the tree.

For example, if the list of priority bigrams is (2, 1, 8), the 2nd bigram position is taken first, followed by the 1st and the 8th. After exploring these three tree nodes the information criteria are considered in order to search for the most informative bigram, if necessary.

The learning algorithm for the names of French cities Verrey-sous-Salmaise, Verrens-Arvey and Verrières-le-Buisson results in the tree shown in Figure 1.

[1] Or rather the bigram position

[2] The entropic (conditional) variation featured in Shannon's information theory.

In the case of the city, the identification of bigrams 2, 1 and 8 leads in three steps to the cities of Verrey-sous-Salmaise, Verrens-Arvey and Verrières-le-Buisson (the "-" and the word "sous" disappear after standardization).

The bigram positions considered here *for the first three tree levels* are fixed. The 2nd bigram of the 1st word is given top priority as this is the bigram that provides the most information in the French language.

Using priority bigrams has several advantages. First and foremost, it sharply **reduces learning time**. It is no longer necessary to calculate all the conditional entropies. When the first two bigrams of the first two words are used as priority bigrams (which is generally the case), processing time is generally divided by three and can even be cut by four or five.

It should also be noted that **coding time does not increase**: the average coding time is about the same as that required for a tree based entirely on the local information criterion. The coding time per unit is a few 10,000ths of a second on a Pentium 90.

Tests on profession descriptions, company names and addresses (for the next population census) have also proved that adding priority bigrams results in a substantial increase in efficiency. In this particular case, it was found to be highly advantageous to choose the address bigrams as the first priority bigrams.

Moreover, the use of several priority bigrams makes the SICORE tree more stable: the first branches of the tree, which correspond to the priority bigrams, cannot be altered by introducing new descriptions into the learning file. Instability resulting from changes in the learning file was one of the main drawbacks of the previous algorithm.

Finally, priority bigrams enable users to find the best compromise between efficiency and reliability. Users can optimise parameters to attain maximum possible efficiency while maintaining acceptable reliability. Such an operation requires users to have a certain level of expertise.

2.4 The Coding Algorithm

2.4.1 The Basic Method

Once the SICORE system has been supplied with sufficient knowledge, it can start coding. The procedure involves three steps.

1. **Standardization** - deleting blank and empty characters and empty words, replacing synonyms and abridging the description.
2. **Questioning** - running through the tree to find a code, which can also be an intermediate code.
3. **Determining the final code** using supplementary variables and the logical coding rules.

The output from SICORE is a file which contains: a standardized description, any potentially useful additional information, a coding response and a coding result (0, 1 or more codes). The coding response gives information on whether the coding process was successful and, if not, then why.

The response can take the form of different values, describing the following situations:

Coding Response	Definition
Simple coding	The description was coded without using any supplementary variables and the redundancy control check was successful.
Redundancy error	As in the preceding case, the questioning was successful, but the redundancy check was not satisfactory. Thus the description is considered as uncoded.
Coding failure	The questioning was not successful, which means that at a certain tree node, the branch corresponding to the value of the available bigram was not present.
Multiple coding	The learning file contained at least one description associated with several different codes. Consequently, when the tree was created, one of its leaves was associated with several codes. When it coded, SICORE logically responded with the codes in question. In this case, it is also considered that SICORE failed.
Simple coding using the logical rules	The description was recognised, the redundancy check was satisfactory and a code was generated using the logical rules without any error in the latter phase.

Error when using the logical rules	The description was recognised, the redundancy check was satisfactory, but in at least one decision table, none of the rules were applicable. It was therefore not possible to continue progressing through the rules to obtain a final code. This is helpful for the expert who has created the knowledge base; it makes it possible to deal with inconsistencies between the description and the supplementary variables.

A description is thus considered as automatically coded if, and only if, the coding response is either *simple coding* or *simple coding using the logical rules*.

2.4.2 Coding Heterogeneous Descriptions

When descriptions are made up of several sub-descriptions - the so-called heterogeneous descriptions, they have to be coded by considering the different components separately. **This method is called complex coding** and can be described as a series of standardizations and codings of the components of the description.

Processing area and city description together consists of first coding the area while considering the city as additional information to be stored. Then the "area code + city description" is coded.

When working on heterogeneous descriptions, coding efficiency can also be improved by applying several questioning trees based on different principles (multiple-tree coding). This method makes use of the fact that some part of the description may be missing or be of poor quality while the rest is good. Thus, we can build a tree where one part of the description is "prioritized" (using priority bigrams) followed by another tree prioritizing another component.

This technique was successfully applied to the company name descriptions in the last population census: efficiency using one questioning tree was 37.5% whereas combined efficiency rose to 49% when the second tree was applied to the uncoded descriptions. Multi-tree coding therefore increased efficiency by 11 points.

3. THE SICORE SYSTEM

If an automatic coding application is to be sustainably efficient, it needs constant maintenance: language evolves, new expressions appear and classifications change (appearance of new kinds of human activities, the merging of cities, new ways of looking at activities, etc.).

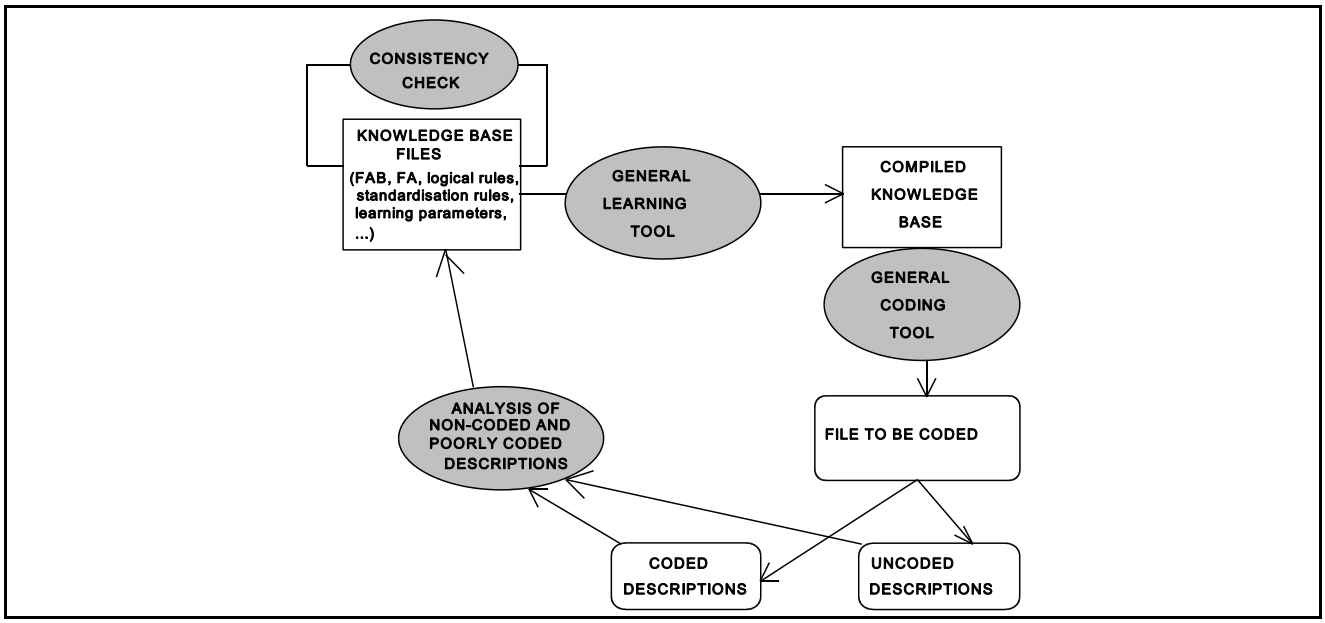
The knowledge bases should be continuously updated with new expressions and new codes. Moreover, new supplementary variables, more detailed logical rules, useful synonyms and appropriate empty words should be developed.

Consequently, an automatic coding system must be concerned not only with the existence of solid knowledge bases but also with their updating, which means considering the methodology and organization used in their development.

3.1 The SICORE Loop

Thus, the SICORE maintenance is based on **permanent feedback**: automatic coding 6 results' analysis 6 knowledge updates and consistency checks 6 creation of a new comprehensive field file 6 automatic coding based on this comprehensive field file.

Figure 2. The SICORE Loop (for a given variable)



The cycle, called the SICORE loop, can be described as in Figure 2.

The automatic coding works based on the compiled knowledge base.

Following automatic coding, the expert **analyses the uncoded descriptions and a sample of coded descriptions** to find some poorly coded descriptions. The expert decides what is coded "well" or "poorly".

The expert **modifies the knowledge bases** according to his or her findings: adds descriptions to the raw learning file, modifies the logical rules, introduces new synonyms, checks the cross-codings, etc. The learning parameters, and in particular the priority bigrams, form part of the knowledge bases.

The next stage is the **consistency check**, once the new knowledge bases have been finalized. If a problem is found, changes are made, which can themselves affect consistency. This process continues until the entire knowledge base is considered to be correct. At this point, a comprehensive field file is created to be used for new automatic coding and a new loop begins.

3.2 Developing Knowledge Bases

The aim of updating knowledge bases is to improve their quality in terms of **efficiency** (the automatic coding rate) and **reliability** (percentage of codes correctly coded). Once a file has been automatically coded, there are three subsequent phases to the operation: detecting problems, modification and checking.

3.2.1 Detecting Problems

The detection phase analyses the uncoded and poorly coded descriptions. The files are displayed so as to highlight the most significant problems. To do this, the files are sorted by decreasing frequency to find what will be added to the learning file or to determine empty words and useful synonyms.

More generally, the SICORE coding results are analysed using the concept of **transformation**. A "transformation" is an application that transforms one file into another file with the help of the following *elementary operations*: sorting, simplification, filtering, housekeeping, sampling, standardization and coding.

3.2.2 Modifying the Knowledge Bases

The modification phase does not actually affect all the bases: in practice, the learning parameters, coding and cross-coding layouts are quickly determined. Thus, the new learning file is enriched with some new descriptions, synonyms are added or removed and some logical rules are refined. The crucial point in the modification phase is to determine what knowledge to update and to what extent.

In this context, the priority bigrams technique facilitates the work, especially if the first bigrams of the first words have been chosen. This technique ensures that the word "roots" are systematically taken into

account. These roots are often sufficient for recognition and it is therefore unnecessary to add synonyms for the root.

Once the priority bigrams have been run through, SICORE searches for the bigram providing the most information according to the local information criterion. *Thus, if two descriptions have different words with identical roots and different codes, there is a good chance that subsequent (post-root) bigrams of the words will be used to distinguish between them.* For single-word descriptions, this is inevitable (as in ?PHYSICIAN" and ?PHYSICIST").

Consequently, *when two words have the same root but completely different meanings, it is better to over-represent them in the learning file* in order to avoid poor coding and to prevent any inaccuracies in the learning algorithm.

It is useful to add short descriptions to the learning file (two words or one word). There are two major reasons for this: firstly, because this often requires the learning algorithm to seek out systematically useful bigrams. Secondly, because interviewees often give brief responses, e.g. when they feel the word 'employee' to be sufficiently descriptive.

In automatic coding, the term synonym simply means that word X should be replaced by word Y wherever X is found. There are several **types of synonyms**:

abbreviations (AUXILIARY = AUXIL, HAUT = HT), *semantic equivalents* (ARE = AM, BE = AM), *common initials* (CHIEF EXECUTIVE OFFICER = CEO) and *inclusions* in a more general category (VIOLIN = MUSIC, CUCUMBER = VEGETABLE).

The abbreviations are worth using whenever the word (or one of its synonyms) is used often and alone does not provide enough information to determine the code. This brings to light two contradictory objectives: to have a description that is short enough to be used as the root, yet long enough to avoid having a number of synonyms (especially as regards the many feminine forms in French). The breakdown into bigrams means that it is preferable to have an even number of characters in the reference synonym. In general, *in the case of ?abbreviation" synonyms*, a reference word of four or six letters is perfectly suitable.

The same rule of thumb is applicable for *semantic equivalents* (words with the same meaning in a given context, but which are not necessarily abbreviations): systematically use the synonym letters) are different and choose the reference word well (length and use as a root).

The *common initials* do not pose any problem with regard to the choice of the reference word or its length: the reference word is made up of the initials. However, they require a careful ordering of the synonyms and attention to preceding synonyms. For example, the description ?CHIEF EXECUTIVE OFFICER" will be written *after* synonyms for the word ?CHIEF", with close attention paid to the fact that the reference word is CHIEF and not CHIE.

Finally, the *inclusions* are extremely efficient at accurately integrating words or groups of words with the following two characteristics:

- they are themselves of little importance;
- they belong in some way to a ?category" and do not provide any further information than the category. Example: BBC and FRANCE INTER belong to the category RADIO for coding descriptions of day-to-day activities.

Yet synonyms should be used with caution as words often have several meanings.

3.2.3 Checking

This is the final stage following the updating of the knowledge bases. It involves a dual analysis: an internal analysis of the bases (i.e. independent of any file to be coded) and a test of automatic coding.

The **internal analysis** is simply a consistency check on each knowledge base taken separately (*in-base consistency*) and on pairs of knowledge bases (*inter-base consistency*). E.g. a learning file will not be internally consistent if it uses invalid codes that are not part of an official list. It will be ambiguous if a single description can generate several different codes. When supplementary variables are available, the pair (FA learning file, logical rules) will not be consistent if there are codes in the learning file that are neither official codes nor names of decision tables.

Thus, it can be seen that the internal analysis helps to resolve ambiguities and bring to light inconsistencies, but does not solve a substantial problem with the knowledge bases: their incompleteness. It is possible to have a perfectly consistent set of knowledge bases with completely inefficient automatic coding.

To solve this, an **automatic coding test** is run. The

initially used file (from which the uncoded and poorly coded descriptions are taken) is coded once again. Previously coded descriptions may turn up no longer coded as the standardization rules may have substantially changed the operation. So this test again refines the knowledge bases in line with overall efficiency and reliability objectives.

3.3 The SICORE Organizational Network

This network is still new as the product has not yet been used much in production.

At present, there are **experts** and **knowledge bases** for the following variables: profession (and socio-economic category), city (as well as areas, countries and nationalities), day-to-day activity (for the survey on work schedules) and establishment (specifically for the population census). Other knowledge bases have been set up, but are not yet operational on a full-time basis: these include mutual funds and places of residence.

The two variables considered most important for automatic coding are 'profession' and 'city'. Each is the focus of a **working party** (*SICORE-Profession* and *SICORE-Commune*, respectively), which meets two to three times a year. The working parties constitute a forum where automatic coding problems are raised and the specific coding features of one or another survey are highlighted.

For example, the SICORE-Profession group has already addressed and taken action regarding the existence of two distinct semantic fields (one for households and one for businesses). Meetings of the SICORE-Commune group have highlighted how important the ?date" supplementary variable is for coding cities, in particular for the Registry Office.

The SICORE system is organized around a SICORE **expert**. The expert is the contact person for any statistician who wants to use automatic coding in an application. The SICORE expert centralizes the knowledge bases, manages the network of variable experts and the corresponding working parties, takes requests about automatic coding into consideration, advises users, organizes training and recommends developments to the tool.

The SICORE expert communicates with the SICORE **computer team**, which is responsible for developing SICORE, integrating it into the survey processing operations, and also for the SICORE ?integration guide".

When automatic coding is carried out (in any survey), the SICORE expert receives the results file and sends it to the variable expert, who analyses it (uncoded and poorly coded descriptions) and updates the knowledge bases accordingly.

This operational procedure, still in its infancy, was used following the survey on living standards, in which the socio-economic category was coded automatically by SICORE: the descriptions were studied thoroughly by the Profession experts, who significantly modified their knowledge base files based on this analysis.

3.4 Using SICORE

The term ?SICORE" can relate to a number of things. It stands for the system as a whole: programs, knowledge bases, methodology and organization. Yet it can also be looked at exclusively from the point of view of the software program. As a software tool, SICORE is *a set of programs forming the building blocks for coding*: reading the knowledge bases, learning, standardization, description recognition, processing of supplementary variables, etc. These building blocks are written in C and function just as well on a PC as on a mainframe.

On PC, all these blocks are combined into a single program operating under Windows. On a mainframe, the separate building blocks are called up by different programs.

Speaking about the **SICORE tool**, we are referring to the PC software consisting of the automatic coding programs and an interface used to create, analyse and update the knowledge bases. The main aim of the PC SICORE program is to develop the knowledge bases for which the automatic coding is one of the tools.

When it comes to *updating knowledge,* this program requires delicate handling and is used by less than ten people: the variable experts (socio-economic category, city, human activity) and the SICORE expert.

In contrast, when the knowledge bases are already available,the automatic coding with SICORE is very easy. However, this is also carried out on the central site, using only the basic building blocks: the rough method, learning, knowledge reading, standardization, coding and transformations.

Finally, one could even envisage a situation involving statisticians, not acting as experts, who simply *carry out a coding*. A user who is neither computer scientist nor expert could only use the system if the

work were carefully prepared beforehand: the computer scientist would prepare the file to be coded, the variable expert would supply the comprehensive field file, the SICORE team would set up the software program and explain how to use it.

3.5 Setting up Automatic Coding in a Survey

Using SICORE to process a given variable in a given survey requires some preparation and a suitable division of labour among the personnel. Experience shows that five main specialists are normally needed: three statisticians and two computer scientists, each of whom may be head of a team:

- the SICORE expert,
- the statistician in charge of the survey,
- the expert on the variable to be coded,
- the SICORE computer manager,
- the survey computer manager.

The two SICORE representatives are always present regardless of the survey or variable considered.

When a variable needs to be coded, the variable expert is always the same regardless of the survey. In contrast, the survey statistician and survey computer scientist are connected to the survey, but are often independent of the variable to be coded; for example, the same statistician-computer scientist team could handle the coding of both socio-economic categories and places of residence.

Among these five specialists, there are both horizontal relations (between statisticians and computer scientists) and vertical relations (between SICORE representatives and survey representatives).

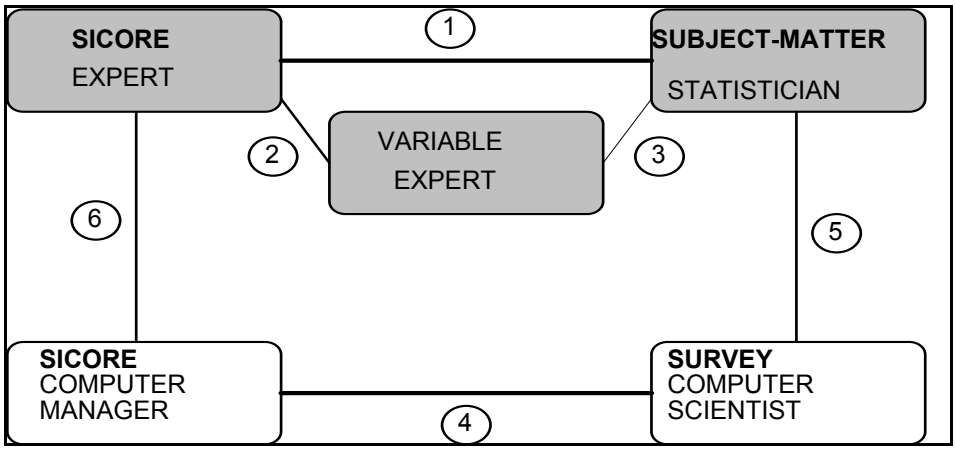
The relative importance of the different specialists depends on the case in hand, and not all of them are necessarily needed. The starting point for setting up automatic coding in the application lies in *determining whether it is feasible to automatically code the variable* in question. If no learning file exists for the variable, it could take a long time to create the required knowledge bases. The survey statistician also has to evaluate the costs of data entry and the computer assisted coding program (even a rough version) that needs to be created. Obstacles could arise concerning the computer system: incompatibility between sites, the fact that SICORE takes up a huge amount of memory, etc. Thus it may become evident at this stage that automatic coding is too expensive compared to the acquired benefits.

If it is determined that it is possible and worthwhile to use automatic coding, the survey statistician will proceed (in liaison with the SICORE expert and the variable expert). *The survey's particularities are determined*: list of supplementary variables used, cross-codings to be defined, file layout. This does not require much time, but must be done carefully.

In the information technology field, the SICORE computer manager and the survey computer manager need to *integrate the SICORE building blocks* into the survey processing programs.

The final stages are more conventional: *tests* (statistical tests on the quality of the results and computer tests on the functioning of the overall procedure) and, finally, *the processing itself*.

Figure 3. SICORE System Personnel and Relationship Links



When the entire automatic coding procedure has been completed, the coding results are sent to the SICORE expert for *updating the knowledge bases*.

4. SICORE APPLICATIONS

4.1 The Variables Tested

SICORE has been tested on the automatic coding of **numerous variables**: socio-economic category, socio-economic category and profession, country, city, area, place of residence, enterprise, human activity, mutual funds and type of business (in English). Most tests were conducted on PCs and combined with the updating of knowledge bases.

A **wide variety of scenarios** were found: homogeneous descriptions without supplementary variables (mutual funds, area), homogeneous descriptions with supplementary variables (profession, human activity), slightly heterogeneous descriptions (area code plus city description), highly heterogeneous descriptions (company name-address, area description plus city description), creation of knowledge bases from scratch (human activity, mutual funds), merge of learning files (place of residence) and the coding of descriptions in several languages (type of business, for Statistics Canada).

The **sizes of the learning files** varied from a few hundred (area) through 11,000 (profession) and 44,000 (cities) to hundreds of thousands (company name-address). Learning time on a PC Pentium 90 with the current version of the learning parameters was around 1 second for small learning files like the mutual funds file (3,000 descriptions), 10 seconds for the professions file, and 40 seconds for the cities file. Learning time on a DEC Alpha 8400 for the whole file of establishments (4 and a half million!) was around 29 minutes.

The **number of synonyms** also varied greatly: from zero (mutual funds) through 500 (human activities) to 1,200 (professions). The cities only made use of a few dozen empty words or synonyms.

For important variables like profession and city, coding was applied in different contexts: the population census, Registry Office civil status certificates and the Annual Social Data statements on employees and wages (ASDS).

The **efficiency of coding** differed greatly according to the context.

For professions, for example, approximately 66% of the descriptions in a file taken from the 1990 population census were coded. The coding rate was 82.5% for socio-economic categories using the ASDS, between 85% and 89% for socio-economic categories using Registry Office civil status certificates and 76% using the Living Standards survey (still for socio-economic categories).

Depending on the case, the efficiency of coding cities varied from 93% to 99.5%.

The automatic coding rate for descriptions of human activities is currently approximately 65% (with satisfactory reliability), but the knowledge bases (built from scratch) change a great deal in this field.

The rate for mutual funds has risen from 61% to 80%, but there is only one file of descriptions to be coded and the learning file has been updated based on the uncoded descriptions in this file. Thus, efficiency can be raised as much as desired and the figure of 80% probably greatly overestimates actual efficiency.

The **coding speed** does not vary much; moreover, the speed variation has no importance as SICORE is an extremely fast coding system. As an average, on a PC Pentium, SICORE codes approximately 5,000 descriptions per second.

The **coding reliability** depends on the learning parameters. Reliability is difficult to measure as it requires an expert coding. As far as we know, it is more than 99% in the case of cities, and more than 96% in the case of occupations.

4.2 Use in Production

SICORE was used for the *survey on the Rhône-Alpes regional public transport system,* SYTRAL. For this survey the cities had to be coded, i.e. the departure and arrival points for each trip on public transport. Out of thousands of descriptions, only three were not coded by SICORE: assisted coding was thus not necessary.

It is difficult to draw any statistical conclusions from this experiment as coding cities is not the most difficult case. However, from the point of view of information technology, computer operator time was very low (1 to 2 days) in this case (where there was no real integrated processing system).

SICORE was also used to code socio-economic categories and places of residence in the Living Standards Survey (in the household living standards section). This situation was ideal for an assessment of SICORE, as manual coding was carried out simultaneously.

Firstly, the tool itself was tested, particularly for how well it was integrated into the computer system. Secondly, the efficiency and quality of automatic coding was analysed.

Due to close liaison between the Living Standards survey team and the SICORE team *no particular problems* arose in integrating SICORE into the survey processing system. 9,992 socio-economic category descriptions and 12,239 place of residence descriptions were coded. SICORE automatically coded 76% of the socio-economic category descriptions and 92% of the place of residence descriptions. This was considered to be a satisfactory efficiency rate.

To compare the quality of automated and manual coding, we had to see how many times the automatic coding differed from the manual coding and then to analyse the deviations on a case-by-case basis. Only 3% of the descriptions coded by SICORE gave different results. An examination of these individual cases showed that the automatic coding was better: the SICORE code was accurate in 82% of the cases and the manual code in the remaining cases.

Coding a socio-economic category, 18% of the professions descriptions coded by SICORE differed from the manual code. The differences were analysed by the Profession experts who calculated that, in contrast, the manual coding was correct in 17.5% of cases. For the rest, in 20% of the cases, both codes could be considered to be suitable; there was ambiguity, which was often due to the fact that SICORE does not use unprocessed activity descriptions.

4.3 The Present and Future of SICORE

It is worth restating that SICORE as a tool for developing knowledge bases is not designed for widespread use, as it requires a certain expertise in both SICORE and in variables to be coded. There is a special training course and three manuals are supplied: the user*s guide for the PC tool, a detailed glossary and a methodological document.

In contrast, this part of SICORE that can be used as pure automatic coding tool (at the central site) is fairly easy to integrate into an application. The SICORE COMPUTER team currently does most of the integration work, but survey computer scientists will eventually be able to do this themselves. The relevant documentation (programmer*s manual and integration guide) has been available since the beginning of 1996.

Work is also continuing on new functions: the ?consistency check" and ?transformation" features are currently being completed.

In the future, SICORE will be used for coding socio-economic category descriptions in the Annual Social Data Statement on Employees and Wages (thus replacing QUID) and will again be used for the Living Standards Survey. Other applications are expected to follow, for example, for other public surveys. A long-term project is planned to use SICORE for the Employment Survey.

Finally, there are plans to market SICORE: several bodies in France and a number of statistical services abroad (in the United States in particular) have shown an interest in the product.

REFERENCES:

[1] Lorigny, J. QUID, une méthode générale de chiffrement automatique. [QUID, A General Automatic Coding Method], *Survey Methodology, December 1988,* Vol. 14, No. 2, pp. 289-298.

[2] Lyberg L., Dean P., Automated Coding of Survey Responses: an International Review. Working Paper, *Conference of European Statisticians, Work Session on Data Editing*, Washington, March 1992.

[3] Riviere, P., Le système de codification automatique SICORE, *Le Courrier des Statistiques* n° 64, 1995.

[4] Wenzowski, M.J., ACTR - A Generalized Automatic Coding System, *Survey Methodology, December 1988*, Vol. 14, No. 2, pp. 299-308.

[5] Riviere, P., SICORE, système sénéral de codage automatique, INSEE - Méthodes - à paraître, 1996.

STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

BIBLIOGRAPHY

on Statistical Data Editing

March 1997

prepared by
UN/ECE Secretariat

Abbate C. (1996). La completezza delle informazioni e l'imputazione da donatore con distanza mista minima, *will be published in Quaderni di Ricerca ISTAT*

ACTR (Automated Coding by Text Recognition) Version 1.06 - User Manuals

Allen, J. D. (1990). An overview of imputation procedures. SMB Staff Report, United States Department of Agriculture.

Alvarez, A., Villan, I. (1991). Automatic Coding in the Spanish Statistical Office. Presented at the Conference of European Statisticians' Work Session on Statistical Data Editing, Geneva, 1991.

Ambler, Carole A., Hyman, S. M. and Mesenbourg, T. L. (1993). Electronic Data Interchange, International Conference on Establishment Surveys, Buffalo, New York, 1993.

Anderson, Christopher (1996). A World Gone Soft: A Survey of the Software Industry, The Economist, May 25, 1996.

Anderson, K. (1989a). Draft, Output Edit Study, Average Weekly Earnings, Statistical Services Branch, Australian Bureau of Statistics, September 1989.

Anderson, K. (1989b). Enhancing Clerical Cost-Effectiveness in the Average Weekly Earnings, Draft, Australian Bureau of Statistics,Statistical Services Branch, 9 November 1989.

Anderson, T.W. (1958). An Introduction to Multivariate Statistical Analysis, Wiley & Sons, pp 126-153.

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P.J., Rogers, W.H. and Tukey, J. W. (1972). *Robust estimates of location: survey and advances*. Princeton, NJ: Princeton University Press.

Armstrong, B. and Pursey, S. (1979). Imputation research: Total imputation methods. Statistics Canada Technical Report.

Ashraf, A. and Macredie, I. (1978). Edit and imputation in the Labour Force Survey. Proceedings of the Section on Survey Research Methods. American Statistical Association, pp.425-430.

Atkinson, D. (1988a). The scope and effect of imputation in quarterly agricultural surveys. USDA Technical Report.

Atkinson, D. (1988b). Travel Notes - Budapest, Hungary. Internal Memorandum, National Agricultural Statistics Service, concerning Atkinson's participation in the second meeting of the Data Editing Joint Group of the United Nations Statistical Computing Project, Phase 2, April 18 and April 22, 1988.

Atkinson, D. (1991). Pakistan October Acreage Survey Edit and Summary System. NASS -- U.S. Department of Agriculture, (UN-ECE, SCP-2 Product No. SCP-2/D.8/f).

Australian Bureau of Statistics (1993). Data Editing - an ABS manual, Belocnnen: Australian Bureau of Statistics.

Aziz, F. and Scheuren, F. (eds.) (1978). Imputation and editing of faulty or missing survey data (Selected papers presented at the 1978 Annual Meeting of the American Statistical Association). U.S. Dpt. of Commerce, Washington D.C.

Badeyan, G. (1992). Quality checking of data entry and coding for the French 1990 Population Census. Presented at the Conference of European Statisticians' Work Session on Statistical Data Editing, Washington, 1992.

Bailar III, J.C. (1978). A discussion of the nonresponse and imputation issues session. Proceedings of the Section on Survey Research Methods. American Statistical Association, pp.230-232.

Bailar III, J.C. and Bailar, Barbara A. (1978). Comparison of two procedures for imputing missing survey values. Proceedings of the Section on Survey Research Methods. American Statistical Association, pp.462-467.

Bailar, Barbara A., and Bailar, III, J.C. (1983). Comparison of the biases of the hot-deck imputation procedures with an "equal-weights" imputation procedure. In Incomplete Data in Sample Surveys. Volume 3, Proceedings of the Symposium. W. G. Madow, and I. Olkin (eds.). Academic Press, New York, pp.299-312.

Bailar, Barbara A. (1983). Error profiles: Uses and abuses. In Statistical Methods and the Improvement of Data Quality. T. Wright (ed.). Academic Press, New York, pp.117-130.

Bailar, Barbara A. (1987). Nonsampling errors. Journal of Official Statistics,. pp.323-326.

Bailey, L., Chapman, D.W. and Kasprzyk, D. (1986). Nonresponse adjustment procedures at the U.S. Bureau of the Census. Survey Methodology 12, pp.161-179.

Bair, R.B. (1981). CONCOR: An edit and automatic correction package. Computer Science and Statistics: Proceedings of the 13th symposium on the Interface, pp.340-343.

Baker, S.G., and Laird, N.M. (1988). Regression analysis for categorial variables with outcome subject to nonignorable nonresponse. Journal of the American Statistical Association. pp.62-69.

Banister, J. (1980). Use and abuse of census editing and imputation. Asian and Pacific Census Forum, 6.

Bankier, M., Luc, M., Nadeau, C. and Newcombe, P. (1995). Imputing Numeric and Qualitative Census Variables Simultaneously, Social Survey Methods Division Report,

Statistics Canada, Dated March 22, 1995.

Banks, Martha J., Andersen, R., and Frankel, M.R. (1983). Total survey error. In Incomplete Data in Sample Surveys. Volume 1, Report and Case Studies. W.G. Madow, H. Nisselson, and I. Olkin, (eds.). Academic Press, New York, pp.391-434.

Barbosa, D.M.R. and Hanono, R.M., Estudo das ferramentas para apuração de dados. Revista Brasileira de Estatística, Rio de Janeiro, 49(191) Jan/Jun 1988,. pp.85-100.

Barcaroli, G. (1990). Project of an integrated system for edit and imputation of data in Italian Statistical Institute. Staff Report, Instituto nazionale di statistica.

Barcaroli, G. (1991). The Automatic Generation of Statistical Incompatibility Rules from E-R Schemes. Technical report from the Italian National Statistical Institute.

Barcaroli, G. (1992) An integrated system for edit and imputation of data in the Italian Statistical Institute, Survey and Statistical Computing, pp.167-177.

Barcaroli, G. (1992). DAISY (Design, Analysis and Imputation System): An integrated system for edit and imputation of data in the Italian National Statistical Institute. Presented at the Conference of European Statisticians' Work Session on Statistical Data Editing, Washington, 1992.

Barcaroli, G., Di Pacel (1992) The automatic generation of statistical incompatibility rules from Entity-Relationship schemes, Proc. of Seminar on New Techniques and Technologies for Statistics, Bonn, February, 1992.

Barcaroli, G., Di Pietro E., **Venturi M.** (1993a) La nuova indagine trimestrale sulle forze di lavoro: aspetti metodologici e analisi dell'impatto delle innovazioni introdotte sulla stima degli aggregati, Politiche del Lavoro n.22-23, Franco Angeli Milano .

Barcaroli, G. (1993). Definition and documentation of data editing rules. Presented at the Conference of European Statisticians' Work Session on Statistical Data Editing, Stockholm, 1993.

Barcaroli, G. and Venturi, M. (1993). An integrated system for edit and imputation of data: An application to the Italian labour force survey. Bulletin of the 49th Session of the International Statistical Institute, Florence, Italy.

Barcaroli G., D'Angiolini G. (1994). Controllo e correzione inter-record, *unpublished internal document, ISTAT Methodological Studies Service*

Barcaroli, G., Ceccarelli, C., Luzi, O., Manzari, A., Riccini, E., Silvestri, F. (1995). The methodology of editing and imputation of qualitative variables implemented in SCIA, *unpublished internal document, ISTAT Methodological Studies Service*

Barcaroli, G. and Venturi, M. (1995). DAISY (Design Analysis and Imputation System):

structure. methodology and first applications. Paper presented at the UN Work Session on Statistical Editing, 6-9 November1995, Athens.

Barcaroli, G. (1996). The Fellegi-Holt methodology for automatic edit and imputation of data: proposals for improvement, Italy.

Barnes, R. (1987). Non-sampling errors: Some approaches adopted in major government surveys in Britain. Journal of Official Statistics, pp.329-333.

Barnett, V. (1983). Principles and methods for handling outliers in data sets. In Statistical Methods and the Improvement of Data Quality. T. Wright (ed.). Academic Press, New York, 117-130.

Barr, J. (1984). Edit Specifications Team Report. Memorandum to Deputy Administrator Raymond R. Hancock, Statistical Reporting Service, U.S. Department of Agriculture.

Bartholomew, D.J. (1961). A method of allowing for 'not at home' bias in sample surveys. Applied Statistics 10, pp.52-59.

Bastelaer von, A., Karssemakers, F., and Sikkel, D. (1988). Data collection with hand-held computers: Contributions to questionnaire design. Journal of Official Statistics.

Beale, E.M.L. and Little, R.J.A. (1975). Missing values in multivariate analysis. Journal of the Royal Statistical Society. Series B. 37, 129-146.

Bean, Jr., E.C. (1988). A Fresh Approach to Information Processing at the U.S. Bureau of the Census. U.S. Bureau of the Census, August 22, 1988.

Becker, R.A., Cleveland, W.S. and Wilks, A.R. (1987) Dynamic Graphics for Data Analysis. Statistical Science, Vol. 2, No. 4, pp.355-395.

Bell, W.R. (1983). A computer program for detecting outliers in time series. Proceedings of the Section of Business and Economic Statistics. American Statistical Association, pp.634-639.

Bemelmans-Spork, M.E.J., and Sikkel, D. (1985b). Data collection with hand-held computers. Bulletin of the International Statistical Institute.

Bemelmans-Spork, E.J. and D. Sikkel (1985a). Observation of prices with hand-held computers. Statistical Journal of the United Nations Economic Commission for Europe, vol. 3, no. 2.

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

Berthelot, J.-M. (1985). Methode de verification statistique pour l'enquete sur le commerce

de gros et le commerce de detail. Document de travail de Statistique Canada.

Berthelot, J.-M. (1987). Data capture and data collection functions: Preliminary edits and follow-up strategies. Statistics Canada Technical Report.

Berthelot, J.-M. (1989). Approche generale pour la sous-fonction de verification et de correction des donnees, version preliminaire. Statistique Canada, rapport technique.

Berthelot, J.-M., Latouche, M., (1990). Strategie de Suivi pour les Enquetes Economiques. Recueil du Symposium 1990 de Statistiques Canada.

Berthelot, J.M. and Latouche, M. (1993). Improving the Efficiency of Data Collection: A Generic Respondent Follow-up Strategy for Economic Surveys, Journal of Business and Economy Statistics, 11:4, pp.417-424

Bethlehem, J.G. and Keller, W.J. (1983). Weighting sample survey data using linear models. Staff Report, Netherlands Central Bureau of Statistics.

Bethlehem, J.G., and Keller, W.J. (1987). Linear weighting of sample survey data. Journal of Official Statistics. pp.141-153.

Bethlehem, J.G. and Keller, W.J. (1989). The Blaise system for computer assisted survey processing. Bulletin of the 47th Session of the ISI, Paris.

Bethlehem, J.G., and Keller, W.J. (1991). The BLAISE system for integrated survey processing. Survey Methodology, pp.43-56.

Bethlehem, J.G., Keller, W.J. (1991). Computer Assisted Statistical Information Processing at the Netherlands Central Bureau of Statistics.

Bethlehem, J.G. and Kersten, H.M.P. (1985). On the treatment of non-response in sample surveys. Journal of Official Statistics 1, pp.287-300.

Bethlehem, J.G. and Kersten, H.M.P. (1986). Werken met Non-respons, Statistische Onderzoekingen M30, Staatsuitgeverij: The Hague.

Bethlehem, J.G., Hundepool, A.J., Schuerhoff , M.H., and Vermeulen,L.F.M. (1989-1993). Blaise version 2 manuals (Automation Department, Statistics Netherlands).

Bethlehem, J.G., Denteneer, D., Hundepool, A.J. and Keller, W.J. (1987). The Blaise system for Computer-Assisted Survey Processing. Proceedings of the Third Annual Research Conference of the Bureau of the Census, Washington, D.C.: U.S. Bureau of the Census, pp.194-203.

Bethlehem, J.G., D. Denteneer, A.J. Hundepool, and W.J. Keller (1987a). BLAISE 1.1/ A first acquaintance, Internal CBS report, Netherlands Central Bureau of Statistics, Voorburg, to appear.

Bethlehem, J.G., Denteneer, D., Hundepool, A.J., Keller, W.J. and Schuerhoff, M.H. (1987). Automating the data editing process with the BLAISE system. Presented at the Conference of European Statisticians Seminar on Statistical Methodology, Geneva.

Bethlehem, J.G. (1987). The data editing research project of the Netherlands Central Bureau of Statistics. Proceedings of the Third Annual Research Conference of the U.S. Bureau of the Census, 194-203.

Bethlehem, J.G. and Hofman, L.P.M.B. (1995). Macro-editing with Blaise III, in V. Kuusela, (ed), Essays on Blaise, Proceedings of the Third Blaise Users' Conference. Statistics Finland, Finland, pp. 1-22.

Bethlehem, J.G. (1997). Control Systems for Computer Assisted Survey Processing, In L. Lyberg, P. Biemer, M. Collins, de Leeuw, Dippo, Schwarz and Trewin (eds), Survey Measurement and Process Quality. Wiley: New York.

Biemer, P.P. (1980). A survey error model which includes edit and imputation error. Proceedings of the Section on Survey Research Methods. American Statistical Association, pp.610-615.

Biemer, P.P. (1988). Modelling matching error and its effect non estimates of census coverage error. Survey Methodology, pp.117-134.

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, November 1995, Working Paper No. 15.

Bilocq, F. (1989). Analysis on grouping of variables and or detection of questionable units. Business Surveys Methods Division, Statistics Canada.

Bilocq, F. and Berthelot, J.-M. (1989a). An editing scheme based on multivariate data analysis. Presented at the annual meeting of the American Statistical Association, Washington, D.C., August 10-14.

Bilocq, F. and Berthelot, J.-M. (1989b). Un schema de verification base sur l'analyse multivariee. Presente au colloque sur les applications statistiques lors du 57eme congres de l'ACFAS, Montreal, 16-19 mai.

Bilocq, F. and Berthelot, J.-M. (1990). Analysis on grouping of variables and on detection of questionable units. Statistics Canada, Methodology Branch Working Paper No. BSMD-005E/F. (Available in English or French).

Binder, D.A. (1984). Some models for non-response and other censoring in sample surveys. Statistics Canada Technical Report.

Bishop, Y.M.M. (1980). Imputation, revision and seasonal adjustment. Proceedings of the Section of Survey Research Methods. American Statistical Association, pp.567-570.

Blair, P.Y. (1978). Technique d'imputation basee sur la regression des donnees d'enquete sur les donees fiscales correspondantes. Statistique Canada, rapport technique.

Blaise - A Survey Processing System : BLAISE III, An Overview - test version. Netherlands Central Bureau of Statistics 1993 (129 pages).

Blom, E. (1993). New technologies in data collection at Statistics Sweden. Presented at the Conference of European Statisticians' Work Session on Statistical Data Editing, Stockholm, 1993.

Blum, O. (1996). Editing an intermediate use file: the 1995 Post enumeration Survey in Israel, Israel.

Bogestrom, B., Larsson, M., Lyberg L. (1983). Bibliography on nonresponse and related topics. In Incomplete Data in Sample Surveys. Volume 2, Theory and Bibliographies. W.G. Madow, I. Olkin, and D.B. Rubin (eds.). Academic Press, New York, pp.479-567.

Bosecker, R.R. (1996). Analytical editing: the NASS interactive data analysis system (IDAS), USA.

Boucaud, W., Dixon, D.P., Michaud, D.B. (1989). The Fields to Impute Problem and Linear Programming, Statistics Canada technical report.

Boucher, L. (1991). Micro-editing for the Annual Survey of Manufactures. What is the Value-added?, Proceedings of the 1991 Annual Research Conference, U.S. Department of Commerce, Bureau of the Census, March 17 - 20, 1991, pp.765-781.

Boucher, L. (1991). Micro-editing for the Annual Survey of Manufactures: What is the value added?, Proceedings of the Bureau of the Census Annual Research Conference, pp.765-781.

Boyes, Barbara A., and Conlon, Margaret E. (1983). The employment cost index: A case study. In Incomplete Data in Sample Surveys. Volume 1, Report and Case Studies. W.G. Madow, H. Nisselson, and I. Olkin, (eds.). Academic Press, New York, pp.123-140.

Brant, J.D., and Chalk, S.M. (1985). The use of automatic editing in the 1981 Census. Journal of the Royal Statistical Society, A, 148, pp.126-146.

Bravo, S. (1990). Editing with MBSPEER. Ministerio de economia y hacienda, Instituto nacional de estadistica, Madrid.

Bravo Cabria, M.S. (1989). SPEER and GEIS Two systems for editing and imputing quantitative data. Instituto Nacional de Estadistica, Spain.

Brewster, K. (1993). Using Blaise for a Customer Satisfaction System. Essays on Blaise 1993: Proceedings of the Second International Blaise Users Conference, Office of Population Censuses and Surveys, London, 1993, pp.1-7.

Bruneau E. (1992) SYNAPSE, serveur de nomenclatures, Le courrier des statistiques n 61-62, June 1992.

BSMD, (1990). An Integrated Approach to Data Editing, Error Correction and Imputation: Summary Report of the Simulation Study. Statistics Canada Technical Report.

Buitenen van, A., Hundepool, A., Jong de, W., and Wetering van de, A. (1993). Electronic dissemination of statistical data. Essays on Blaise 1993: Proceedings of the Second International Blaise Users Conference, Office of Population Censuses and Surveys, London, 1993, pp.8-14.

Bureau, M., Michaud, S. and Sistla, M. (1986). A comparison of different imputation techniques for quantitative data. Statistics Canada, Methodology Branch Working Paper No. BSMD-87-002.

Bureau, M., Michaud, S. and Kovar, J. (1988). Edit and imputation of tax data. Proceeding of the Section on Business and Economic Statistic. American Statistical Association, pp.372-375.

Burg T. (1995). Current Status of Automated Coding in the Austrian Central Statistical Office, EUROSTAT Workshop on Census Processing, Fareham 1995

Burns, E.M. (1983). Editing and imputation for the EIA weekly petroleum surveys. Proceedings of the Section on Survey Research Methods. American Statistical Association, pp.539-543.

Burns, E.M. (1990). Multiple and replicate item imputation in a complex sample survey. Proceedings of the U.S. Bureau of the Census 1990 Annual Research Conference, pp.655-665.

Burns, E.M. and Goldberg, M.L. (1988). Imputation strategies for complex sample survey with many variables. Presented at the Conference of European Statisticians Seminar on Statistical Methodology, Geneva.

Burns, E.M. (1980). Procedures for the detection of outliers in weekly time series. Proceedings of the Section of Business and Economic Statistics. American Statistical Association, pp.560-563.

Bushnell, Diane (1995). Computer Assisted Occupation Coding, Office of Population Censuses and Surveys, UK

California Scientific Software (1993). BrainMaker Professional User's Guide and Reference Manual. 4th Edition. Nevada City, Ca.

Cassel, C.M., Sarndal, C-E., and Wretman, J.H. (1983). Some uses of statistical models in connection with the nonresponse problem. In Incomplete Data in Sample Surveys. Volume 3, Proceedings of the Symposium. W.G. Madow, and I. Olkin (eds.). Academic Press, New

York, pp.143-160.

Catlin, G., Ingram, S. and Hunter, L. (1988). The Effects of CATI on Data Quality: A Comparison of CATI and Paper Methods, Proceedings of the Bureau of the Census Annual Research Conference, pp.291-299.

CBS (1993). Blaise III, An Overview, Netherlands Central Bureau of Statistics.

CBS (1994). Samenvattend overzicht van de industrie 1992 [Summary of Manufacturing, 1992] (SDU/CBS-publications, The Hague).

Census of Agriculture (1981a). 1981 imputation specifications: Appendix I. Statistics Canada Technical Report.

Census of Agriculture (1981b). 1981 imputation specifications: Appendix L. Statistics Canada Technical Report.

Chambers, J.M., Cleveland, W.S., Kleiner, B., and Tukey, P.A. (1983). Graphical Methods for Data Analysis. Duxburry Press: Boston, USA.

Chapman, D.W., and Weinstein, R.B. (1990). Sampling design for a monitoring plan for CATI interviewing. Journal of Official Statistics, pp.205-211.

Chapman, D.W. (1976). A survey of non-response imputation procedures. Proceedings of the Social Statistics Section, American Statistical Association, pp.245-251.

Chapman, D.W. (1983a). An investigation of nonresponse imputation procedures for the health and nutrition examination survey. In Incomplete Data in Sample Surveys. Volume 1, Report and Case Studies. W.G. Madow, H. Nisselson, and I. Olkin, (eds.). Academic Press, New York, pp.435-484.

Chapman, D.W. (1983b). The impact of substitution on survey estimates. In Incomplete Data in Sample Surveys. Volume 2, Theory and Bibliographies. W.G. Madow, I. Olkin, and D. B. Rubin (eds.). Academic Press, New York, pp.45-62.

Cheng B and D. M. Titterington,(1994). Neural Networks: A review from a statistical perspective. Statistical Science 9, pp.2-54.

Cheng, B. and Titterington, D.M. (1994). Neural Networks: A review from a Statistical Perspective. Statistical Science, Vol. 9, No. 1, pp.3-54.

Chernick, M.R. (1982). The influence function and its application to data validation. American Journal of Mathematical and Management Sciences, pp.263-288.

Chernick, M.R. (1983). Influence functions, outlier detection, and data editing. In Statistical Methods and the Improvement of Data Quality. T. Wright (ed.). Academic Press, New York, pp.167-176.

Chernick, M.R., Downing, D.J., and Pike, D.H. (1982). Detecting outliers in time series data. Journal of the American Statistical Association, 77, pp.743-747.

Chernick, M.R., and Murthy, V.K. (1983). The use of influence functions for outlier detection and data editing. American Journal of Mathematical and Management Sciences, 3, pp.47-61.

Chernikova, N.V. (1964). Algorithm for finding a general formula for the nonnegative solutions of a system of linear equations. U.S.S.R. Computational Mathematics and Mathematical Physics 4, pp.151-158.

Chernikova, N.V. (1965). Algorithm for finding a general formula for the nonnegative solution of a system of linear inequalities. U.S.S.R. Computational Mathematics and Mathematical Physics 5, pp.228-233.

Cheung, S. and Seko, C. (1986). A study of the effects of imputation groups in the nearest neighbour imputation method for the National Farm Survey. Survey Methodology 12, pp.99-106.

Chiu, H.Y. and Sedransk, J. (1986). A Bayesian procedure for imputing missing values in sample surveys. Journal of the American Statistical Association 81, 667-676.

Christianson, A. and Tortora, R.D. (1993). Issues in surveying establishments. In the Monograph of the International Conference on Establishment Surveys, Buffalo, NY, June 28-30. To be published by Wiley in 1994.

Chvatal, V. (1983). Finding all vertices of a polyhedron. Linear Programming, W.H. Freeman and Company, New York, pp.271-287.

Cimermanović, Branka (1992). Computer assisted data editing: A comparative study. Proceedings of the 14th International Technology Interfaces (ITI'92), Pula, Croatia, 1992, pp.437-442.

Ciok, R. (1991). The Use of Automated Coding in the 1991 Canadian Census of Population, Paper presented at the 1991 Annual Meeting of the American Statistical Association, Atlanta, Georgia

Ciok, R. (1992). Spider - Census Edit and Imputation System, Social Survey Methods Division Report, Statistics Canada, Dated September 1992.

Ciok, R. (1993). The results of automated coding in the 1991 Canadian Census of Population. Paper presented at the 1993 Annual Research Conference, organized by the US Bureau of the Census.

Clark, R.G. (1995). Winsorization methods in sample surveys. Masters thesis, Department of Statistics, Australian National University.

Clark, A. and Street, L. (1996). Planning for the 2001 Census of the United Kingdom. Presentation for the US Bureau of the Census Annual Research Conference 1996. Washington DC.

Clayton, Richard L. (1994). Electronic Data Interchange: Automated Data Collection for the Next Decade, Proceedings of the First Eurostat Conference on Panel Surveys, February 1994.

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

Cochran, W.G. (1983). Historical perspective. In Incomplete Data in Sample Surveys. Volume 2, Theory and Bibliographies. W.G. Madow, I. Olkin, and D.B. Rubin (eds.). Academic Press, New York, pp.11-28.

Coder, J., Feldman, A. and Nelson, C. (1978). Reporting of quarterly earnings amounts in the income survey development program site research sample. Presented at the Annual Meeting of the American Statistical Association.

Colledge, M.J., Johnson, J.H., Pare, R. and Sande, I.G. (1978). Large scale imputation of survey data. Survey Methodology 4, pp.203-224.

Cook, L.W. (1993). Graphical editing: Direct and indirect benefits. Bulletin of the 49th session of the International Statistical Institute, Florence, Italy.

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

Corby, C. (1987). 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.

Cotton, C. (1991). Functional Description of the Generalized Edit and Imputation System, Business Survey Methods Division Report, Statistics Canada, Dated July 25, 1991.

Cotton, C. (1991). "Generalized Edit and Imputation System Functional Description," Statistics Canada Technical Report.

Cotton, P. (1988). A Comparison of software for editing survey and census data. Proceedings of Symposium 88, The Impact of High Technology on Survey Taking, Ottawa, Ontario, Canada: Statistics Canada, pp.211-241.

Coulter, J. (1982). A proposed comparison of hot-deck imputation and weighting as methods of correcting for non-response. Statistics Canada Technical Report.

Couper, M., Groves, R.M. (1990). Interviewer expectations regarding CAPI: results of laboratory tests II. Report to Bureau of Labour Statistics.

Cox, B.G. (1980). The weighted sequential hot deck imputation procedure. Proceedings of the Section on Survey Research Methods. American Statistical Association, pp.721-726.

Cox, B.G. and Cohen, S.B. (1985). Imputation procedures to compensate for missing responses to data items. In Methodological Issues for Health Care Surveys, Marcell Dekker, New York.

Cox, B.G. and Folsom, R.E. (1978). An empirical investigation of alternative item non-response adjustments. Proceedings of the Section on Survey Research Methods. American Statistical Association, pp.219-223.

Cox, B.G. and Folsom, R.E. (1981). An evaluation of weighted hot-deck imputation for unreported health care visits. Proceedings of the Section on Survey Research Methods. American Statistical Association, pp.412-417.

Cox, N.W.P., and Croot, D.A. (1994). Data editing in a mixed DBMS environment. Statistical Data Editing Methods and Techniques, Volume No. 1, UN Economic Commission for Europe, Statistical Division, Conference of European Statisticians, 1994.

Cox, N.W.P. (1991). Data Editing in a Mixed DBMS Environment: Report to the joint group on data editing. Statistics Canada Technical Report.

Cox, N. (1986). Generalized edit and imputation system. Technical Report, Statistics Canada.

Cox, N., and Kovar, J. (1987). Generalized edit and imputation system (GEIS). Presentation from Statistics Canada.

Criado, I.V., Bravo Cabria, M.S. (1990). Procedimento de Depuracion de Datos Estadisticos. Euskal Estatistika Erakundea Instituto Vasco de Estadistica (EUSTAT), Spain.

CRIPTAX - Users Manual - IBGE

Cruddas, M. and Kokic, P. (1996). The treatment of outliers in ONS business surveys. Proceedings of the GSS(M) methodology conference. ONS, Newport.

Cruddas, M., Thomas, J., Thorogood, D. (1996). Editing and imputation research for the 2001 Census in the United Kingdom, UK.

Cumberbirch, P. (1979). Donor record selection procedure for "hot-deck" imputation in the Census of Agriculture. Statistics Canada Technical Report.

Cushing, J. (1988). A report on recent survey processing in developing countries: The demographic and health surveys microcomputer approach. Proceedings of Symposium 88, The Impact of High Technology on Survey Taking, Ottawa, Ontario, Canada: Statistics Canada, pp.201-210.

Czajka, J.L. (1987b). Turning the tables: Imputing for item non-response when donors are scarce. Presented at the Fourth Statistics Canada International Symposium on the Statistical Uses of Administrative Data, November 23-25,1987.

Czajka, J.L. (1987a). Predicting edit outcomes: The strategic use of imputation in estimating corporate income statistics. Proceedings of the Section on Survey Research Methods. American Statistical Association, pp.312-317.

Daffin, C., Duncombe, P., Hills, Mary G., North, P.M., and Wetherill G.B. (1986). A microcomputer survey analysis package designed for the developing countries. The Statistician. pp.505-523.

Dalenius, T. (1977a). Bibliography on nonsampling errors in surveys-I, International Statistical Review,. pp.71-85.

Dalenius, T. (1977b). Bibliography on nonsampling errors in surveys-II, International Statistical Review, pp.181-197.

Dalenius, T. (1977c). Bibliography on nonsampling errors in surveys-III, International Statistical Review, pp.303-317.

Dalenius, T. (1983a). Errors and other limitations of surveys. In Statistical Methods and the Improvement of Data Quality. T. Wright (ed.). Academic Press, New York, pp.1-24.

Dalenius, T. (1983b). Informed consent or R. S. V. P. In Incomplete Data in Sample Surveys. Volume 3, Proceedings of the Symposium. W. G. Madow, and I. Olkin (eds.). Academic Press, New York, pp.85-106.

Dalenius, T. (1987). Error control in surveys: Some informal remarks, Journal of Official Statistics, pp.327-328.

Daoust, P. (1984). Etude en rapport avec les variables selectives utilisees lors de l'etape d'imputation du recensement de l'agriculture du Canada de 1981. Statistiques Canada, Document du Travail.

Data Editing Joint Group (1993). Data Editing System Guidelines for Concepts and Specifications: Glossary of Terms on The Statistical Computing Project (ECE/ UNDP/ SCP/ H.2 and New terms proposed to be added to the ECE/ UNDP/ SCP/ H.2 Glossary). Room paper presented at the Conference of European Statisticians' Work Session on Statistical Data Editing, Stockholm, 1993.

Data Editing Joint Group (1992). Data Editing System Guidelines for Concepts and Specification. Presented at the Meeting of Joint Group on Data Editing, Washington, 1992.

Data Editing Group, FCSM (1992). 1990-1992 Data Editing Experience in U.S. Federal Agencies.Presented at the Meeting of Joint Group on Data Editing, Washington, 1992.

Data Editing Joint Group (1991). Economic Commission for Europe. Statistical Computing Project Phase 2: List of Products (revision 2).

David, M., Little, R.J.A., Samuhel, M.E., and Tries, R.K. (1986). Alternative methods for CPS income imputation. Journal of the American Statistical Association, pp.29-41.

David, M. and Triest, R. (1983). The CPS hot-deck: An evaluation using IRS records. Proceedings of the Section on Survey Research Methods. American Statistical Association, pp.421-426.

Davila, E.H. (1994). Macro-editing -- The Hidiroglou-Berthelod method. Statistical Data Editing Methods and Techniques, Volume No. 1, UN Economic Commission for Europe, Statistical Division, Conference of European Statisticians, 1994.

De Jong, P. (1996). Designing a complete edit strategy, Netherlands.

De Waal, A.G. (1995). Developing an edit and imputation system. Paper presented at the UN Work Session on Statistical Editing, 6-9 November, Athens.

De Waal, T., (1995). Developing an edit and imputation system. Report, Statistics Netherlands, Voorburg.

De Waal, A.G. (1996). CHERRYPI: A computer program for automatic edit and imputation, paper presented at the UN Work Session on Statistical Data Editing, 4-7 November, 1996, Voorburg, the Netherlands.

Degerdal, H., Hoel, T., Thirud, T. (1995). The CAI system of Statistics Norway, Statistics Norway.

Deming, W.E. (1953). On a probability mechanism to attain an economic balance between the resultant error of nonresponse and the bias of nonresponse. Journal of the American Statistical Association.

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

Denteneer, D., Bethlehem, J.G., Hundepool, A.J., and Schuerhoff (1994). M. S. (1992). Blaise - A new approach to computer assisted survey processing. Statistical Data Editing Methods and Techniques, Volume No. 1, UN Economic Commission for Europe, Statistical Division, Conference of European Statisticians, 1994.

Denteneer, D. Bethlehem, J.G., Hundepool, A.J., and Schuerhoff, M.S. (1987). The Blaise system for computer assisted survey processing. Proceedings of the Third Annual Research Conference of the Bureau of the Census, Washington, D.C.: U.S. Bureau of the Census, pp.112-127.

Deville, J-C. and Särndal, C-E. (1994). Variance estimation for the regression imputed Horvitz-Thompson estimator. *Journal of Official Statistics, **10*** (4), pp.381-394.

Diederen, B., and P. Michels, in preparation. Het gebruik van tijdreeksen bij het controle-, correctie- en publikatie-proces na het herontwerp van de statistieken van de internationale handel; methodologie en eerste experimentele resultaten. (internal report, Statistics Netherlands, Heerlen).

Dillman, D.A. (1978). Mail and Telephone Surveys: The Total Design Method, New York, Wiley-Interscience.

Dinh, K.T. (1987). Application of the spectral analysis in editing a large data base. Journal of Official Statistics , 3, pp.431-438.

Dixon, D.P., Todor, C.L. (undated). Vertex Generation and Chernikova's Algorithm.

Dolson, D. (1994). On Using a Very Current Source of Frame Information in Canada's Establishment Based Employment Survey, Paper presented at the First Eurostat Workshop on Techniques of Enterprise Panel, Luxembourg, February 21-23,1994.

Donda, A. (1989). Relations between the mode of data capture and the quality of surveys. Bulletin of the 47th Session of the ISI, Paris.

Doucet, J.E. General Survey Function Design, Informatics Overview and Update. Internal Working Paper, Statistics Canada, April 1990.

Dowling, T.A., and Shachtman, R.H. (1975). On the relative efficiency of randomized response models. Journal of the American Statistical Association.

Downer, R. (1990). An integrated approach to data editing, error correction and imputation: detailed report of the simulation study. Statistics Canada, Methodology Branch Working Paper.

Draper, L ., Greenberg, B., & Petkunas, T. (1990). On-line capabilities in SPEER (Structured Programs for Economic Editing and Referrals). Proceedings of Statistics Canada Symposium 90: Measurement and Improvement of Data Quality, pp.235-44. Ottawa: Statistics Canada.

Drew, J.D. (1991). Research and testing of telephone survey methods at Statistics Canada. Survey Methodology, pp.57-68.

Drew, J.H. and Fuller, W.A. (1980). Modelling nonresponse in surveys with callbacks. Proceedings of the Section on Survey Research Methods. American Statistical Association, pp.639-642.

Drew, J.H. and Fuller, W.A. (1981). Nonresponse in complex multiphase surveys. Proceedings of the Section on Survey Research Methods. American Statistical Association,

pp.623-628.

Dumičić, S., Kecman, Nataša and Dumičić, Ksenija (1992). An Implementation of Sampling Method on Optical Reading Control in the Census 1991. Proceedings of the 14th International Conference "Information Technology Interfaces" (ITI'92). Pula, 1992.

Dumičić, S. (1992a). One Solution to Data Editing: IBM Mainframe. Data Editing System Guidelines for Concepts and Specifications. Joint Group on Data Editing, Washington, 1992.

Dumičić, S. (1992b). The Automatic Deterministic Data Correction Methods. Data Editing System Guidelines for Concepts and Specifications. Joint Group on Data Editing, Washington, 1992.

Dupont, F., and Rivière, P. (1993). Imputation procedures for qualitative variables and hierarchical qualitative variables. Presented at the Conference of European Statisticians' Work Session on Statistical Data Editing, Stockholm, 1993.

Dussert, F., Luciani G. (1995). CAPI PLUS et Blaise III: Une organisation génèrale pour les enquêtes de l'INSEE, INSEE, France.

Economic Commission for Europe, (1991). Evaluation Criteria for Software on Data Editing, SCP-2, United Nations ECE Statistical Division, Geneva, 1991.

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

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

Economic Commission for Europe, UNDP, CSP/H.2. Glossary of Terms on The Statistical Computing Project. United Nations,

Ekholm, A., and Palmgren, J. (1987). Correction for misclassification using doubly sampled data. Journal of Official Statistics, pp.419-430.

Elias, P. (1992). Computer-assisted and computer-automated coding of low quality occupation information. Presented at the Conference of European Statisticians' Work Session on Statistical Data Editing, Washington, 1992.

Elvers, Eva (1993). Imputations in the Swedish Industrial Survey utilizing auxiliary information. Presented at the Conference of European Statisticians' Work Session on Statistical Data Editing, Stockholm, 1993.

Engström, P. (1995). A study on using selective editing in the Swedish survey on wages and employment in industry. *Conference of European Statisticians, Work Session on Statistical Data Editing, room paper No. 11, November 6-9, 1995, Athens Greece.*

Engström, P. (1996). Monitoring the editing process, Sweden.

Engström, P. and Ängsved, C. (1994). A Description of a Geographical Macro-editing Application, Conference of European Statisticians, Work Session on Data Editing, Cork, Ireland.

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

Ericksen, E.P., and Kanade, J.B. (1985). Estimating the population in a census year: 1980 and beyond. Journal of the American Statistical Association.

Ericksen, E.P., and Kanade, J.B. (1986). Using administrative lists to estimate census omissions. Journal of Official Statistics.

Ernst, L.R. (1980). Variance of the estimated mean for several imputation procedures. Proceedings of the Section on Survey Research Methods. American Statistical Association, 716-720.

Ernst, L.R. (1978). Weighting to adjust for partial nonresponse. Proceedings of the Section on Survey Research Methods. American Statistical Association, pp.468-473.

Esposito, R., Lin, D., and Tidemann (1993). "The ARIES Review System in the BLS Current Employment Statistics Program," ICES Proceedings of the International Conference on Establishment Surveys, June 27-30, 1993, Buffalo, New York.

Esposito, R., Fox, J.K., Lin, D., and Tidemann, K. (1994). ARIES: a Visual Path in the Investigation of Statistical Data, Journal of Computational and Graphical Statistics 3, pp. pp.113-125.

Estevao, V. (1988). Donor Imputation Specifications. Statistics Canada Technical Paper.

Eurostat and Statistics Denmark (1995). Statistics on persons in Denmark - a register based statistical system, statistical document.

Fay, R.E. (1986). Causal models for patterns of nonresponse. Journal of the American Statistical Association, pp.354-365.

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

Federal Committee on Statistical Methodology (1993). Survey nonresponse in the federal statistical system. Preliminary Committee report.

Fellegi, I.P. (1975). Automatic editing and imputation of quantitative data. Statistics Canada

Technical Report.

Fellegi, I.P. and Holt, D. (1976). A Systematic Approach to Automatic Editing and Imputation. Journal of the American Statistical Association, March 1976, Vol.71, No. 353, pp. 17-35.

Fellegi, I.P. (1988). Data, statistics, information: Some issues of the Canadian social statistics scene. Journal of Official Statistics.

Fellegi, I.P. (1991). Maintaining public confidence in official statistics. Journal of the Royal Statistical Society, Ser. A. 1-6.

Ferguson, Dania P. (1987). Why a New New Edit System Internal Memorandum, National Agricultural Statistics Service, U.S. Department of Agriculture. Presented at the 1987 June Enumerative Survey school.

Ferguson, Dania P. (1991). SAS use in data editing. Statistical Journal of the United Nations Economic Commission for Europe, Vol. 8, No. 2, (Special Issue on Data Editing), pp.167-174.

Ferguson, Dania P. (1993). Review of methods and software used in data editing, second edition (proposed new title: An introduction to the data editing process). Presented at the Conference of European Statisticians' Work Session on Statistical Data Editing, Stockholm, 1993.

Ferguson, Dania P. (1994a). Introduction to the data editing process. Statistical Data Editing Methods and Techniques, Volume No. 1, UN Economic Commission for Europe, Statistical Division, Conference of European Statisticians, 1994.

Ferguson, Dania P. (1994b). SAS usage in data editing. Statistical Data Editing Methods and Techniques, Volume No. 1, UN Economic Commission for Europe, Statistical Division, Conference of European Statisticians, 1994.

Fernandez, F. and Quintan, P. (1988). Extension of Chernikova's Algorithm for solving general mixed linear programming problems. Institut national de recherche en informatique et en automatique, Rapport de recherche, No. 943.

Ferrari, P. and Bailey, L. (1981). Preliminary results of the 1980 decennial census telephone follow-ups of nonresponse experiments. Proceedings of the Section on Survey Research Methods. American Statistical Association, pp.264-269.

Fillion, J.M. and I. Schiopu-Kratina, On the use of Chernikova's algorithm for error localisation. Statistics Canada.

Folsom, R.E. (1981). The equivalence of generalized double sampling regression estimators, weight adjustments and randomized hot-deck imputations. Proceedings of the Section on Survey Research Methods. American Statistical Association, pp.400-405.

Ford, B.L., Kleweno, D.G. and Tortora, R.D. (1980). The effects of procedures which impute for missing items: A simulation study using an agricultural survey. Proceedings of the Section on Survey Research Methods. American Statistical Association, pp.251-256.

Ford, B.L. (1976). Missing data procedures: A comparative study. Proceedings of the Social Statistics Section. American Statistical Association. pp.324-329.

Ford, B.L. (1978). A general overview of the missing data problem. Statistical Research Division, U.S. Department of Agriculture, Washington, D.C.

Ford, B.L. (1983). An overview of hot-deck procedures. In Incomplete Data in Sample Surveys. Volume 2, Theory and Bibliographies. W.G. Madow, I. Olkin, and D.B. Rubin (eds.). Academic Press, New York, pp.185-209.

Forsman, G. (1991). Guide for Evaluation of the Editing Process (in Swedish), Editing Memo 24, 1991

Frane, J.W. (1976). Some simple procedures for handling missing data in multivariate analysis. Psychometrika. pp.409-415.

Frankel, L.R., and Dutka, S. (1983). Survey design in anticipation of nonresponse and imputation. In Incomplete Data in Sample Surveys. Volume 3, Proceedings of the Symposium. W.G. Madow, and I. Olkin (eds.). Academic Press, New York, pp.69-84.

Freie, R.L. (1983). USDA Livestock inventory surveys. In Incomplete Data in Sample Surveys. Volume 1, Report and Case Studies. W.G. Madow, H. Nisselson, and I. Olkin, (eds.). Academic Press, New York, pp.141-172.

Freund, R.J., and Hartley, H.O. (1967). A procedure for automatic data editing. Journal of the American Statistical Association, 62, pp.341-352.

Friedman, J.H., Bentley, J.L. and Finkel, R.A. (1977). An algorithm for finding best matches in logarithmic expected time. ACM Transaction on Mathematical Software 3, pp.209-226.

Fuchs, C. (1982). Maximum likelihood estimation and model selection in contingency tables with missing data. Journal of the American Statistical Association. pp.270-278.

Fuller, W.A. (1987). Measurement Error Models. Wiley, New York.

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

Garås, T. (1993). Evaluation of the Editing Process in the Annual Statistics of Government Employees (in Swedish), Statistics Sweden, Bakgrundsfakta till arbetsmarknads- och utbildningsstatistiken, 1993

Garcia-Rubio, E., and Peirats, V. (1992). Evaluation of data editing procedures: Results of a simulation approach. Statistical Data Editing Methods and Techniques, Volume No. 1, UN Economic Commission for Europe, Statistical Division, Conference of European Statisticians, February 1992, pp.92-116.

Garcia-Rubio, E., and Peirats, V. (1991). Evaluation of data editing procedures - Results of simulation approach. Statistical Journal of the United Nations Economic Commission for Europe, Vol. 8, No. 2, (Special Issue on Data Editing), pp.175-190.

Garcia-Rubio, E., and Villan, I. (1988). The DIA System, (An Automatic Editing and Imputation System), Volume I, DIA System Description. National Statistical Institute, Madrid.

Garcia-Rubio, E. and Villan, I. (1990). DIA System: Software for the automatic imputation of qualitative data. Proceedings of the U.S. Bureau of the Census 1990 Annual Research Conference, pp.525-537.

Garfinkel, R.S. (1979). An Algorithm for optimal imputation of erroneous data, College of Business Administration Working Paper Series, The University of Tennessee, Knoxville.

Garfinkel, R.S., Kunnathur, A.S. and Liepins, G.E. (1986). Optimal imputation of erroneous data: Categorical data, general edits. Operations Research 34, pp.744-751.

Garfinkel, R.S., Kunnathur, A.S. and Liepins, G.E. (1988). Error localization for erroneous data: Continuous data, linear constraints. SIAM Journal of Scientific Statistical Computing, pp.922-931.

Gary Houston, Andrew G. Bruce (1993). GRED: Interactive Graphical Editing for Business Surveys, Journal of Official Statistics, pp.81-90

GEIS Development Team (1990). Generalized Edit and Imputation System specifications. Statistics Canada Technical Report.

Giles, P. and Patrick, C. (1986). Imputation options in a generalized edit and imputation system. Survey Methodology 12, pp.49-60.

Giles, P. (1985a). PSTAT E&I system for numeric data: Details on determining the "closest" donor. Statistics Canada Technical Report.

Giles, P. (1985b). Effect of choice of distance function on forming candidate-donor pairs for imputation. Statistics Canada Technical Report.

Giles, P. (1985c). Dealing with a combination of qualitative and quantitative variables in an imputation system. Statistics Canada Technical Report.

Giles, P. (1985d). Comparison of one tree versus many trees: PSTAT edit and imputation system. Statistics Canada Technical Report.

Giles, P. (1986a). "Generalized" edit and imputation. Statistics Canada Technical Report.

Giles, P. (1986b). "Generalized" edit and imputation - Part II. Statistics Canada Technical Report.

Giles, P. (1986c). Methodological specifications for a generalized edit and imputation system. Statistics Canada Technical Report.

Giles, P. (1986d). Checking of equality edits. Statistics Canada Technical Report.

Giles, P. (1987). Towards the development of a generalized edit and imputation system. Proceedings of the Census Bureau Third Annual Research Conference of the Bureau of the Census, Washington, D.C., pp.185-193.

Giles, P. (1988). A model for generalized edit and imputation of survey data. Special Issue of the Canadian Journal of Statistics 16, Supplement, pp.57-73.

Giles, P. (1989). Analysis of edits in a generalized edit and imputation system. Statistics Canada, Methodology Branch Working Paper No. SSMD-89-004E.

Gillam, S. (1991). Statistical Data Drawn From Administrative Systems - Experiences With Social Security in Great Britain. Presented at the Conference of European Statisticians' Work Session on Statistical Data Editing, Geneva, 1991.

Gissel, Katherine C., Wray, Martha L., and Hansard, Martha S. (1983). Health and Mortality Study error detection, reporting, and resolution system. In Statistical Methods and the Improvement of Data Quality. T. Wright (ed.). Academic Press, New York, pp.321-353.

Gogic, G., Egersdorfer, D., and Pesic, R. (1989). GoDar - Program product for automation of statistical surveys. Selected Papers from The Sixth Meeting of Data editing Joint Group, Madrid, 1990.

Gogic, G., Egersdorfer, D. (1992). Interactive Checking and Correction. Data Editing System Guidelines for Concepts and Specifications.

Gonzales, M.E. et al. (1975). Standards for discussion and presentation of errors in survey and census data. Journal of the American Statistical Association.

Goodger, Chris (1995). Training Interviewers in the Use of Blaise, Office of Population Censuses and Surveys, UK.

Gosselin, J.-F., Chinnappa, B.N., Ghangurde, P.D. and Tourigny, J. (1978). A Compendium of Methods of Error Evaluation in Censuses and Surveys, Ottawa: Statistics Canada.

Gower, A. (1979). Non-response in the Canadian Labour Force Survey. Survey Methodology 5, pp.29-58.

Granquist, L. (1982). On generalized editing programs and solution of the data quality problems, UNDP Statistical Computing Project, Data Editing Joint Group.

Granquist, L. (1983). On the Role of Editing. SCP/DE/SP.39.

Granquist, L. (1984a). On the role of editing. Statistic Tidskrift 2, pp.105-118.

Granquist, L. (1984b). Data editing and it's impact on the further processing of statistical data. Presented at the Workshop on the SCP in Budapest, November 12-17, 1984.

Granquist, L. (1987). A report of the main features of a macro-editing procedure which is used in Statistics Sweden for detecting errors in individual observations. Presented at the Data Editing Joint Group Meeting in Madrid, April 22-24,1987.

Granquist, L. (1988). A Report of the Main Features of a Macro-editing Idea Applied on the Monthly Survey on Employment and Wages in Mining, Quarrying and Manufacturing. Report presented at the Data Editing Joint Group Meeting in Budapest, April 18-22, 1988. Statistics Sweden.

Granquist, L. (1990). A review of some macro-editing methods for rationalizing the editing process. Proceedings of Statistics Canada Symposium 90, Measurement and Improvement of Data Quality, pp.225-34. Ottawa: Statistics Canada.

Granquist, L. (1990a). Status report on the Swedish data editing project. Selected Papers from The Sixth Meeting of Data Editing Joint Group, Madrid, 1990.

Granquist, L. (1990b). Contents of lectures on data editing and quality. Presented at the Workshop on Data Editing and Imputation Methods in Rio de Janeiro, February 12-16,1990.

Granquist, L. (1990c). Evaluations/Experiences. Presented at the Workshop on Data Editing and Imputations Methods in Rio de Janeiro, February 12-16,1990.

Granquist, L. (1990d). Definitions on editing and imputation and the traditional microediting procedure. Presented at the Workshop on Data Editing and Imputation Methods in Rio de Janeiro, February 12-16,1990.

Granquist, L. (1990e). Generalized numeric edit and imputation system. Presented at the Workshop on Data Editing and Imputation Methods in Rio de Janeiro, February 12-16,1990.

Granquist, L. (1990f). Macro-editing methods: The Hidroglou-Berthelot method (Statistical Edits). Presented at the Workshop on Data Editing and Imputation Methods in Rio de Janeiro, February 12-16,1990.

Granquist, L. (1990g). A review of some macro-editing methods for rationalizing the editing process. Proceedings of Statistics Canada Symposium 90, Measurement and Improvement of Data Quality, pp.225-34. Ottawa: Statistics Canada.

Granquist, L. (1991). Progress Report on the Swedish Data Editing Project. Presented at the Conference for European Statisticians' Work Session on Statistical Data Editing, Geneva, 1991.

Granquist, L. (1992). A Review of Studies on Impact of Data Editing on Estimates and Quality. UN-ECE Joint Group on Data Editing. Washington, 1992.

Granquist, L. (1993a). International review of research on data editing strategies. Bulletin of the 49th session of the International Statistical Institute, Florence, Italy.

Granquist, L. (1993b). Improving the Traditional Editing Process. In the Monograph of the International Conference on Establishment Surveys, Buffalo, NY, June 28-30,1993. To be published by Wiley in 1994.

Granquist, L. (1993c). Data editing activities in Statistics Sweden. Presented at the Conference of European Statisticians' Work Session on Statistical Data Editing, Stockholm, 1993.

Granquist, L. (1994). Improving the Traditional Editing Process, in Business Survey Methods, B.G.Cox, D.A.Binder, N.Chinnappa, A.Christianson, M.J.Colledge and P.S.Kott (eds.), New York: Wiley, pp.385-401.

Granquist, L. (1994a). On the need for generalized numeric and imputation systems. Statistical Data Editing Methods and Techniques, Volume No. 1, UN Economic Commission for Europe, Statistical Division, Conference of European Statisticians, 1994.

Granquist, L. (1994b). Macro-editing - A review of methods for rationalizing the editing of survey data. Statistical Data Editing Methods and Techniques, Volume No. 1, UN Economic Commission for Europe, Statistical Division, Conference of European Statisticians.

Granquist, L. (1994c). Macro-editing - The aggregate method. Statistical Data Editing Methods and Techniques, Volume No. 1, UN Economic Commission for Europe, Statistical Division, Conference of European Statisticians, 1994.

Granquist, L. (1994d). Macro-editing - The top-down method. Statistical Data Editing Methods and Techniques, Volume No. 1, UN Economic Commission for Europe, Statistical Division, Conference of European Statisticians, 1994.

Granquist, L., (1995). Improving the traditional editing process. In: Cox, Binder, Chinnappa, Christianson, Colledge and Knott, eds., Business Survey Methods (John Wiley, New York), pp.385-401.

Granquist, L. (1996). The new view on editing, Sweden.

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

appear.

Graves, R.B. (1976). CAN-EDIT: A generalized edit and imputation system on a data base environment. Presented at the Electronic Data Processing Working Party of the Conference of European Statisticians, Geneva.

Gray, James (1995). An Object-Based Approach for the Handling of Survey Data, Office of Population Censuses and Surveys, UK.

Gray, J. (1993). Outputs from Blaise surveys. Essays on Blaise 1993: Proceedings of the Second International Blaise Users Conference, Office of Population Censuses and Surveys, London, 1993, pp.15-26.

Greenberg, B. (1981). Developing an edit system for industry statistics. Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface, New York: Springer-Verlag, pp.11-16.

Greenberg, B. (1982a). Issues in editing continuous data. Regional Meeting of the Southeastern Chapter of the Institute of Management Sciences, Myrtle Beach, California.

Greenberg, B. (1982b). Examples illustrating the need for implied edits for continuous data. Internal Working Paper, U.S. Bureau of the Census.

Greenberg, B. (1982c). Using an edit system to develop editing specifications. Proceedings of the Section on Survey Research Methods. American Statistical Association, pp.366-371.

Greenberg, B. (1982d). Discussion of the paper: Imputing for missing survey responses by G. Kalton and D. Kasprzyk. Proceedings of the Section on Survey Research Methods. American Statistical Association, pp.32-33.

Greenberg, B. (1985). Example illustrating the need for implied edits for categorical data. Internal Working Paper, U.S. Bureau of the Census.

Greenberg, B. (1986a). An evaluation of edit and imputation procedures used in the 1982 Economic Censuses in Business Division. Statistical Research Division Report Series, U.S. Bureau of the Census. CENSUS/SRD/RR-86-06.

Greenberg, B. (1986). The use of implied edits and set covering in automated data editing. Statistical Research Division Report Series, U.S. Bureau of the Census. CENSUS/SRD/RR-86-02.

Greenberg, B. (1987a). Discussion of the papers: Towards the development of a generalized edit and imputation system by P. Giles; and The data editing research project of the Netherlands Central Bureau of Statistics by J. Bethlehem. Proceedings of the U.S. Bureau of the Census Third Annual Research Conference, 204-210.

Greenberg, B. (1987b). Edit and imputation as an expert system. Statistical Policy Working

Paper Number 14: Statistical Uses of Microcomputers in Federal Agencies, Washington, D.C.: Office of Management and Budget, pp.85-92.

Greenberg, B. (1987c). Discussion, session on designing automated data editing systems. Proceedings of the Third Annual Research Conference of the Bureau of the Census, Washington, D.C.: U.S. Bureau of the Census, pp.204-212.

Greenberg, B. and Surdi, R. (1984). A flexible and interactive edit and imputation system for ratio edits. Proceedings of the Section on Survey Research Methods. American Statistical Association, pp.421-426.

Greenberg, B., and Petkunas, T. (1987). An evaluation of Edit and Imputation Procedures Used in the 1982 Economic Censuses in Business Division, 1982 Economic Censuses and Census of Governments Evaluation Studies, U.S. Department of Commerce, Washington, pp.85-98.

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

Greenberg, B., and Petkunas, T. (1990). Structured Programs for Economic Editing and Referrals (SPEER), presented at 1990 annual meetings of the ASA. To appear in "Proceedings of the Section on Research Methods".

Greenberg, B., Draper, L., and Petkunas, T. (1990). On-line Capabilities in SPEER, presented at Statistics Canada Symposium, 1990.

Greenless, J.S., Reece, W.S. and Zieschang, K.D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. Journal of the American Statistical Association 77, pp.251-261.

Gross, W.F., Bode, G., Taylor, J.M. and Lloyd-Smith, C.W. (1986). Some finite population estimators which reduce the contribution of outliers. Proceedings of the Pacific Statistical Congress, Auckland, New Zealand, 20-24 May, 1985.

Groves, R.M. (1989). Survey Errors and Survey Costs. Wiley, New York.

Groves, R.M. and Nicholls II, W.L. (1986). The status of computer assisted telephone interviewing: Part II - Data Quality Issues, Journal of Official Statistics 2, pp.117-134.

GSFD, Generalized Survey Function Development Team (1989). Methodological and Operational Concepts in the Collection and Capture Module, Technical report, Statistics Canada, Ottawa.

GSFD (1989a). Investigation of Applicability of GSFD Concepts to the 1988 Annual Survey of Manufactures. Statistics Canada Technical Report.

Hanono, R.M. and Barbosa, D.M.R., A Tool for the Automatic Generation of Data Editing

and Imputation Application for Surveys Processing. Survey and Statistical Computing (SGCSA) - North Holland - pag. 449-456.

Hanono, R., Barbosa, D. (1996). CRIPTAX - A Generalized Editing Application Generator, Brazil.

Harris, K. (1996). Documentation and evaluation of data editing practices at the National Center for Health Statistics, USA.

Hasselblad, V., Creason, J.P., and Stead, A.G. (1983). Applications of the missing-information principle. In Incomplete Data in Sample Surveys. Volume 3, Proceedings of the Symposium. W.G. Madow, and I. Olkin (eds.). Academic Press, New York, pp.183-202.

Hedges, L.V., and Olkin, I. (1983). Selected annotated bibliography. In Incomplete Data in Sample Surveys. Volume 2, Theory and Bibliographies. W.G. Madow, I. Olkin, and D.B. Rubin (eds.). Academic Press, New York, pp.417-478.

Hedlin, D. (1993). Raw data compared with edited data. Presented at the Conference of European Statisticians' Work Session on Statistical Data Editing, Stockholm, 1993.

Hedlin, D. (1993). Raw Data Compared with Edited Data, Conference of European Statisticlans' Work Session on Statistical Data Editing, Stockholm 1993.

Heitjan, D.F. and Rubin, D.B. (1990). Inference from coarse data via multiple imputation with application to age heaping. Journal of the American Statistical Association 85, pp.304-314.

Heller, J.-L. (1993a). CAPI and BLAISE pour l'enquete Emploi en France. Essays on Blaise 1993: Proceedings of the Second International Blaise Users Conference, Office of Population Censuses and Surveys, London, 1993, pp.27-37.

Heller, J.-L. (1993b). The use of CAPI and BLAISE in the French Labour Force Survey. Presented at the Second International Blaise Users Conference, London, 1993.

Hellerman, E. (1982). Overview of the Hellerman I&O Coding System. Internal document, US Bureau of the Census.

Herzog, T.N. (1980). Multiple imputation of individual Social Security amounts Part II. Proceedings of the Section on Survey Research Methods. American Statistical Association, pp.404-407.

Herzog, T.N. and Lancaster, C. (1980). Multiple imputation of individual Social Security amounts, Part I. Proceedings of the Section on Survey Research Methods. American Statistical Association, pp.398-403.

Herzog, T.N., and Rubin, D.B. (1983). Using multiple imputations to handle nonresponse in

sample surveys. In Incomplete Data in Sample Surveys. Volume 2, Theory and Bibliographies. W.G. Madow, I. Olkin, and D.B. Rubin (eds.). Academic Press, New York, pp.210-248.

Hidiroglou, M.A. (1983). Imputations for the monthly retail trade survey. Statistics Canada Technical Report.

Hidiroglou, M.A. and Berthelot, J.-M. (1986). Statistical editing and imputation for periodic business survey. Survey Methodology 12, pp.73-83.

Hidiroglou, M. (1991). Structure of the Generalized Estimation System (GES). Statistics Canada.

Hidiroglou, M.A., and Berthelot, J.-M. (1986). Statistical editing and imputation for periodic business surveys, Survey Methodology 12, pp.73-83.

Hill, E., Jr. and French, C. (1981). Editing very large data bases. Proceedings of Conference on Information, Science, and Systems, pp.70-78.

Hill, C.J. (1978). The application of a systematic method of automatic edit and imputation to the 1976 Canadian Census of Population and Housing. Survey Methodology 4, pp.178-202.

Hinkins, S. and Scheuren, F. (1986). Hot-deck imputation procedure applied to a double sampling design. Survey Methodology 12, pp.181-196.

Hinkins, S. (1983). Matrix sampling and the related imputation of corporate income tax returns. Proceedings of Section on Survey Research Methods. American Statistical Association, pp.427-433.

Hoaglin, D.C., Mosteller, F., & Tukey, J. W. (Eds.) (1983). Understanding Robust and Exploratory Data Analysis. NY: Wiley.

Hocking, R.R. (1983). The design and analysis of sample surveys with incomplete data: Reduction of respondent burden. In Incomplete Data in Sample Surveys. Volume 3, Proceedings of the Symposium. W.G. Madow, and I. Olkin (eds.). Academic Press, New York, pp.107-125.

Hocking, R.R., Huddleston, H.F. and Hunt, H.H. (1974). A procedure for editing survey data. Technical Report.

Hodges, B.S.III (1983). Using the computer to edit and code survey data. Proceedings of the Section on Statistical Computing. American Statistical Association, pp.238-240.

Hogan, H., and Wolter, K. (1988). Measuring accuracy in a post-enumeration survey. Survey Methodology, pp.99-116.

Höglund D. E., (1989). A report on a study on the Hidiroglou-Berthelot Method (Statistical

Edits) applied to the Swedish Survey of the Delivery and Orderbook Situtation in the Swedish Industry. (Statistics Sweden).

Hollins, D. (1984). 1981 Census of Agriculture data processing methodology. Statistics Canada Technical Report.

Holt, D. (1974). On the editing and imputation of sample survey data. Technical Report.

Hooft van Huijsduijnen, J. and A. van Zijl, (1989). Surfox, release 1.0, user's manual (in Dutch).

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association, **47***, pp.663-685.

Houston, G., & Bruce, A. G. (1992, February). Graphical editing for business and economic surveys. Technical report, New Zealand Department of Statistics, Mathematical Statistical Branch.

Houston, G. and Bruce, A.G. (1993). Geographical Editing for Business and Economic Surveys, Journal of Official Statistics 9, pp.81-90.

Huang, E.T. (1984). An imputation study for the monthly retail trade survey. Proceedings of the Section on Survey Research Methods. American Statistical Association, pp.610-615.

Huang, E.T. (1986). Report on the imputation research for the monthly retail trade survey. Statistical Research Division Report Series, U.S. Bureau of the Census. CENSUS/SRD/RR-pp.86-09.

Huddleston, H.F. and Hocking, R.R. (1978). Imputation in agricultural surveys. Proceedings of the Section on Survey Research Methods. American Statistical Association, pp.480-485.

Hughes, P.J., McDermid, I. and Linacre, S.J. The Use of Graphical Methods in Editing, Australian Bureau of Statistics.

Hughes, P.J., Linacre, S.J. and McDermid, I.M. (1990). The use of graphical methods in editing. Proceedings of the U.S. Bureau of the Census Annual Research Conference, pp.538-550.

Hunter, P. (1993). Introducing CAI standards within OPCS. Essays on Blaise 1993: Proceedings of the Second International Blaise Users Conference, Office of Population Censuses and Surveys, London, 1993, pp.38-48.

Iannacchione, V.G. (1982). Weighted sequential hot deck imputation macros. Presented at the 1992 SAS User's Group International Conference, San Francisco, California, February 1982.

Iannacchione, V.G., Milne, J.G. and Folsom, R.E. (1991). Response probability weight adjustment using logistic regression. Proceedings of the Survey Methods Section, ASA, pp.637-642.

Informatica Comunidad de Madrid SA, 1993, Lince, Sistema de validación e imputatión automatica de datos estadisticos; manual de usuario (ICM, Madrid).

Israëls, A. (1996). Simultaneous imputation under balancing constraints, Netherlands.

Israëls, A.Z. (1995). Imputeren ten behoeve van voorlopige cijfers bij het Samenvattend Overzicht van de Industrie [Imputation for preliminary figues at the Summary of Manufacturing] (internal CBS-note).

Jabine, T.B. (1987). Nonsampling errors: Some reflections. Journal of Official Statistics,. pp.335-338.

Jinn, J.-H., Sedransk, J., (undated). Effect on Secondary Data Analysis of Different Imputation Methods.

Johansson, L.A. (1993). AKK. ("Automatisk Kodning av diagnoser i Klartext", or in English: "Automated Coding of Diagnostic Expressions"). Presented at the Conference of European Statisticians' Work Session on Statistical Data Editing, Stockholm, 1993.

Johnson, J.H., and Wade, M.J. (1990). Use of editing for DC2 - current and projected. Statistics Canada.

Jones, R.G. (1983). An examination of methods of adjusting for nonresponse to a mail survey: A mail-interview comparison. In Incomplete Data in Sample Surveys. Volume 3, Proceedings of the Symposium. W.G. Madow, and I. Olkin (eds.). Academic Press, New York, pp.271-290.

Kaiser, J. (1983). The effectiveness of hot-deck procedures in small samples. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp.523-528.

Kalpić, D. (1994). Automated Coding of Census Data, Journal of Official Statistics, Vol. 10. No. 4, 1994, pp.449-463, Statistics Sweden

Kalsbeek, W.D. (1980). A conceptual review of survey error due to nonresponse. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp.131-136.

Kalton, G. (1983). Compensating for Missing Survey Data. Survey Research Centre, University of Michigan.

Kalton, G. (1986). Handling wave nonresponse in panel surveys. Journal of Official Statistics, pp.303-314.

Kalton, G. and Kasprzyk, D. (1982). Imputing for missing survey responses. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp.22-31.

Kalton, G. and Kasprzyk, D. (1986). The treatment of missing survey data. Survey Methodology 12, pp.1-16.

Kalton, G., Kasprzyk, D. and Santos,R. (1980). Some problems of nonresponse and nonresponse adjustment in the Survey of Income and Program Participation. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp.501-506.

Kalton, G. and Kish, L. (1981). Two efficient random imputation procedures. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp.146-151.

Kalton, G. and Kish, L. (1984). Some efficient random imputation procedures. Communications in Statistics 13, pp.1919-1939.

Kalton, G., and Schuman, H. (1982). The effect of the question on survey responses: A review. Journal of the Royal Statistical Society, Ser. A.

Keller, W.J. and Bethlehem, J.G. (1990). The Impact of Microcomputers on Survey Processing at the Netherlands Central Bureau of Statistics. Proceedings of the Sixth Annual Research Conference of the Bureau of the Census, to appear.

Kent, Jean-Pierre (1995). Performance and Design, Statistics Netherlands.

Kiregyera, B. (1987). Types and some causes of nonsampling errors in household surveys in Africa. Journal of Official Statistics, pp.349-357.

Kirkendall, N.J. (1988). An Application of a Kalman Filter for Estimation and Edit. Proceedings of the business and Economic Statistics Section, American Statistical Association, pp.510-515.

Kish, L. (1967). Survey Sampling. Wiley: New York.

Knaus, R. (1987). Methods and Problems in Coding Natural Language Survey Data, Journal of Official Statistics, Vol 3, No. 1, pp.45-67, Statistics Sweden

Knaus, R. (1981). Pattern-based Semantic Decision Making. Empirical Semantics, edited by B Rieger, Bochum, West Germany.

Koeijers, Elly and Willeboordse, Ad (ed) (1995). Reference Manual on Design and Implementation of Bussiness Surveys, Statistics Netherlands.

Kokic, P.N. and Bell, P.A. (1994). Optimal winsorizing cut-offs for a stratified finite population estimator. Journal of Official Statistics, 10, pp.419-435.

Kott, P.S. (1987). Nonresponse in a periodic sample survey. Journal of Business and

Economic Statistics 5, pp.287-293.

Kott, P.S. (1988). Robust small domain estimation using random effects modelling. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp.377-380.

Kott, P.S. (1990). Nonresponse adjustments in NASS agricultural surveys. USDA Technical Report.

Kovar, J.G., (undated). Imputation in Surveys and Censuses, presented at the Statistics Canada Course #416E - "Survey Methodology".

Kovar, J.G. (1981). Edit and imputation package for the Integrated Agriculture Survey. Statistics Canada Technical Report.

Kovar, J.G. (1982a). Imputing agriculture data. Presented at the Methodological Interchange, a joint meeting of U.S. Bureau of the Census and Statistics Canada, Ottawa, Canada.

Kovar, J.G. (1982b). A closer look at the nearest neighbour/hot deck imputation methods: An empirical study. Statistics Canada Technical Report.

Kovar, J.G. (1990a). Generalized Edit and Imputation System: An overview. Presented at the Workshop on Data Editing and Imputation Methods, Rio de Janeiro, Brasil, February 12-16,1990.

Kovar, J.G. (1990b). Generalized Edit and Imputation System: Applications. Presented at the Workshop on Data Editing and Imputation Methods, Rio de Janeiro, Brasil, February 12-16, 1990.

Kovar, J.G. (1990c). Generalized Edit and Imputation System: Algorithms. Presented at the Workshop on Data Editing and Imputation Methods, Rio de Janeiro, Brasil, February 12-16, 1990.

Kovar, J.G. (1990d). Automatic editing: A discussion. Proceedings of the U.S. Bureau of the Census 1990 Annual Research Conference, pp.551-554.

Kovar, J.G., (1991). The Impact of Selective Editing on Data Quality. Presented at the Conference of European Statisticians' Work Session on Statistical Data Editing, Geneva, 1991.

Kovar, J.G. (1992). Recent Experiences with Data Editing. Presented at the Meeting of the UN/ECE Joint Group on Data Editing, Washington, 1992.

Kovar, J.G. (1993). Use of generalized systems in editing of economic survey data. Bulletin of the 49th session of the International Statistical Institute, Florence, Italy, 1993.

Kovar, J. (1996). Canada's report on developments and progress in Statistical Data Editing, Canada.

Kovar, J.G., MacMillan, J. and Whitridge, P. (1988). Overview and strategy for the Generalized and Imputation System. Statistics Canada, Methodology Branch Working Paper No. BSMD-88-007E.

Kovar, J.G., MacMillan, J.H. and Whitridge, P. (1991). Overview and Strategy for the Generalized Edit and Imputation System, Statistics Canada, Methodology Branch Working Paper BSMD 88-007E (updated in 1991).

Kovar, J.G. and Mayda, J.E. (1990). Edit and Imputation: An Unannotated Sampling of the Literature. Revista Brasileira de Estatistica.

Kovar, J.G., and Whitridge, P. (1990). Generalized Edit and Imputation System: Overview and Applications, Revista Brasileira Estatistica 51, pp.85-100.

Kovar, J.G. and Whitridge, P.J. (1993). Imputation of establishment survey data. In the Monograph of the International Conference on Establishment Surveys, Buffalo, NY, June 28-30. To be published by Wiley in 1994.

Kovar, J.G., Whitridge, P. and MacMillan, J. (1988). Generalized Edit and Imputation System for economic surveys at Statistics Canada. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp.627-630.

Kovar, J., and Winkler, W. E. (1996). Editing Economic Data, *American Statistical Association*, *Proceedings of the Section on Survey Research Methods*, to appear.

Kovar, J., and Winkler, W. (1996). Editing economic data, Canada, USA.

Kovašević, M. (1991). Kontrola kvaliteta obuhvata popisnog materijala optićkim ćitaćem. Radni materijal, Republièki zavod za statistiku Republike Hrvatske, Zagreb.

Kozak, R. (1993). Selective editing and its impact on data quality for the Canadian Annual Survey of Manufactures. Presented at the Conference of European Statisticians' Work Session on Statistical Data Editing, Stockholm, 1993.

Kozjek, P. (1996). Report on progress on CAI surveys at the Statistical Office of Slovenia, Slovenia.

Kusch, G.L. and Clark, D.F. (1979). Annual survey of Manufacturers: General statistics edit. Proceedings of the Section on Business and Economic Statistics, American Statistical Association, pp.183-187.

Kuusela, Vesa (1995). Interviewer interface of the.CAPI-system of Statistics Finland, Statistics Finland.

Kuusela, V., Virtanen, H., and Heiskanen, M. (1993). Use of prior information in Blaise CATI surveys. Essays on Blaise 1993: Proceedings of the Second International Blaise Users Conference, Office of Population Censuses and Surveys, London, 1993, pp.49-55.

Laflamme, F., Barrett, C., Johnson, W., and Ramsay, L. (1996). Experiences in Re-engineering the Approach to Editing and Imputing Canadian Imports Data, Proceedings of the Bureau of the Census Annual Research Conference and Technology Interchange, pp.1025-1037.

Lallande, D. (1988). Outlier detection system for survey of shipments, inventories and orders. Statistics Canada Technical Report.

Latouche, M. and Berthelot, J.-M. (1990). Use of a score function for error correction. To be presented at the Measurement Errors in Survey Conference, Tuscon, Arizona, November 11-14, 1990.

Latouche, M., and J.-M. Berthelot (1992). Use of a Score Function to Prioritize and Limit Recontacts in Editing Business Surveys, Journal of Official Statistics, Vol. 8, No. 3, 1992, pp.389-400.

Law, P., (1990). Special surveys: Imputation project (design specifications, Ver. 4), Statistics Canada Technical Report.

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

Lee, H. (in press). Outliers in survey sampling. In B. Cox et al. (Eds.), Survey Methods for Business, Farms, and Institutions. NY: Wiley.

Lehner, V. (1993). Activities of the Czech Statistics in data entry on microcomputers. Room paper presented at the Conference of European Statisticians' Work Session on Statistical Data Editing, Stockholm, 1993.

Lepp, H. and Linacre, S. (1993). Improving the efficiency and effectiveness of editing in a statistical agency. Bulletin of the 49th session of the International Statistical Institute, Florence, Italy.

Lessler, Judith T. (1983). An expanded survey error model. In Incomplete Data in Sample Surveys. Volume 3, Proceedings of the Symposium. W.G. Madow, and I. Olkin (eds.). Academic Press, New York, pp.259-270.

Liepins, G. (1983). Can automatic data editing be justified? One person's opinion. In Statistical Methods and the Improvement of Data Quality. T. Wright (ed.). Academic Press, New York, pp.205-214.

Liepins, G.E., (1989). Sound data are a sound investment. Quality Progress, pp.61-64.

Liepins, G.E. and Pack, D.J. (1980). An Integrated Approach to Data Editing. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp.777-781.

Liepins, G.E. and Pack, D.J. (1981). Maximal Posterior Probability Error Localization. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp.192-195.

Liepins, G.E., Garfinkel, R.S., Kunnathur, A.S., (1982). Error Localization for Erroneous Data: a Survey, TIMS/Studies in the Management Sciences 19, pp.205-219.

Lina, M. (1993). Blaise 2.5 / Interactive Coding. Netherlands Central Bureau of Statistics, Voorburg, The Netherlands.

Linacre, S. J. (1991). Approaches to Quality Assurance in ABS Business Surveys, Invited Papers Booklet, Vol. 2, 48th International Statistical Institute Session, Cairo, pp.297-321

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

Lindblom, A. (1990). A review of the macro-editing procedure Top-Down, Data Editing Joint Group Product NR SCP2/D.12/f, June 1990.

Lindell, K. (1995). Impact of editing on the salary statistics for employees in county councils, *Conference of European statisticians, Work Session on Statistical Data Editing, working paper No. 28, Athens 1995*.

Lindell, K. (1994). Evaluation of the editing process of the salary statistics for employees in country councils, paper presented at the UN congress on data editing in Cork, Ireland. To appear in Statistical Data Editing, Vol. 2. Geneva: UN Statistical Commission and Economic Commision for Europe.

Lindstrom, K. (1990). Functions of Macro-Editing: The Aggregate Method.

Lindstrom, K. (1991). A macro-editing - A review of some methods for rationalizing the editing of survey data. Statistical Journal of the United Nations Economic Commission for Europe, Vol. 8, No. 2, Special Issue on Data Editing, pp.155-166.

Lindstrom, K. (1994). A macro-editing application developed in PC-SAS. Statistical Data Editing Methods and Techniques. Volume No. 1, UN Economic Commission for Europe, Statistical Division, Conference of European Statisticians, 1994.

Lippmann, Richard P. (1987). An Introduction to Computing with Neural Nets. IEEE ASSP Magazine, April 1987, pp.4-22.

Little, R.J.A. (1979). Maximum likelihood inference for multiple regression with missing values: A simulation study. Journal of the Royal Statistical Society, Series B. 41, pp.76-87.

Little, R.J.A. (1982). Models for nonresponse in sample surveys. Journal of the American Statistical Association 77, pp.237-250.

Little, R.J.A. (1983). Superpopulation models for nonresponse. In Incomplete Data in Sample Surveys. Volume 2, Theory and Bibliographies. W.G. Madow, I. Olkin, and D.B. Rubin (eds.). Academic Press, New York, pp.337-413.

Little, R.J.A. (1984). Survey nonresponse adjustments. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp.1-18.

Little, R.J.A. (1985a). Nonresponse adjustments in longitudinal surveys: Models for categorial data. Bulletin of the International Statistical Institute. Amsterdam.

Little, R.J.A. (1985b). A note about models for selectivity bias. Econometrica 53, pp.1469-1474.

Little, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. International Statistical Review 54, pp.139-157.

Little, R.J.A. (1988a). A test of missing completely at random for multivariate data with missing values. Journal of the American Statistical Association 83, pp.1198-1202.

Little, R.J.A. (1988b). Missing-data adjustments in large surveys. (With discussion). Journal of Business and Economic Statistics 6, pp.287-301.

Little, R.J.A. (1988c). Robust estimation of the mean and covariance matrix from data with missing values. Applied Statistics 37, pp.23-38.

Little, R.J.A. and Rubin, D.B. (1983). Missing data in large data sets. In Statistical Methods and the Improvement of Data Quality. T. Wright (ed.). Academic Press, New York, pp.215-244.

Little, R.J.A. and Rubin, D.B. (1987). Statistical Analysis With Missing Data. J.Wiley & Sons, New York.

Little, R.J.A. and Rubin, D.B., (1989). The analysis of social science data with missing values. In: Sociological methods and research, volume 18, number 2&3, November 1989 / February 1990, pp.292-326.

Little, R.J.A. and Samuhel, M.E. (1983). Alternative models for CPS income imputation. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp.415-420.

Little, R.J.A., and Schluchter, M.D. (1985). Maximum likelihood estimation for mixed continuous and categorial data with missing values. Biometrika, pp.497-512.

Little, R.J.A. and Smith, P.J. (1983). Multivariate edit and imputation for economic data.

Proceedings of the Section on Survey Research Methods, American Statistical Association, pp.518-521.

Little, R.J.A., and Smith, P.J. (1987). Editing and Imputation for Quantitative Survey Data, Journal of the American Statistical Association 82, pp.58-68.

Lorigny J. (1982) Questionnaire theory applied to wording recognition, IEEE Congress at Les Arcs, Ed. CNRS GR23, Paris VI.

Lorigny, J. (1988). QUID - A general automatic coding method. Survey Methodology. December 1988, Vol. 14, No. 2, pp.289-298.

Luc, M. (1993). Preliminary Results on Analysing Age, Sex, Marital Status, Common-law Partner Status and Relationship to Person 1 for Some 1991 Six person Household Data, Social Survey Methods Division Report, Statistics Canada, Dated February 9, 1993.

Lyberg, L. (1981). Control of the Coding Operation in Statistical Investigations - Some Contributions. Doctoral Dissertation, Urval, CSB Statistika Centralybran, Stockholm.

Lyberg, L. (1985). Plans for computer-assisted data collection at Statistics Sweden. Bulletin of the International Statistical Institute.

Lyberg, L., and Dean, Patricia, (1991). International Review of Approaches to Automated Coding. Presented at the Conference of European Statisticians' Work Session on Statistical Data Editing, Geneva, 1991.

Lyberg, L., and Dean, Patricia (1992). Automated coding of survey responses: An international review. Presented at the Conference of European Statisticians' Work Session on Statistical Data Editing, Washington, 1992.

Lyberg, L., and Sundgren, B. (1987). The Impact of the Development of EDP on Statistical Methodology and Survey Techniques, R & D Report. Statistics Sweden, Stockholm.

Madow, W.G., Nisselson, H., and Olkin, I. (eds.) (1983). Incomplete Data in Sample Surveys. Volume 1, Report and Case Studies. Academic Press, New York.

Madow, W.G., Olkin, I., and Rubin, D.B. (eds.) (1983). Incomplete Data in Sample Surveys. Volume 2, Theory and Bibliographies. Academic Press, New York.

Madow, W.G., and Olkin, I. (eds.) (1983). Incomplete Data in Sample Surveys. Volume 3, Proceedings of the Symposium. Academic Press, New York.

Madsen, B. (1996). Data editing, imputation and statistical match at Statistics Denmark, Denmark.

Magnas, H.L. (1989). An Expert System to Assist in the Disposition of Computer Edit Error Flags. Presented at American Statistical Association Committee on Energy Statistics spring

meeting.

Manners, Tony (1995). The UK Family Expenditure Survey in Blaise: Lessons from Trials and Experience, Office of Population Censuses and Surveys, UK.

Manners, T., Cheesbrough, Sara, and Diamond, A. (1993). Integrated field and office editing in Blaise: OPCS's experience of complex financial surveys. Essays on Blaise 1993: Proceedings of the Second International Blaise Users Conference, Office of Population Censuses and Surveys, London, 1993, pp.56-69.

Manning, A. (1993). Conversion of a major NASS probability survey to Blaise. Essays on Blaise 1993: Proceedings of the Second International Blaise Users Conference, Office of Population Censuses and Surveys, London, 1993, pp.70-82.

Martin, J. (1993). PAPI to CAPI: the OPCS experience. Essays on Blaise 1993: Proceedings of the Second International Blaise Users Conference, Office of Population Censuses and Surveys, London, 1993, pp.96-117.

Martin, B. (1993a). The Blaise coding facilities - a closer look. Presented at the Second International Blaise Users Conference, London, 1993.

Martin, B. (1993b). Blaise applications in Statistics New Zealand. Essays on Blaise 1993: Proceedings of the Second International Blaise Users Conference, Office of Population Censuses and Surveys, London, 1993, pp.83-95.

Mattheiss, T.H., (1970). An Algorithm for Determining Irrelevant Constraints and all Vertices in Systems of linear Inequalities.

Mattheiss, T.H. and Rubin D.S. (1980). A survey and comparison of methods for finding all vertices of convex polyhedral sets. Mathematics of Operations Research 5, pp.167-185.

Mayda, J.E., Whitridge, P. and Berthelot, J.-M. (1990). An integrated approach to editing. Proceedings of the Business and Economics Statistics Section, American Statistical Association, to appear.

Mazur, C. (1990). Statistical Edit System for Livestock Slaughter Data. Nass Staff Report, SRB-90-01, National Agricultural Statistics Service, U.S. Department of Agriculture.

McKeown, P.G. (1975). A vertex ranking procedure for solving the linear fixed-charge problem. Operations Research 23, pp.1183-1191.

McKeown, P.G. (1984). A mathematical programming approach to editing of continuous survey data. Siam Journal of Scientific Statistical Computing 5, pp.784-797.

McKeown, P.G. and Schaffer, J.R. (1981). Using linear programming to find approximate solutions to the fields to impute problem for industry data. Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface, pp.288-291.

McNeil, D. R. (1977). Interactive Data Analysis. NY: Wiley.

Michaud, S. (1985a). The data transformation in the E&I system in PSTAT. Statistics Canada Technical Report.

Michaud, S. (1985b). Comparison of weighting and imputation: Proposal for the study. Statistics Canada Technical Report.

Michaud, S. (1985c). Study of the PSTAT edit and imputation system. Statistics Canada Technical Report.

Michaud, S. (1986). Revision of the proposed strategy for comparing different imputation techniques. Statistics Canada Technical Report.

Michaud, S. (1987). Weighting versus imputation: A simulation study. Presented at the Annual Meeting of the American Statistical Association.

Michaud, S. and Bureau, M. (1988). Edit and imputation of tax data: An overall strategy. Statistics Canada, Methodology Branch Working Paper No. METH-88-018E, May 1988.

Miller,A.K. (1985). Imputation requirements: Integrated agriculture system. Statistics Canada memorandum to R. Harris dated March 11, 1985.

Miller, R., Meroney, W. and Titus, E. (1987). Identification of Anomalous Values in Energy Data. Proceedings of the Business and Economic Statistics Section, American Statistical Association, pp.241-246.

Moody, J.E. (1993). Prediction Risk and Architecture for Neural Networks. In Charkassy, V., Friedman, J.H. and Wechsler, H. (Eds.). From Statistics to Neural Networks, Theory and Pattern Recognition Applications, Springer, Berlin.

Morris, C. N. (1983). Nonresponse issues in public policy experiments, with emphasis on the Health Insurance Study. In Incomplete Data in Sample Surveys. Volume 3, Proceedings of the Symposium. W.G. Madow, and I. Olkin (eds.). Academic Press, New York, pp.313-326.

Mosteller, F., & Tukey, J. (1977). Data Analysis and Regression. Reading, MA: Addison Wesley.

Mowry, S., and Estes, A. (1995). Graphical Interface Tools in Data Editing/Analysis, Washington Statistical Society Seminar presentation, March 10, 1995

Murphy, Patrick (1995). Role of Manipula Programs in Support of a Blaise III v 1.05, Institutional CATI Study, Battelle/Survey Research Associates, USA.

Murtagh, F. (1994) Neural Networks and Related 'Massively Parallel' Methods for Statistics: A Short Overview. Intemational Statistical Review, 1994, Vol. 62, No. 3, pp.275-

288.

Murthy, M.N. (1983). A framework for studying incomplete data, with a reference to the experience in some countries of Asia and the Pacific. In Incomplete Data in Sample Surveys. Volume 3, Proceedings of the Symposium. W.G. Madow, and I. Olkin (eds.). Academic Press, New York, pp.7-24.

Nadeau, C. (1992). Statistics on the Utilisation of Different Imputation Techniques in CANEDIT, Social Survey Methods Division Memo, Statistics Canada, Dated October 26, 1992.

National Agricultural Statistics Service, U.S. Department of Agriculture (1993). NASS experiences with CAPI. Room paper presented at the Conference of European Statisticians' Work Session on Statistical Data Editing, Stockholm, 1993.

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

Naus, J.I., Johnson, T.G. and Montalvo, R. (1972). A Probabilistic Model for Identifying Errors in Data Editing, Journal of the American Statistical Association, 67 (December 1972), pp.943-50.

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

Neiswanger, W.A. (1947). Elementary Statistical Methods. The Macmillan Company, New York.

Nesich, R. (1980). General Methodological Approach to Edit and Imputation in the 1981 Census of Agriculture. Statistics Canada.

Neumann, K. (1989). New meta-information aspects in a changing environment. Bulletin of the International Statistical Institute. Paris.

Neural Technologies Limited - Final Report from Phase 1 of the Neural Imputation Trial.

Nichols II, W.L., and Groves, R.M. (1985). The status of computer-assisted telephone interviewing. Bulletin of the International Statistical Institute.

Nordbotten, S. (1963). Automatic Editing of Individual Statistical Observations. Statistical Standards and Studies, Handbook No.2, United Nations, N.Y.

Nordbotten, S., (1963). Automatic Editing of Individual Statistical Observations, Presented at the Conference of European Statisticians, United Nations Statistical Commission and Economic Commission for Europe.

Nordbotten, S. (1965). The Efficiency of Automatic Detection and Correction of Errors in Individual Observations as Compared with Other Means for Improving the Quality of

Statistics, Bulletin of the International Statistical Institute, Proceedings of 35th Session, Helgrade, 41 (September 1965), pp.417-41.

Nordbotten, S. (1995). Editing Statistical Records by Neural Networks, Journal of Official Statistics. Vol. 11. No.4. Stockholm. pp.391-411.

Nordbotten, S. (1996). Neural network imputation applied to the Norwegian 1990 population census data, Norway.

Nordbotten, S. (1995). Editing Statistical records by neural networks. Norway. Paper prepared for the Work Session on Statistical Data Editing, Athens, Greece, 6-11 November 1995

Nordbotten, S. (1996). Predicting the Accuracy of Imputed Proportions, Department of Information Science, University of Bergen, Bergen.

Nordbotten, S. (1996a). Editing Statistical Records by Neural Networks. Journal of Official Statistics, No. 3, Stockholm.

Nordbotten, S. (1996a). Editing and Imputation by Means of Neural Networks, Statistical Journal of the United Nations ECE, Vol. 13, No. 2, pp.119-129.

Nordbotten, S. (1996b). Neural Network Irnputation Applied on Norwegian 1990 Population Census Data Utilizing Administrative Registers. Department of Information Science, University of Bergen, Bergen.

Nordbotten, S. (1996c). Small Area Statistics Based on Imputations from Survey Data and Administrative Registers, Department of Information Science, University of Bergen, Bergen.

Nordbotten. S. (1996d). Preliminary Report from Editing Experiments with Data from the Swedish Industrial Statistics, Department of Information Science, University of Bergen, Bergen.

Norris M.J. and Coyne S. (1991). Automated coding of Mobility Place Name data for the 1991 Census. Proceedings of Symposium 91 -Spatial Issues in Statistics, pp.83-94.

O'Reilly, James (1995). WIC Infant Feeding Practice Study: Development of a Complex longitudinal computer-assisted questionnaire, Research Triangle Institute, USA.

O'Reilly, J. M. (1993). Lessons learned programming a large, complex CAPI instrument. Presented at the Second International Blaise Users Conference, London, 1993.

Office of Management and Budget. (1987). Standard Industrial Classification Manual. Available from National Technical Information Service, Springfield, VA (Order no. PB 87-100012).

Oh, H.L., Scheuren, F.J. and Nisselson, H. (1980). Differential bias impacts of alternative

census bureau hot-deck procedures for imputing missing CPS income data. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp.416-420.

Ordinas, J.P. (1988). Generalized System Permitting Disaggregation in Cascade of Series Tables. I.N.E. Spain.

Oh, H.L. and Scheuren, F.J. (1978a). Multivariate raking ratio estimation in the 1973 exact match study. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp.716-722.

Oh, H.L. and Scheuren, F.J. (1978b). Some unresolved application issues in raking ratio estimation. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp.723-728.

Oh, H.L. and Scheuren, F.J. (1980). Estimating the variance impact of missing CPS income data. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp.408-415.

Oh, H.L. and Scheuren, F.J. (1983). Weighting adjustment for unit nonresponse. In Incomplete Data in Sample Surveys. Volume 2, Theory and Bibliographies. W.G. Madow, I. Olkin, and D.B. Rubin (eds.). Academic Press, New York, pp.143-184.

Ono, M. and Miller, H.P. (1969). Income non-response in the current population survey. Proceedings of the Social Statistics Section, American Statistical Association, pp.277-288.

Outrata, E. and Chinnappa, B.N. (1989). General survey functions design at Statistics Canada. Bulletin of the 47th Session of the ISI, Paris.

Pageau, F., (1989). Traitement de la Non-Reponse dans les Enquetes par Sondage, Memoire d'Initation a la recherche.

Pageau, F. (1992). Features of the CANEDIT Software, Social Survey Methods Division Report, Statistics Canada, Dated September 1992.

Pare, R.M. (1978). Evaluation of 1975 methodology: Simulation study of the imputation system developed by BSMD. Statistics Canada Technical Report.

Parkanová, M., Loutocký D. (1996). Future development of the DataMan system, Czech Republic.

Passe, R.M., Carpenter, M.J. and Passe H.A. (1987). Operational Outlier Detection. Communications in Statistics (Theory & Methods) 16, pp.3379-91.

Patrick, C.A., (1978). Estimation, imputation, randomization, and risk equivalence. Presented at the Annual Meeting of the American Statistical Association.

Perron, S., Berthelot, J.-M., and Blakeney, R. D. (1992). New technologies in data

collection for business surveys. Presented at the Conference of European Statisticians' Work Session on Statistical Data Editing, Washington, 1992.

Pettersson, H. (1993). Collection and editing of administrative data - some experiences from The Department of Demographic and social Statistics at Statistics Sweden. Presented at the Conference of European Statisticians' Work Session on Statistical Data Editing, Stockholm, 1993.

Pfeffermann, D. (1988). The effect of sampling design and response mechanism on multivariate regression-based predictors. Journal of the American Statistical Association 83, pp.824-833.

Phipps, P.A., and Tupek, A.R. (1991). Assessing measurement errors in a touchstone recognition survey. Survey Methodology, pp.15-26.

Pieper, D., (1991). Transposed File Controller (TrafiC): Status Report.

Pierce, D.A. and Bauer, L.L. (1989). Tolerance-Width Groupings for Editing Banking Deposits Data: An Analysis of Variance of Variances. Finance and Economics Discussion Series, No. 72.

Pierzchala, M. (1990). A Review of the State of the Art in Automated Data Editing and Imputation, Journal of Official Statistics 6, pp. 355-377.

Pierzchala (eds.). Proceedings of the Data Editing Workshop and Exposition, Washington D.C.: Bureau of Labor Statistics.

Pierzchala, M. (1990a). A review of the state of the art in automated data editing and imputation. Journal of Official Statistics, pp.355-377.

Pierzchala, M. (1990b). A review of three editing and imputation systems. Proceedings of the Section on Survey Research Methods, American Statistical Association, to appear.

Pierzchala, M. (1991). One Agency's Experience with the Blaise Survey Processing System. Proceedings of the Section on Survey Research Methods.

Pierzchala, M. (1992). Generating Multiple Versions of Questionnaires, Proceedings of the First International Blaise Users Meeting, Voorburg, The Netherlands: Netherlands Central Bureau of Statistics, pp.131-145.

Pierzchala, M. (1992). A review of the art in automated data editing and imputation (with Glossary of terms found in the editing literature). Statistical Data Editing Methods and Techniques. Volume No. 1, UN Economic Commission for Europe, Statistical Division, Conference of European Statisticians, February 1992, pp.1-57.

Pierzchala, M. (1993). Computer Generation of Mega-Version Instruments for Data Collection and Interactive Editing of Survey Data, Proceedings of the Annual Research

Conference, Washington, DC: U.S. Bureau of the Census, pp.637-645.

Pierzchala, M. (1993). The role of editing systems and software in improving survey methods and productivity. In the Monograph of the International Conference on Establishment Surveys, Buffalo, NY, June 28-30. To be published by Wiley in 1994.

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.

Pierzchala, Mark (1995). The 1995 June Area Frame Project, NASS, USA.

Pierzchala, M. (1995). The 1995 June Area Frame Project, *Proceedings of the Third International Blaise Users* Conference*, Helsinki, Finland, Statistics Finland (to be given).

Pietilä, P., Niemi, H. (1995). Total Quality and Computer-Assisted Interviewing; frames, definitions and tools for planing a total quality system, Statistics Finland.

Platek, R. (1985). Some important issues in questionnaire development. Journal of Official Statistics,. pp.119-136.

Platek, R. and Gray, G.B. (1978). Non-response and imputation. Survey Methodology 5, pp.144-177.

Platek, R.P. and Gray, G.B. (1983). Imputation methodology - Total survey error. In Incomplete Data in Sample Surveys. Volume 2, Theory and Bibliographies. W.G. Madow, I. Olkin, and D.B. Rubin (eds.). Academic Press, New York, pp.249-336.

Podehl, W.M. (1974). Introduction to the Generalized Editing Imputation System using Hot-Deck Approach. General Survey Statistics Division, Statistics Canada.

Polak, J., Axhausen, K., and Oldham, J. (1993). Experience with using the Blaise system for conjoint measurement of travel behaviour. Presented at the Second International Blaise Users Conference, London, 1993.

Pons Ordinas, J. Proceso de Macroedición, Analisis y Transferencias, Macro-Micro en la Encuesta Nacional, Desagregación en Cascada de Tablas de Series, Instituto Nacional de Estadística, España, Documento de Trabajo, Diciembre 1988.

Poulsen, M. E. (1995). The LFS and the register based labour force statistics - a quality assessment, contributed paper to the SMPQ-conference, Bristol, 1-4 April, 1995.

Pregibon, D. (1978). A discussion of the Survey Imputation and Editing session. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp.492-493.

Present and future of data editing in the INE, Ministerio de economia y hacienda, Instituto

nacional de estadistica, Madrid, 1990.

Pring-Mill, F. and Emery, D. (1985). Proposal for future work on Shell-2 SPIDER. Statistics Canada, Research and General Systems Technical Report.

Pritzker, L., Ogus, J. and Hansen M.H. (1965). Computer Editing Methods - Some Applications and Results, Bulletin of the International Statistical Institute, Proceedings of 35th Session, Belgrade, 41 (September 1965), pp.442-65.

Proctor, C.H. (1978). More on imputing versus deleting when estimating scale scores. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp.209-211.

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.

Radner, D.B. (1978). The development of statistical matching in economics. Presented at the Annual Meeting of the American Statistical Association.

Ramos, M., Waite,J.P., Cole, S.J., (undated). Evaluation Study of the Imputation of Data for Small Manufacturing Companies in the 1982 Census of Manufactures.

Rao, J.N.K., Hughes, E. (1983). Comparison of domains in the presence of nonresponse. In Incomplete Data in Sample Surveys. Volume 3, Proceedings of the Symposium. W.G. Madow, and I. Olkin (eds.). Academic Press, New York, pp.215-226.

Rao, P.S.R.S. (1981). Nonresponse in sample surveys: The imputation technique. Unpublished manuscript.

Rao, P.S.R.S. (1983a). Callbacks, follow ups, and repeated telephone calls. In Incomplete Data in Sample Surveys. Volume 2, Theory and Bibliographies. W.G. Madow, I. Olkin, and D.B. Rubin (eds.). Academic Press, New York, pp.33-44.

Rao, P.S.R.S. (1983b). Randomization approach (to nonresponse and double sampling). In Incomplete Data in Sample Surveys. Volume 2, Theory and Bibliographies. W.G. Madow, I. Olkin, and D.B. Rubin (eds.). Academic Press, New York, pp.97-106.

Rauch, L., and Vogel, J. (1989). The software environment for the central statistical data base and the impact on the statistical data processing. Bulletin of the International Statistical Institute.

Riviere P. (1994). The SICORE automatic coding system, Working Paper, Conference of European Statisticians, ISIS 94 Seminar, Bratislava, May 1994

Rivière, P. (1996). The new annual enterprise surveys in France, France.

Rizvi, M.H. (1983). An empirical investigation of some item nonresponse adjustment procedures. In Incomplete Data in Sample Surveys. Volume 1, Report and Case Studies. W.G. Madow, H. Nisselson, and I. Olkin, (eds.). Academic Press, New York, pp.299-366.

Robinson, E.L. and Richadrson, W.J. (1978). Editing and imputation of the 1977 truck inventory and use survey. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp.203-208.

Rockwell, R.C. (1975). An investigation of imputation and differential quality of data in the 1970 Census. Journal of the American Statistical Association, pp.39-42.

Roddick, L.H. (1993). Data Editing Using Neural Networks, Statistics Canada, Ottawa.

Rodgers, W.L., and Herzog, A.R. (1987). Covariances of measurement errors in surveys responses. Journal of Official Statistics, pp.403-418.

Roessingh, M., and Bethlehem, J. (1993). Trigram coding in the Family Expenditure Survey of the CBS. Presented at the Conference of European Statisticians' Work Session on Statistical Data Editing, Stockholm, 1993.

Roessingh, M., and Bethlehem, J. (1993). Trigram coding in the Family Expenditure Survey of the CBS. Essays on Blaise 1993: Proceedings of the Second International Blaise Users Conference, Office of Population Censuses and Surveys, London, 1993, pp.118-132.

Rousseeuw, P.J., and A.M. Leroy, 1987, Robust regression and outlier detection (Wiley, New York).

RSG, (1985). SPIDER User Guide. Statistics Canada Technical Report.

Rubin, D.B. (1974). Characterizing the estimation of parameters in incomplete data problems. Journal of the American Statistical Association, pp.467-474.

Rubin, D.S., (1975). Vertex generation and cardinality constraint linear programs. Operations Research, 23, pp.555-565.

Rubin, D.B. (1976). Inference and missing data. Biometrika, pp.581-592.

Rubin, D.B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. Journal of the American Statistical Association.

Rubin, D.B. (1978). Multiple imputations in sample surveys: A phenomenological Bayesian approach to non-response. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp.20-34.

Rubin, D.B. (1979). Illustrating the use of multiple imputations to handle nonresponse in sample surveys. Bulletin of the International Statistical Institute, pp.517-532.

Rubin, D.B. (1980). Using multiple imputations to handle nonresponse. National Academy of Sciences's Panel on Incomplete Data. Draft.

Rubin, D.B. (1983). Conceptual issues in the presence of nonresponse. In Incomplete Data in Sample Surveys. Volume 2, Theory and Bibliographies. W.G. Madow, I. Olkin, and D.B. Rubin (eds.). Academic Press, New York, pp.125-142.

Rubin, D.B. (1986). Basic ideas of multiple imputation for nonresponse. Survey Methodology 12, pp.37-47.

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

Rubin, D.B. (1990). Imputation: A discussion. Proceedings of the U.S. Bureau of the Census 1990 Annual Research Conference, pp.676-679.

Rubin, D.B., Schafer, J.L. and Schenker, N. (1988a). Imputation strategies for estimating the undercount. Proceedings of the Fourth Annual U.S. Bureau of the Census Research Conference, pp.151-159.

Rubin, D.B., Schafer, J.L. and Schenker, N. (1988b). Imputation strategies for missing values in post-enumeration surveys. Survey Methodology 14, pp.209-221.

Rubin, D.B. and Schenker, N. (1984). Multiple imputation methods for simple random samples with ignorable nonresponse. Department of Statistics, University of Chicago.

Rubin, D.B. and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. Journal of the American Statistical Association, pp.366-374.

Rubin, D.B. and Schenker, N. (1987). Interval estimation from multiply-imputed data: A case study using census agriculture industry codes. Journal of Official Statistics 3, pp.375-387.

Rubin, D.B. and Zaslavsky, A.M. (1989). An overview of representing within-household and whole-household misenumerations in the Census by multiple imputations. Proceedings of the U.S. Bureau of the Census Fifth Annual Research Conference, pp.109-117.

Rumelhart, D.E. and McClelland, J.L. (1986). Parallel Distributed Processing - Explorations in Microstructure of Cognition. Vol. 1: Foundation. MIT Press, Cambridge, Mass.

Sande, G. (1976a). Diagnostic capabilities for a numerical edit specification analyzer. Statistics Canada Technical Report.

Sande, G. (1976b). Searching for numerically matched records. Statistics Canada Technical Report.

Sande, G. (1977). Numerical edit and imputation test system. Statistics Canada Technical Report.

Sande, G. (1978). An algorithm for the fields to impute problems of numerical and coded data. Statistics Canada Technical Report.

Sande, G. (1979). Numerical edit and imputation. Presented at the 42nd Session of the International Statistical Institute, Manila, Philippines.

Sande, G. (1981). Descriptive statistics used in monitoring edit and imputation process. Presented at the 13th Symposium on the Interface, Pittsburgh, Pennsylvania.

Sande, Innis G. (1983). Hot-deck imputation procedures. In Incomplete Data in Sample Surveys. Volume 3, Proceedings of the Symposium. W.G. Madow, and I. Olkin (eds.). Academic Press, New York, pp.339-350.

Sande, Innis G. (1987). A Statistics Canada perspective on numerical edit and imputation in business surveys. Presented at the Conference of European Statisticians, Geneva.

Sande, Innis G. (1979). A personal view of hot-deck imputation procedures. Survey Methodology 5, pp.238-258.

Sande, Innis G. (1982). Imputation in surveys: Coping with reality. American Statistician, pp.145-152.

Sanderse, S. (1995). Simultaan imputeren. Imputeren onder restricties in het Samenvattend Overzicht van de Industrie 1992 Industrie [Simultaneous imputation. Imputation under constraints at the Summary of Manufacturing, 1992] (internal CBS-report).

Santa, J., (1991a). On the Fellegi-Holt Rule Analysis. Prepared for the 8th meeting of the Joint Group on Data Editing.

Santa, J., (1991b). private communications, (two handwritten documents).

Santa, J., (1991c). Rule Analyzer. Prepared for the 8th meeting of the Joint Group on Data Editing.

Santos, R.L. (1981). Effects of imputation on regression coefficients. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp.140-145.

Särndal, C-E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology, **18*** (2) pp.241-252.

Särndal, C-E., Swensson, B. and Wretman, J. (1992*). Model Assisted Survey Sampling*. New York: Springer-Verlag.

SAS Institute Inc. (1990). Technical Accessing Data Effiiciently with PROC SQL Views,

SAS Communication Magazine, Vol. XVI, No. 1, Third Quarter 1990, P.3.

SAS, (1993), SAS/INSIGHT user's guide, version 6, 2nd ed. (SAS Institute, Cary NC, USA)

Schaible, W.L. (1983). Estimation of finite population totals from incomplete sample data: Prediction approach. In Incomplete Data in Sample Surveys. Volume 3, Proceedings of the Symposium. W.G. Madow, and I. Olkin (eds.). Academic Press, New York, pp.131-142.

Scheiber, S.J. (1978). A comparison of three alternative techniques for allocating unreported Social Security income on the survey of the Low-Income Aged and Disabled. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp.212-218.

Schenker, N. (1988). Handling missing data in coverage estimation, with application to the 1986 test of adjustment related operations. Survey Methodology 14, pp.87-97.

Schenker, N. (1989). The use of imputed probabilities for missing binary data. Proceedings of the U.S. Bureau of the Census Fifth Annual Research Conference, pp.133-141.

Schenker,N., Treiman,D. and Weidman, L. (1988). Multiple imputation of industry and occupation codes for public-use data files, Tables 1-6. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp.85-92.

Schiopu-Kratina, I., and Kovar, J. (1989). Use of Chernikova's algorithm in the Generalized Edit and Imputation System, Working paper no. BSMD- 89-001E, Ottawa: Statistics Canada.

Schiopu-Kratina, I., and Kovar, J.G. (1989). Use of Chernikova's algorithm in the Generalized Edit and Imputation System. Statistics Canada, Methodology Branch Working Paper No. BSMD-89-001E.

Schiopu-Kratina, I., and Srinath, K.P. (1986). The Methodology of the Survey of Employment, Payroll and Hours, Internal Working Paper, Ottawa: Statistics Canada.

Schou, Roger (1995). Developing a Multi-Mode Survey System, NASS, USA.

Schou, R., and Pierzchala, M. (1993). Standard multi-survey shells in NASS. Essays on Blaise 1993: Proceedings of the Second International Blaise Users Conference, Office of Population Censuses and Surveys, London, 1993, pp.133-142.

Schuerhoff, M. (1989). Changes in automated survey processing. Selected Papers from The Fifth Meeting of Data Editing Joint Group, Moscow, 1989.

Schuerhoff, M. (1990a). Trends in automation. Survey processing in the nineties. Netherlands Central Bureau of Statistics.

Schuerhoff, M. (1990b). New developments in the BLAISE system for survey processing. Netherlands Central Bureau of Statistics.

Schuerhoff, M. (1991). Developments in the Blaise System for Survey Processing.

Schuerhoff, M., Roessingh, M., Hofman, L., (1991). Examples of Computer Assisted Coding. Presented at the Conference of European Statisticians' Work Session on Statistical Data Editing, Geneva, 1991.

Schuerhoff, M., Roessingh, M., and Hofman, L. (1992). Examples of computer assisted coding. Presented at the Conference of European Statisticians' Work Session on Statistical Data Editing, Washington, 1992.

Schulte, E., Nordholt, J., and van Huijsduijnen, Hooft. (1996). The treatment of item nonresponse during the editing of survey results, Netherlands.

Searle, S.R. (1971). Linear models (Wiley, New York).

Shanks, J.M. (1989). Information technology and survey research: Where do we go from here? Journal of Official Statistics, pp.3-22.

Searls, D.T. (1966). An estimator which reduces large true observations. Journal of the American Statistical Association, 61, pp.1200-1204.

Sebestik, J., Zelon, H., DeWitt, D., O'Reilly, J.M. and McGowan, K. (1988). Initial Experiences with CAPI, Proceedings of the Bureau of the Census Annual Research Conference, pp.357-365.

Shapiro, G.M. (1987). Interviewer - respondent bias resulting from adding supplemental questions. Journal of Official Statistics.

Shasha, D., Wang, T.-L., (1990). New Techniques for Best-Match Retrieval, ACM Transactions on Information Systems, Vol. 8, No. 2, p.140-158.

Shimizu, I.M., Gonzalez, J.F. and Jones, G.K. (1980). Alternative adjustments for nonresponse in the National Hospital Discharge survey. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp.649-651.

Silva, P.L.N. and Bianchini, Z.M. (1993). Data editing issues and strategies at the Brazilian Central Statistical Office. Bulletin of the 49th session of the International Statistical Institute, Florence, Italy.

Siminoff, J.S. (1984). The Calculation of Outlier Detection Statistics. Communications in Statistics, Simulation & Computation, 13, pp.275-285.

Singh, A.C. (1989). Log-linear imputation. Proceedings of the U.S. Bureau of the Census Fifth Annual Research Conference, pp.118-132.

Singh, A.C., Armstrong, J.B. and Lemaitre, G.E. (1988). Log-linear imputation and its application to file merging. Statistics Canada Technical Report.

Singh, A.C., Mantel, H., Kinack, M., Rowe, G., (1990). On Methods of Statistical Matching With and Without Auxiliary Information, Statistics Canada Technical Report.

Singh, B. (1983). Bayesian approach (to nonresponse and double sampling). In Incomplete Data in Sample Surveys. Volume 2, Theory and Bibliographies. W.G. Madow, I. Olkin, and D.B. Rubin (eds.). Academic Press, New York, pp.107-124.

Singh, B., and Sedransk, J.H. (1983). Bayesian procedures for survey design when there is nonresponse. In Incomplete Data in Sample Surveys. Volume 3, Proceedings of the Symposium. W.G. Madow, and I. Olkin (eds.). Academic Press, New York, pp.227-248.

Sirken, M.G. (1983). Handling missing data by network sampling. In Incomplete Data in Sample Surveys. Volume 2, Theory and Bibliographies. W.G. Madow, I. Olkin, and D.B. Rubin (eds.).Academic Press, New York, pp.81-92.

Siver, E., and R. Peterson, (1985). Decision systems for inventory management and production planning, second edition (Wiley, New York).

Smith, James E. (1995). From Products to Systems: Addressing the Needs of CAI Surveys at Westat, WESTAT, USA.

Smith, P., Kokic, P. (1996). Winsorisation in ONS business surveys, UK.

Smouse, E.V. (1982). Bayesian estimation of a finite population total using auxiliary information in the presence of nonresponse. Journal of the American Statistical Association. pp.97-102.

Spiers, E.F. and Knott, J.J. (1969). Computer method to process missing income and work experience information in the Current Population Survey. Proceedings of the Social Statistics Section, American Statistical Association, pp.289-297.

Stasny, Elisabeth A. (1986). Estimating gross flows using panel data with nonresponse: An example from the Canadian Labour Force Survey. Journal of the American Statistical Association.

Stasny, Elisabeth A. (1987). Some Markov-chain models for nonresponse in estimating gross labour force flows. Journal of Official Statistics. pp.359-374.

Statistical Data Editing Methods and Techniques. Volume No.1. Economic Commission for Europe, Statistical Division, Geneva.

Statistical Office of the Republic of Slovenia (1993). VEGA-STAT. Presented at the Conference of European Statisticians' Work Session on Statistical Data Editing, Stockholm, 1993.

Statistical Policy Working Paper 19 (1990). Computer Assisted Survey Information Collection, Office of Management and Budget.

Statistics Canada (1988). Generalized Data Collection and Capture proposed production system environment. Statistics Canada Technical Report.

Statistics Denmark (1992). Statistiske efterretninger - Arbejdsmarked (1992:20), Arbejdsstyrkeundersøgelsen 1991.

Statistics Denmark (1993). Statistiske efterretninger - Arbejdsmarked (1993:17), Registerbaseret arbejdsstyrkestatistik ultimo November 1991.

Statistics Denmark (1994). Gross national product; comparison of labour force surveys with data from administrative registers with special emphasis on the full coverage of economic activity, Report (unpublished).

Statistics Denmark (1996). Industry Survey, Sample Survey 1996, Working Paper no. 1.

Statistics Netherlands (1996). Blaise Developers Guide. Voorburg: Statistical Informatics Department.

Statistisk- sentralbyrå (1992). Folke- og boligtellingen 1990, Oslo.

Steeh, Charlotte G., Groves, R.M., Comment, R., and Hansmire, Evelyn (1983). Report on the Survey Research Centre's surveys of consumers attitudes. In Incomplete Data in Sample Surveys. Volume 1, Report and Case Studies. W.G. Madow, H. Nisselson, and I.Olkin, (eds.). Academic Press, New York, pp.173-208.

Stol, R. H. (1993). An architecture for EDI in business surveys based on the use of Blaise. Essays on Blaise 1993: Proceedings of the Second International Blaise Users Conference, Office of Population Censuses and Surveys, London, 1993, pp.143-153.

Stuart, W.J. (1966). Computer editing of survey data: Five years of experience in BLS manpower. Journal of the American Statistical Association. pp.375-384.

Subcommittee on Data Editing in Federal Statistical Agencies (1990). Data editing in federal statistical agencies. Statistical Policy Working Paper.

Subcommittee on Data Editing in Federal Statistical Agencies (1990). Data Editing in Statistical Agencies, Statistical Policy Working Paper 18, Office of Management and Budget.

Sunter, A.B., Patrick, C.A. and Binder, D.A. (undated). On the editing of survey data. Statistics Canada Technical Report.

Sweet, E., Russell, C. (1996). A Discussion of Data Collection Via the Internet, Proceedings of the Section on Survey Research Methods, American Statistical Association, in print.

Szameitat, K. and Ziendler, H.J. (1965). The Reduction of Errors in Statistics by Automatic Corrections, Bulletin of the International Statistical Institute, Proceedings of 35th

Session, Belgrade, 41)September 1965), pp.442-65.

Taeuber, C. and Hansen, M.H. (1963). A preliminary evaluation of the 1960 Censuses of Population and Housing. Proceedings of the Section on Social Statistics, American Statistical Association, pp.56-71.

Tambay, J.L. (1986). Study of Outiers in the C.S.I.O., Regional Offices, Technical Report, Statistics Canada.

Teague, A. and Thomas, J. (1996). Neural Networks as a Possible Means for Imputing Missing Census Data in the 2001 British Census of Population, In Banks, R. et al. (eds). Survey and Statistical Computing, 1996.

Teufel T.: Informationsspuren zum numerischen und graphischen Vergleich von reduzierten natürlichsprachlichen Texten, Dissertation der ETH Zürich, 1989

Thomas, J. (1996). Neural networks as a possible means of imputing census data in the 2001 British Census of Population, United Kingdom.

Thomas, J. (1996). Statistical measurement and monitoring of data editing and imputation in the 2001 British Census of Population, United Kingdom.

Thompson, K. J. (1996). "Statistical Methods for Developing Ratio Edit Tolerances for Economic Censuses," *American Statistical Association*, *Proceedings of the Section on Survey Research Methods*, to appear.

Thomsen, I. and Siring, E. (1983). On the causes and effects of nonresponse: Norwegian experiences. In Incomplete Data in Sample Surveys. Volume 3, Proceedings of the Symposium. W.G. Madow, and I. Olkin (eds.). Academic Press, New York, pp.25-68.

Thomsen, I., and Tefsu, D. (1988). On the use of models in sampling from finite population. In Handbook of Statistical: Sampling. P.R. Krishanaiah and C.R. Rao (eds.). North-Holland, Amsterdam, pp.369-398.

Thornberry Jr., O.T. and Massey, J.T. (1978). Correcting for undercoverage bias in random digit dialled national health surveys. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp.224-229.

Toro, V. and Chamberlain, K. (1988). The use of Microcomputers for Census Processing in Developing Countries: An Update. Proceedings of Symposium 88, The Impact of High Technology on Survey Taking, Ottawa, Ontario, Canada: Statistics Canada, pp.181-199.

Tortora, R.D. (1985). CATI in an Agricultural Statistical Agency, Journal of Official Statistics 1, pp.301-314.

Tortora, R.D. (1987). Quantifying nonsampling errors and biases. Journal of Official Statistics, pp.339-342.

Tourigny, J.Y., Moloney, Joanne, and Miller, Diane (1993). The 1991 Canadian Census of Population experience with automated coding, (including the annex: Rick Ciok.: The results of automated coding in The 1991 Canadian Census of Population). Presented at the Conference of European Statisticians' Work Session on Statistical Data Editing, Stockholm, 1993.

Trebjesanin, B., (1990). Statistical Computing Project, Phase 2. Report from the seventh meeting of the Joint Group on Data Editing.

Trewin, D. (1987). How do we reduce nonsampling errors? Journal of Official Statistics. pp.343-348.

Tukey, J.W. (1977). Exploratory Data Analysis., Addison-Wesley Publishing Company, 1977.

Tukey, J.W. (1977). Exploratory Data Analysis. London: Addison-Wesley.

Tupek, A.R. and Richardson, W.J. (1978). Use of ratio estimates to compensate for nonresponse bias in certain economic surveys. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp.197-202.

Turner, R., and Lawes, M. (1983). Incomplete data in the survey of consumer finances. In Incomplete Data in Sample Surveys. Volume 1, Report and Case Studies. W.G. Madow, H. Nisselson, and I. Olkin, (eds.). Academic Press, New York, pp.269-298.

Tyrkkö, A., Vihavainen, H. (1995). CAPI and the Collection of Living Condition Data; a user's view, Statistics Finland.

U.S. Bureau of the Census. (1994, April). Combined Annual and Revised Monthly Wholesale Trade, January 1987-December 1993. Washington, DC: U.S. Government Printing Office (Current Business Reports, Item BW/93-RV).

U.S. Bureau of the Census. (1992). Annual Survey of Communication Services: 1992 . Washington, DC: U.S. Government Printing Office (Current Business Reports, Item BC/92).

United Nations, 1994, Statistical Data Editing Vol. 1: Methods and Techniques. Conference of European Statisticians, Statistical Standards and Studies No. 44. (UN Statistical Commission and Economic Commission for Europe, Geneva).

Valliant, R., Tomasino, R., and Hansen, M.H. (1983). Treatment of missing data in an office equipment. In Incomplete Data in Sample Surveys. Volume 1, Report and Case Studies. W.G. Madow, H. Nisselson, and I. Olkin, (eds.). Academic Press, New York, pp.209-236.

Van Bastelaer, A., Kerssemakers, F. and Sikkel, D. (1987). A test of the Continues Labour Force Survey with hand-held computers: Interviewer behavior and data quality. Netherlands Central Bureau of Statistics, Voorburg, Netherlands.

Van Bochove, C.A. (1996). From assembly line to electronic highway junction: a twin-track transformation of the statistical process. Netherlands Official Statistics. Statistics Netherlands, Voorburg/Heerlen.

Van Buuren, S., van Rijckevorsel, (1991). Fast Least Squares Imputation of Missing Data, from Leiden Psychological Reports (Psychometrics and Research Methodology PRM 01-91).

Van de Pol, Frank (1995). Data editing of business surveys: an overview. Report. CBS, Voorburg.

Van de Pol, Frank (1995). Data Editing of Business Surveys: an Overview, Vol. 10, Netherlands Official Statistics, Winter 1995.

Van de Pol, Frank (1995). Selective and automatic editing with CADI-applications, Statistics Netherlands. In: V. Kuusela, ed., Essays on Blaise 1995. Proceedings of the third International Blaise Users's Conference (Statistics Finland, Helsinki), pp.159-168.

Van de Pol, F. (1995). Selective editing in the Netherlands annual construction survey. *Conference of European statisticians, Work Session on Statistical Data Editing, working paper No. 26, Athens 1995*.

Van de Pol, F. and Diederen, B. (1996). A priority Index for macro-editing the Netherlands foreign trade survey. Report. CBS, Voorburg, Netherlands.

Van de Pol, F. and Molenaar, W. (1995). Selective and automatic editing with CADI-applications, In V. Kuusela, (ed)., Essays on Blaise 1995, Proceedings of the third International Blaise Users's Conference, Heslinki: Statistics Finland, pp.159-168.

Van de Pol, F. and Molenaar, W. (1996). Selective Editing: Where is the limit? Report. CBS, Voorburg.

Van Toor, Leo (1995). Use of Blaise and Manipula in the Annual Survey of Employment and Wages, Statistics Netherlands.

Vaughan, D.R. (1978). Errors in reporting supplemental security income recipience in a pilot household survey. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp.288-293.

Velleman, P.F., & Hoaglin, D. (1981). Applications, Basics, and Computing of Exploratory Data Analysis. Boston: Duxbury Press.

Venetiaan, S.A., (1990). Relaties tussen enkele methoden die het binnenwerk van een kruistabel aanpassen aan bekende randtotalen [Relations between several methods for adjustment of a table to known marginals] (internal CBS-report).

Verboon, P., (1994). Automatisch imputeren van deelposten bij bedrijfsenquêtes [Automatic imputation of subitems at business surveys] (internal CBS-note).

Verboon, P. (1996). Estimating correlation coefficients under nonresponse, Netherlands.

Verboon, P. and A.Z. Israëls, (1994). A simulation study on the treatment nonresponse in continuous data (Research Report, Statistics Netherlands).

Verboon, P. and Schulte Nordholt, E., (1994). Simulation experiments for hot deck imputation (Research report, Statistics Netherlands).

Viglino, L. (1992). QUID: An automatic coding method application to the Census in the French Overseas Department. Presented at the Conference of European Statisticians' Work Session on Statistical Data Editing, Washington, 1992.

Vogel, Fred, H. Bynum, G. Hanuschak, R. Murphy, W. Dowdy, C. Hudson, and J. Steinberg (1985). Crop Reporting Board Standards. Report of the Crop Reporting Board Policy and Procedures Working Group, Statistical Reporting Service, U.S. Department of Agriculture.

Wahlström, C. (1990). The Effects of Editing: A Study of The Financial Statistics at Statistics Sweden (in Swedish), Statistics Sweden, F-metod 27, 1990.

Walker, W.E. (1976). A heuristic adjacent extreme point algorithm for the fixed charge problem. Management Science 22, pp.587-596.

Walukiewicz, S., (1983). The Ellipsoid Algorithm for Linear Programming

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.

Weir, P., Emery, R., and Walker, J., (1996). The graphical editing analysis query system. To appear in: Proceedings of the Section on Survey Research Methods of the American Statistical Association.

Welniak, E.J. and Coder, J.F. (1980). A measure of the bias in the March CPS earnings imputation system. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp.421-425.

Wensing, Fred (1995). Update from "Downunder"; History, plans and functions we've built for CAI and Blaise in Australia, Australian Bureau of Statistics.

Wenzowski, M.J. (1988). ACTR - A generalized automated coding system. Survey Methodology. pp.299-308.

Wenzowski, Michael J. (1995). Advances in Automated Computer Assisted Coding Software at Statistics Canada, Statistics Canada.

Werbos, P. (1974). Beyond Regression: New Tools for Prediction and Analysis in

Behaviour Sciences. Ph.D. dissertation, Harvard University.

Werking, George S. (1994). Establishment Surveys: Designing the Survey Operations of the Future, Proceedings of the Section on Survey Research Methods, Invited Panel on the Future of Establishment Surveys, American Statistical Association, pp.163-169.

Werking, George S., and Clayton, R.L. (1991). Enhancing Data Quality Through the Use of Mixed Mode Collection, Survey Methodology, June 1991, 17, No. 1, pp.3-14.

Werking, G., Tupek, A. and Clayton, R. (1988). CATI and touchstone self response applications for establishment surveys. Journal of official statistics, Vol. 4, No. 4, 1988, pp.349-362.

Werner, B. (1977). The development of automatic editing for the next Census of Population, Statistical News No 37, UK Central Statistical Office, pp.3710-3715.

Wesolowska, Maria (1993). Selected problems of statistical data correctness check methods and techniques on the basis of Retail Prices Survey. Presented at the Conference of European Statisticians' Work Session on Statistical Data Editing, Stockholm, 1993.

Wesseling, H. (1993). Retrospective questioning with the aid of Blaise. Netherlands Central Bureau of Statistics, Heerlen.

West, S.A., Butani, S. and Witt, M. (1993). Alternative imputation methods for wage data. Proceedings of the International Conference on Establishment Surveys, Buffalo, New York, June 28-30.

Wetherill, G.B. and Gerson, M.E. (1987). Computer aids to data quality control. The Statistician 36, pp.589-592.

Whitridge, P., Kovar, J.G. and MacMillan, J. (1988). Systeme generalise de verification et d'imputation pour les enquetes economiques a Statistique Canada. Presented at the ACFAS, Moncton, New Brunswick.

Whitridge, P. and Kovar, J.G. (1990). Applications of the Generalized Edit and Imputation System at Statistics Canada. Proceedings of the Section on Survey Research Methods, American Statistical Association, to appear.

Whitridge, P., Bureau, M. and Kovar, J.G. (1990b). Use of mass imputation to estimate for subsample variables. Proceedings of the Business and Economic Statistics Section, American Statistical Association, to appear.

Whitridge, P., Bureau, M. and Kovar, J.G. (1990a). Mass imputation at Statistics Canada. Proceedings of the U.S. Bureau of the Census 1990 Annual Research Conference, pp.666-675.

Willenborg, L. C. R. J. (1988a). Computational Aspects of Survey Data Processing.

Catholic University of Brabant, Tilburg, The Netherlands.

Wings, H., Hofman, L. (1995)**.** MANILUS; A powerful environment for managing Blaise III applications, Statistics Netherlands.

Winkler, W.E. and L.R. Draper (1995). Application of the SPEER edit system. US Bureau of the Census, Washington DC, USA.

Winkler, W. E., and Petkunas, T. (1996). "The DISCRETE Edit System," in *Data Editing, Volume 2*," U.N. Statistical Commission and Economic Commission for Europe, Geneva, Switzerland, to appear.

Willenborg, L.C.R.J. (1988b). Computational Aspects of Survey Data Processing. Centre for Mathematics and Computer Science Tract 54, Amsterdam.

Winkler, W. E. (1996). "SPEER Edit System," computer system and unpublished documentation, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C., USA. 1991 Census Report for Great Britain, (Part 1), ISBN 0-11-691536-6

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.

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. (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, unpublished document, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C., USA.

Winkler, W. (1996). The new structured programs for economic editing and referrals (SPEER), USA.

Winkler, W., and Draper, L.R. (1994). Application of the SPEER edit system, Research paper, Washington D.C.: U.S. Bureau of the Census.

Woodbury, M.A. (1983). Statistical record matching for files. In Incomplete Data in Sample Surveys. Volume 3, Proceedings of the Symposium. W.G. Madow, and I. Olkin (eds.). Academic Press, New York, pp.173-182.

Woolf, M., (1991). The Implications of Changing Automatic Systems For Survey Practioners. Presented at the Conference of European Statisticians' Work Session on Statistical Data Editing, Ottawa, 1991.

Wright, T. (ed.) (1983). Statistical Methods and the Improvement of Data Quality. Academic Press, New York.

Yates, F. (1971). The Use of Computers for Statistical Analysis: A Review of Aims and Achievements, Bulletin of the International Statistical Institute, Proceedings of 38th Session, Washington, 44 (August 1971), pp.39-53.

Yielding, J. (1993). Adjustment-cell methods for estimation of finite population quantiles under nonresponse. Texas A&M University, Technical Report No. 195.

Yielding, J. and Yansaneh, I.S. (1993). Weighting adjustments for income nonresponse in the U.S. Consumer Expenditure Survey. Texas A&M University, Technical Report No. 202.

Zvirblis, E. (1996). Development and progress in statistical data editing in the Lithuanian statistical offices, Lithuania.