11-522E
1990
c. 3


# SYMPOSIUM 90 

Measurement and Improvement of Data Quality

## PROCEEDINGS

APR 111995
LisRAITY
BIBLIOTHEQUE

## SYMPOSIUM 90

Measurement and Improvement of Data Quality

October 29-31, 1990
Ottawa, Ontario, Canada

## PROCEEDINGS

September 1991
Symposium 90 Organizing Committee
Mary March
Normand Laniel
Robert Lussier
Jeffrey Smith

Published by authority of the Minister responsible for Statistics Canada

- Minister of Industry,

Science and Technology, 1991

## PREFACE

Symposium 90 was the seventh in the series of international symposia held annually at Statistics Canada since 1984. Each year the symposium focuses on a particular theme. The 1990 theme was data quality.

The 1990 symposium was attended by over 350 persons from several countries who met over three days in the Simon Goldberg Conference Centre in Ottawa to hear experts in data quality from numerous government agencies, universities, and private industry. During the symposium, participants heard 30 invited papers presented in 10 different sessions which addressed a wide spectrum of quality issues including coverage measurement in censuses, measuring and improving the quality of administrative records, improving data processing and variance estimation to name just as few. In addition, John Early (Juran Institute) delivered the keynote address "Managing Quality in National Statistics Programs" on the first day and professor Carl Särndal (Université de Montréal\} gave the special invited lecture "Methods for Estimating the Precision of Survey Estimates When Imputation Has Been Used" on the final day.

Aside from translation and formatting, the Proceedings of Symposium 90 contain a record of the papers as submitted by the authors. The order of presentation of the papers is the same as that of the symposium itself.

The Symposium 90 Proceedings Sub-Committee would like to acknowledge the contributions of the many persons involved in the preparation of these Proceedings.

Naturally, thanks go to the presenters at Symposium 90 who took the time to put their ideas into words and submit them for inclusion in this volume. The efforts of many others were vital in the publication of these Proceedings. Processing of the manuscripts was expertly handled by Christine Larabie, Carmen Lacroix and Judy Clarke. Proofreading was done by numerous methodologists and subject matter specialists: B. Allard, J. Armstrong, S. Auger, Y. Beaucage, J.-M. Berthelot, R. Boyer, M. Bureau, J. Dumais, S. Giroux, H. Gough, S. Krawchuk, G. Laflamme, D. Lalande, E. Langlet, J. Mayda, S. Michaud, C. Morin, F. Pageau, G. Parent, G. Parsons, E. Rancourt, G. Reinhardt, G. St. Louis, L. Swain, J. Tremblay, P. Whitridge.

Statistics Canada's eighth annual international symposium will be held November 12-14, 1991 in Ottawa. The topic will be "Spatial Issues in Statistics".

Symposium 90 Proceedings Sub-Committee
Normand Laniel and Jeffrey Smith
August 1991

Extracts from this publication may be reproduced for individual use wilhout permission provided the source is fully acknowledged. However. reproduction of this publicstion in whole or in purt for the purposes of resale or redistribution requires written permission from Chief. Author Services, Publications Division. Statistics Canada. Ottawa, Ontario, Canada KIA OTG.

## STATISTICS CANADA SYMPOSIUM SERIES

1984 - Analysis of Survey Data
1985 - Small Area Statistics
1986. Missing Data in Surveys

1987-Statistical Uses of Administrative Data

1988 - The Impact of High Technology on Survey Taking
1989. Analysis of Data in Time
1990. Measurement and Improvement of Data Quality

## STATISTICS CANADA INTERNATIONAL SYMPOSIUM SERIES PROCEEDINGS ORDERING INFORMATION

To order additional copies of the proceedings of Symposium 90: Measurement and Improvement of Data Quality, use the order form on this page. A limited number of copies of back issues from the 1987, 1988, and 1989 symposia are also available. To order, complete this form and send it (or a copy) to:

```
SYMPOSIUM }90\mathrm{ PROCEEDINGS
STATISTICS CANADA
FINANCIAL OPERATIONS DIVISION
R.H. COATS BUILDING, 6't FLOOR
TUNNEY'S PASTURE
OTTAWA, ONTARIO
KTA OTG
CANADA
```

Please include payment with your order (cheque or money order, in Canadian funds or equivalent, payable to "The Receiver General for Canada - Symposium 90 Proceedings").

## SYMPOSIUM PROCEEDINGS: ISSUES AVAILABLE

1987. Statistical Uses of Administrative Data - ENGLISH
1988. Statistical Uses of Administrative Data - FRENCH
@ $\$ 10$ EACH
1987 - SET OF 1 ENGLISH AND 1 FRENCH
@ $\$ 10$ EACH
1989. The Impact of High Technology on Survey Taking - BILINGUAL $\qquad$ @ $\$ 10 \mathrm{EACH}$
1989-Analysis of Data in Time - BILINGUAL $\qquad$ @ $\$ 20$ EACH

1990- Measurement and Improvement of Data Quality - ENGLISH
@ \$20 EACH
1990. Measurement and Improvement of Data Quality - FRENCH

1990- SET OF 1 ENGLISH AND 1 FRENCH $\qquad$ © $\$ 20$ EACH

PLEASE ADD \$2 PER VOLUME FOR SHIPPING
\$
TOTAL AMOUNT OF ORDER
\$
Prices include GST; GST registration no. R121491907)

HIGHLIGHTS OF THE 1989 PROCEEDINGS: Analysis of Repeated Surveys (W.A. Fuller), Unique Features and Problems of Rolling Samples (L. Kish), A Time Series Model for Estimating Housing Price Indexes Adjusted for Changes in Quality (D. Preffermann, L. Burke, S. BenTuvia), Analysis of Seasonal ARIMA Models from Survey Data (D.A. Binder, J.P. Dick), Small Area Estimation Using Models that Combine Time Series and Cross-Sectional Data (G.H. Choudhry, J.N.K. Rao), Mapping Aggregate Birth Data (D.R. Brillinger), Analysis of CrossClassified Categorical Time Series (A.C. Singh, G.R. Roberts), Alternative Approaches to the Analysis of Time Series Components (W.R. Bell, M.G. Pugh), Adjustment for Reporting-Delay of AIDS and Estimation of the Size of the HIV Infected Population in the U.S.A. (I.B. MacNeill, Q.P. Duong, V.R. Jandhyala, L. Liu), Some Statistical Methods for Panel Life History Data (J.D. Kalbfleisch, J.F. Lawless).

PLEASE INCLUDE YOUR FULL MAILING ADDRESS WITH YOUR ORDER
NAME
ADDRESS

CITY $\qquad$ PROV/STATE $\qquad$ COUNTRY $\qquad$
POSTAL CODE $\qquad$ TELEPHONE $\qquad$ FAX $\qquad$

Please note: Each Symposium 90 registrant not employed by Statistics Canada receives one free copy of the Symposium 90 Proceedings

## TABLE OF CONTENTS ${ }^{1}$

OPENING REMARKS ..... 3
1.P. Fellegi, Chief Statistician of Canada
KEYNOTE ADDRESS
Chairperson: B. Slater, Statistics Canada
Managing Quality in National Statistics Programs ..... 9
J. Early, Juran Institute
SESSION 1: The Challenge of Reducing Resources and Improving Data Quality
Chairperson: C.D. Cowan, Opinion Research Corporation
The Application of the Total Survey Design Approach to Determining Strategies for Resource Allocation in the Survey of Motor Vehicle Usage ..... 23
S. Linacre and P. Bell, Australian Bureau of Statistics
An Experience with a Productivity Incentive Plan at Statistics Canada ..... 33
J.-F. Gosselin, Statistics Canada
A Systematic Approach to Quantifying the Quality of Repeated Surveys ..... 43
R. D. Tortora, U.S. Bureau of the Census
SESSION 2: Data Integration
Chairperson: J.F. Gentleman, Statistics Canada
Integration of Economic Data: Benefits and Problems ..... 51
M. Colledge, Statistics Canada
Data Quality Checks for Telephone Usage Data paper not provided
S.T.T. Au, AT\&T Bell Laboratories
Meta-Analysis: Current Issues in Research Synthesis paper not provided
I. Olkin, Stanford University
SESSION 3: Coverage Measurement in Censuses of Population
Chairperson: E.T. Pryor. Statistics Canada
Differential Coverage in the United States Census of Population: A Historic Review ..... 67
G. Robinson and H. Hogan, U.S. Bureau of the Census
The Measurement of Net Coverage Error in Canadian Censuses ..... 79
R.G. Carter, Statistics Canada
Small-Area Prediction of Undercount Using the General Linear Model ..... 93
N. Cressie, Iowa State University

[^0]SESSION 4: Measuring Survey ErrorChairperson: S. Linacre, Australian Bureau of Statistics
Measurement of Content Data Quality in the 1990 Census ..... 109
H. Woltman and K.F. Thomas, U.S. Bureau of the Census
Comparing Three Bootstraps for Survey Data ..... 117
R.R. Sitter, Carleton University
Simple Correction for the Bias of Traditional Variance Estimators ..... 133
Y. Leblond, Statistics Canada
SESSION 5: Improving Data Collection
Chairperson: M. Levine, Statistics Canada
Enhancing the Quality of Time Critical Estimates Through the Use of Mixed Mode CATI/CASI Collection ..... 145
G.S. Werking and R.L. Clayton, U.S. Bureau of Labor Statistics
Follow-up Strategy for Economic Surveys ..... 155
J.-M. Berthelot and M. Latouche, Statistics Canada
The Quality of Data in Crop Yield Surveys ..... 167
R. Fecso, U.S. National Agricultural Statistics Service
SESSION 6: Measurement and Improvement of Quality of Administrative Records When Used as a Replacement to Traditional Survey TakingChairperson: F. Scheuren, U.S. Internal Revenue Service
A Computerized Quality Review System and the 1988 Statistics of Income
Partnership Program paper not provided
G.E. Moglen, U.S. Internal Revenue Service
Canadian Taxfiler Family Data: Assessment of Qualitative Criteria and Trends ..... 175J.M. Leyes, Statistics Canada
Survey Methods for the Evaluation of National Cause of Death Data paper not provided
L. Curtin, F. Chevarley, U.S. National Center for Health Statistics,
P. Royston, U.S. Health Resources and Services Administration, and
D. Gibbs, Research Triangle Institute
SESSION 7: Improving the Quality of Frame Data and Their Use
Chairperson: D. Pfeffermann, Hebrew University
Sampling Flows of Mobile Human Populations ..... 187G. Kalton, University of Michigan
Improving the Quality of the U.S. Census of Agriculture List ..... 197
C.Z.F. Clark, U.S. Bureau of the Census
Improving the Quality of Standard Statistical Establishment List Data ..... 211
P.S. Hanczaryk and M.L. Trager, U.S. Bureau of the Census
SESSION 8: Improving Data Processing and Estimation
Chairperson: J. Coombs, Statistics Canada
A Review of Some Macroediting Methods for Rationalizing the Editing Process ..... 225L. Granquist, Statistics Sweden
On-line Capabilities in SPEER (Structured Programs for Economic Editing and Referrals) ..... 235L. Draper, B. Greenberg and T. Petkunas, U.S. Bureau of the Census
M-Estimators and Outlier Resistant Alternatives to the Ratio Estimator ..... 245
L.-P. Rivest and E. Rouillard, Université Laval
SESSION 9: Are Our Statistical Products Fit for Use?
Chairperson: R. Lussier, Statistics Canada
Moving From Descriptive Statistics to Inference ..... 261
K. O'Conor, U.S. Internal Revenue Service, B.K. Atrostic, and R. Gillette, U.S. Department of the Treasury
An Examination of Statistics Canada's Data Quality Release Criteria ..... 273R. Burgess, Statistics Canada
Statistics: The Balance Between Desired Precision and Intended Use ..... 287B.L. Khuong, Bureau de la Statistique du Québec
SESSION 10: Quality Assurance
Chairperson: D. Beecroft, University of Waterloo
Applying Industry Quality Improvement Techniques to the Statistics Canada Environment ..... 295
D. Williams, D.N. Williams and Associates
Weighted CUSUM Schemes and Their Applications ..... 303
E. Yashchin, IBM Corporation
Techniques to Control and Improve Quality of Data in Large Databases ..... 319
T.C. Redman and R.W. Pautke, AT\&T Bell Laboratories
SPECLAL INVITED LECTURE
Chairperson: G. Brackstone, Statistics Canada
Methods for Estimating the Precision of Survey Estimates When Imputation
Has Been Used ..... 337
C.E. Särndal, Université de Montréal
CLOSING REMARKS ..... 351
G. Brackstone, Statistics Canada
NOTE: The original language of all Symposium 90 papers was English except forthe following: Opening Remarks, I.P. Fellegi (bilingual); Session 4,Y. Leblond (French), Session 5, J.-M. Berthelot, (French); Session 9,B.L. Khuong, (French).

OPENING REMARKS

Proceedings of Statistics Canada Symposium 90
Measurement and Improvement of Data Quality
October 1990

## OPENING REMARKS

I.P. Fellegi ${ }^{1}$

Welcome to Symposium '90. A methodology symposium has become an annual event at Statistics Canada. Since 1984, we have organized seven symposia, each one focusing on a different methodology topic. Topics covered in previous symposia were:

The Analysis of Survey Data;<br>Small Area Statistics;<br>Missing Data in Surveys;<br>Statistical Uses of Administrative Data;<br>The Impact of High Technology on Survey Taking; and<br>The Analysis of Data in Time.

This year's symposium topic is "Measurement and Improvement of Data Quality". In choosing this topic, we did not want to limit ourselves to considering the quality of survey data only. All data that are collected, processed, analyzed and used would be in the scope of this meeting. In addition to survey or census statistics, data collected for some administrative purpose are very much in scope. But so are also measurements taken as part of an experiment, or measures used for a diagnosis (e.g. blood pressure, heart rate or measures obtained from the machine that the mechanic hooks up to your car to diagnose its ills).

World-wide concerns about quality indicate that this is an issue whose time has come. Quality has become a key advertising slogan, in preference to other attributes such as styling or price - such as "Quality is Job 1 "(Ford) or "People Delivering Quality" (CP Express) or "Quality. Who could ask for anything more?" (Toyota).

In a few days, winners in the Quality Category of the Canada Awards for Business Excellence will be announced. This award, like Japan's "Deming Prize" and the U.S. "Baldridge" and "NASA Supplier" awards, is now a prestigious and much sought-after distinction.

In fact, quality of our goods and services is an issue that has assumed national importance. In declaring October to be Canadian Quality Month, Prime Minister Brian Mulroney has said that: "Every Canadian should do his or her part to make Canada synonymous with quality."

One may logically ask: why this burgeoning interest in quality? The post-war mentality preferred disposability: the short term disposability of paper napkins, diapers, plastic goods, glass and tin containers, and the medium term disposability implied by model years, drastically changing fashions, suburban homes and office buildings designed to be written off in 15-20 years. These days, the rapid growth of disposable incomes is over, and at the same time there are heightened concerns about environmental issues and sustainable development. These factors fundamentally re-oriented consumer preferences towards quality, durability, dependability, and long term economies.

At the same time the Japanese have proved that building in quality is fundamental to being competitive. This is partly due to the fact, briefly indicated above, that quality is increasingly demanded by consumers. They have also realized that designing a production process that builds quality into a product is more economical than relying on final-stage quality control, even from the point of view of pure cost-efficiency.

[^1]A fundamental aspect of building quality into products involves measuring errors throughout the production process and designing a production system that keeps these errors under control. The first of these tasks is explicitly statistical - allthough in order to be able to measure errors the statistician must attempt to understand the underlying processes. The second task - keeping errors under control - must critically involve statisticians, together with production staff, engineers and managers. The statistician's role here is not only to design process control charts but to help with an intelligent balancing of the costs end and errors through the various stages of processing - what we in survey statistics call minimizing total error for a given cost.

So the issue of quality is in a very basic sense statistical - even though it cannot be addressed by statisticians working alone.

In our own work, as official statisticians, we have a long history of working to improve data quality in the dual sense outlined above. The early history of probability sampling was motivated by the desire to develop surveys whose errors (in this case sampling errors) are measurable, hence potentially controllable. Following the early breakthroughs in sampling, the long and arduous process of improving other aspects of quality began. Some major examples are: research on sampling frames and registers to improve coverage; measuring and controlling coverage errors; the development of models to measure, explain, and reduce the impact of interviewer errors; cognitive research to try to understand the reactions of individual respondents to questionnaires and methods of enumeration with the aim of reducing reporting errors; the development of sophisticated methodologies of editing and imputation to reduce individual response and non-response biases; survey data analysis which permits the estimation of second moments even for complex designs; and a variety of quality assurance methods necessary to certify the final product as being "fit to use".

As government statisticians we bave a deep motivation for improving and maintaining quality, and this goes beyond the drive to apply our tools of trade to our own practice. This derives from the fact that information whose reliability is highly doubtful becomes basically useless. Yet, few users have an opportunity to replicate or otherwise directly assess the quality of our output: they must rely, in the final analysis, on the reputation of the producing agency. Our reputation for producing reliable statistical information is therefore a direct determinant of the usefulness of our output. Reputation, therefore, becomes quite crucial. But it is exceedingly difficult to maintain: the sheer volume of output and the multiplicity of users renders us particularly vulnerable to well publicized "traffic accidents" - not to speak of users of data for purposes they were not designed. Our professionalism in the face of such problems, indeed in anticipation of them, will determine whether such occasional incidents damage our reputation.

An important aspect of this type of professionalism is our ability and willingness to describe fully our methods and our typically incomplete knowledge of data limitations. These are, in fact, the rationales behind Statistics Canada's Policy on Informing Users of Data Quality and Methodology which requires that all data released should be accompanied by a description of the concepts and the methodology used to obtain them and by an assessment of their quality.

A tremendous amount of work has been done and we should be proud of our successes.
I think all of us would agree, however, that a lot of work still can be done to improve both the quality or fitness for use of our data as well as our ability to measure and report quality. Development work also continues to be needed in the areas of data collection, use of administrative records, data processing and estimation, survey frames and quality control of collection and processing. There is also room for improvement in our methods of measuring coverage, content and sampling errors. We need to substantially improve our ability to optimize complete survey systems - indeed, our knowledge here is entirely rudimentary. We have to expand our ability to juxtapose data from different sources so that their value is improved, and we need to learn to exploit the quality enhancing capabilities of such integrated information systems.

Lastly, but not least, we have to develop our management skills to adequately manage the ways in which we plan for quality and our handling of our quality improvement efforts. In looking for ways to manage our quality improvement programs we will undoubtedly find that we can learn a lot from industry - and perhaps vice-versa. This is why meetings such as this one are so important.

We are very fortunate at this meeting to be able to draw on an especially rich mix of experiences. We have papers from Canada, Australia, Sweden and the United States. There are representatives from government, industry and the academic world. Clearly, the organizers have hit upon not only a fashionable topic, but one where there is a great deal of interest and relevant work going on. I welcome all of you and express my sincere hope that you will profit by your attendance at this Symposium - as I know Statistics Canada will.

I would like to express our special thanks to the Carleton University and University of Ottawa Laboratory for Research in Statistics and Probability for their co-sponsorship of this event. Their practical assistance, advice, and financial and moral support have been greatly appreciated.

KEYNOTE ADDRESS


# MANAGING QUALITY IN NATIONAL STATISTICS PROGRAMS 

J.F. Early ${ }^{\text { }}$

## WHY QUALITY?

Why are we here? Without intending to be presumptuous, let me suggest four of many possible motivating factors. First, as professionals, we all have a desire to maintain and improve the performance of our disciplines. Second, statistical agencies are under considerable pressure to improve their performance. Third, there is a rapidly growing demand for more and better data. And finally, there is a growing awareness throughout the world that quality has been neglected in all aspects of modern industrial society.

We live behind quality dikes holding back the deluge of disaster. Quality stands between us and the melt-down of a nuclear reactor core. Quality stands between us and the crash of a jumbo jet. Quality stands between us and a prolonged loss of electric power over wide geographic areas.

In national statistics, the integrity of our quality dikes help protect us from misinformed or erroneous decisions on:

- Reductions (or increases) in interest rates
- Tax increases (or decreases)
- Creation (or elimination) of costly public programs to "fix" the pressing social problem in
-- Health
-- Education
-- Criminal justice
- Making (or withholding) capital investments
- Employer decisions to hire, layoff, train
- Or even individual decisions on whether to move, take a new job, or go back to school.

In all these and many other cases, disasters both large and small are avoided partly by the quality of the data available for the decisions.

Today, I am going to discuss the role of management in quality. In general, I believe we have far greater technical capability to achieve quality than we have managerial ability to create significant quality improvement. Much of what I am going to say may not look new to you. You may be doing much of it, especially at the technical level. What I will describe are managerial processes that need to be observed systematically and comprehensively.

## HOW TO THINK ABOUT QUALITY

But just what is quality? One common short definition is "fitness for use." That immediately raises the question of "Whose use?" Here are some of the generic users we must consider.

The processors: Those who take our data and combine or adapt it in different ways to create new information.

[^2]In-house or internal: Those who work with us are every much customers of what we do as those external folks whom we may think of as out clients.

The merchant chain: These are the people who are the channel of distribution. Now that is an easy enough concept when your are thinking about quality for breakfast cereal or television sets. You want to satisfy the needs of the merchant chain for ease of handling and selling, so that they will be eager to sell your product. But there are also distribution chains for national statistics: the press, the universities, the professiomal societies.

The ultimate user: Those who use that data for literally thousands of applications. This is not a homogeneous set of people, it is a varied and even divisive cast of characters.

Society at large: Even those who mever see one of our numbers are affected--their incomes, their health, and their security.

In the quality disciplines we have adopted the convention of calling all those who are affected by a product the customers. Many organizations who have not traditionally thought about their customers will sometimes find this to be an awkward term and adopt another one. There is certainly no magic in the specific word, but thinking about all those who are affected by our work as our customers can be a very enlightening experience.

Dividing customers into internal and external is useful because it reminds us to look in both places. Believe it or mot, many sincere but misguided organizations have failed in their quality revolution because they truncated one or the other of these two groups. Some become so inward directed that they forget the principle reason for their existence is service to the external customer. Others forget the converse, that we cannot expect our employees (our internal customers) to serve the external customer any better than they themselves have been served internally.

Among our external customers, we must remember the needs of:
Those who analyze the data to reach various critical decisions.
-- Those who analyze primarily for public policy purposes
-- Those in academia who may have a policy interest, but also have many investigatory interests beyond immediate application
-- Users making business decisions throughout the economy

- Some customers use our statistics to create new statistics--for example the gross national product estimates depend on many other economic statistics, most population-based surveys require controls and sampling frames from population censuses and surveys.
- Professional societies are customers not only of the actual estimates that we produce but also of the methods and documentation. The customer professions include not only the appropriate statistical science, but also the subject matter professions: public health, law enforcement, environmental sciences, economics, etc.
- I mention society at large once again to highlight the critical impact our statistics have on millions of people who do not articulate their needs well in either the technical or the political jargons which we use most of the time.

Now a national statistical agency faces some unique specific problems in meeting its customer meeds and these should not be minimized. Dealing with internal customers is essentially the same as for any other organization, but external customers raise additional issues. Most organizations face multiple publics with conflicting demands, but a government agency cannot decide unilaterally to serve only a subset of these. And while there is some room to produce alternative sets of information, in the end there are certain data that can exist only as a single authoritative answer. There can be only one official census count. The official rate of inflation for all consumers
must be exactly one number. In the market place a company can segment and differentiate, that is not an option generally available to government.

A political scientist David Easton once described the political system as the "authoritative allocation of values." The census count dictates voting strength in elections and allocation of national budgetary resources. The estimate of inflation controls cost of living adjustments in transfer payments. Market penetration estimates trigger retaliatory trade restrictions.

It is this unitary authority on the one hand, and the important values at stake on the other that make customer satisfaction more difficult for a public agency than a private market-drive corporation.

Elected officials are one class of customers that are especially important on the one hand, and especially difficult to deal with on the other. While they represent the needs of the rest of our customers, they cannot generally be relied on to be effective in that regard. This observation is not intended as an elitist view that somehow "we" know what the customers need better than "they" do. Rather it is simply the realization that while elected official should and do set many of the broad policies that define our customer needs, there is much that they will never do in that regard.

Quality, then, is exactly what all these various customers say it is--what they think is fit for their use.

## TWO SPECIES OF QUALITY

But that fitness for use really breaks down into two different species: features and absence of deficiencies. This is not an academic distinction. It is very important because the way one manages quality is directly affected by this distinction. What one does to reduce defects is very different from what one does to optimize feature quality.

Once again I must warn against some of those well meaning but misguide folk heroes of quality who seem to be constantly lurking on the horizon. Often they have been unable to distinguish between these two facts of life. As a result, their prescriptions, while flashy, tend to be confused, shallow and destined for disappointment if one is really seeking world-class quality.

Product features are what respond to the customers' needs. (And by product I mean both good and services because they are both produced.) The features we offer are why they come to us. It is these features that satisfy their needs. In the market place, the company that offers the better the package of features will gain greater market share and/or be able to charge a premium price. In terms of features, higher quality generally cost more. First class air travel costs more than tourist. Jaguar automobiles cost more than Hondas.

Some of the quality features of national statistics include the following:
Frequency: Monthly estimates may be more useful than annual estimates, but they also cost more.
Timeliness: All other things being equal, estimates released within 20 calendar days of the reference date are to be preferred to those released 40 days later. But the more timely estimates may require greater resource expenditure.

However, here there is a relationship between feature quality and absence-of-defect quality. We may be able to speed up preparation of the data, in part, by reducing the number of defects in our processing that consume time and resources to correct.

Precision: The sampling error is driven by the sample size and design. In general, lower sampling error will be more expensive.

| Accuracy: | For some types of design biases, the decisions relate to feature quality. Explicit trade- <br> offs that minimize root-mean-squared error and accept some bias as the price for <br> much lower variance, are features of a particular statistic. Other kinds of biases that <br> arise from ineffective operations are defects and must be addressed as such. |
| :--- | :--- |
| Concept: | Because we are a "statistical" agency, we may become overwhelmed with the technical <br> statistical issues and not give sufficient attention to whether the best concept is being <br> measured in the first place. For example, a consumer price index should measure <br> price changes for consumption, and not include capital costs. |
| Change, Level, |  |
| or Longitudinal: $\quad$Do the customers primarily need measures of change in a phenomenon? Measures <br> of the current level? Or be able to study gross flows through longitudinal data? The <br> way we do our work is greatly affected by these choices. |  |
| Documentation:The amount of documentation provided users is a feature, I would say it is a very <br> important feature, but it must be considered in the total cost of the statistics. Freeing <br> the documentation from defects is another matter. No matter what is provided it must <br> be correct. |  |
| Analysis:Some of the national statistics policy debates swirl around how much analysis should <br> be provided and by whom. This discussion needs to be viewed in the context of <br> satisfying our customers, not ourselves. |  |
| Media/mode of | How does the customer want the data: Glossy report? Massive tables? Short <br> summaries? On floppy disk? On electronic bulletin board? Ask the customer! |
| Delivery: |  |

Freeing our statistics and services from deficiencies avoids dissatisfaction by our customers. Now satisfaction and dissatisfaction are linguistic opposites, but they are not operationally opposites in the market place or in the work of a statistical agency.

A customer buys a first-class airline ticket to get china plates, 4 course meals, and wide seats. These satisfy his or her travel needs. But if the plane arrives two hours late because of "mechanical difficulties," no amount of amenities will erase the fundamental dissatisfaction.

Or consider automobiles again, we said that the difference in price between a Jaguar and a Honda was a difference in feature quality. If you want premium leathers and genuine walnut on the dash board and lots of sound insulation and lots of admiring looks at the yacht club, you can satisfy those "needs" with a Jaguar, you will be mighty dissatisfied, however, if it doesn't run!

Now for the key point about quality as absence of defect, higher quality in this sense always costs less! If you do it right the first time, you do not need to spend the resources to check it, fix it, and mollify the irate customer (whether among the public or within your own agency). Incorrect data, late responses, inappropriate interpretation, and discourteous handling of inquiries are all defects.

## COSTS OF POOR QUALITY

Consider for a moment all the activities that your agency engages in because something was not done perfectly the first time. Also think about all those costs you incur because you know that something is likely to go wrong and you want to find that something before your customer does. Those are the costs of poor quality. In manufacturing companies with no integrated total quality management process driven from the top executives, those costs run $25 \%$ of sales or more. In banks, those costs run $33 \%$ of total operating revenues and higher. And in one statistical program, a short search uncovered at least $30 \%$ of the budget consumed by the costs of poor quality.

Not in my shop you say? That's what they all say. One reason some of these costs are not readily visible is because if the murder is committed in one office, the body often surfaces in another. You usually wind up fixing my errors.

Here are some examples of costs of poor quality in a statistical agency, arranged by the three standard classes of costs of poor quality.

Inspection: Costs associated with reviewing data, double keying data to verify it, any kind of proof reading. These costs would be reduced or eliminated if we were confident that the original work had been correct.

Internal Failures: Computer reruns are a very expensive and time consuming costs of poor quality.
Any time resources stand idle because the preceding activity was late or incorrect we have a cost of poor quality. Field staff waits for samples to collect, computer operations waits for data, and sample designers wait for the unit variances.

Samples with high non-response rates or lower-than-expected yields for other reasons are costs of poor quality.

External Failures: Publishing incorrect data means that the correct statistic must be calculated, then published. Questions must be answered for the users and complaints satisfied.

The Consumer Price Index program in the United States in the early 1980's averaged about 10 corrections per year to published data-never to a major item at the national level, almost always to local area data. The work associated with making these corrections and supporting the users cost about one-quarter million dollars. Over a three year period, systematic quality management eliminated these corrections to published data completely.

Late release of data usually creates additional costs--overtime, messenger services, express printing charges, and above all many more special inquiries from users.

Effective management of our feature quality can raise the satisfaction of our customers and improve our service vis-à-vis our budgeted appropriations. Reducing deficiencies will lower costs by eliminating wasted effort and rework and minimizing the costs of user complaints. Reducing deficiencies also creates more responsive service. The lowered costs from reduced deficiencies may help us manage effects of fiscal tightening better, but, on a more positive note, it also allows us to fund more service to our customers. Our customers are happier and our budgets healthier if we manage quality effectively.

## A STRATEGY FOR QUALITY

Most managers have a fairly clear picture of how one manages finance.
First, one develops a financial plan-a budget of what financial resources are going to be spent on what activities.

Second, one controls financial activities by that plan. Reports are developed showing income and spending rates by appropriate categories and actions are taken to bring them into conformance with the plan when necessary.

Finally, projects are frequently undertaken to improve financial performance--such as specific object class costs reductions.

Most managers are also familiar with one or more of the standard processes for managing projects-critical path methods, PERT charts etc. Project management has the same three basic steps.
(1) Planning the project.
(2) Using the plan to control the project and keep it on schedule.
(3) Setting about to improve project performance by establishing task forces to eliminate bottlenecks or identify tasks that can be performed in parallel.

In short, most good managers know how to manage for finance (the cost of what they do) and how to manage the project activities (the time it takes to do them). Most managers, however, have not learned the strategy for managing quality. Fortunately, it involves the same three basic process-planning, control, and improvement.

Quality Planning begins by identifying our customers and their needs. We have already talked about some guidelines for doing that in the statistical program enviroment.

Those needs should be obtained in the customers' language, don't expect them to speak ours. Then we must explicitly translate those needs into our language.

Next, we need to develop product features that meet those customer needs. For example, in order to meet customer needs for having estimates of both level and rate of change in the labor force we might establish as critical features the root-mean-squared error (RMSE) of both level and change and then set specific goals for each.

Next, we need to develop a process that will meet those goals. Specific sample designs and estimators are developed that will enable us to meet our RMSE product feature goals. The big difference in this step and "business as usual" in survey design is that it is imbedded in the overall quality planning framework. On the front end, the design is explicitly tied to the needs of customers. On the back end, the design is proved capable of delivering the product before it is turned over to operations.

In business as usual, we tend to slight the two ends of this process, customer needs are either assumed or taken only from one small set of customers. At the other end, we rarely give the needed attention to demonstrating that a process is capable of providing the desired result before we turn it over to operations. And throughout we often ignore the need for participative planning in preference for hallowed "expert opinion."

Quality Control requires that we identify the control subjects--what we want to control-and the appropriate units of measure. These choices are very important. The control subjects must be variables that are important to the quality of the process. In fact, there selection is really part of the quality planning process as well. There are two rules for selecting control subjects. First, each control subject must have a direct process tie to critical customer needs. Second, each critical customer need must be met by a product feature, which must be delivered by appropriate process features, which in turn is a control subject.

One can vigorously control a trivial parameter without much effect on the quality of the product in the customers' eyes. It may be culturally and mechanically easy to control the typographical errors in executive correspondence. So we control them with great passion. The weightier matter of the content of these messages to our customers may, however, be harder to deal with and, thus, tends to be ignored. Executive management must insure itself that the important matters are under control.

The next steps are classical quality control--(1) establish a standard, (2) measure the actual performance, (3) interpret the difference between standard and actual, and (4) take action on that difference to bring the process within the established operating characteristics.

Quality Improvement is the process for making things better. It begins with identifying the most pressing and appropriate quality problems facing the organization. A project team is then organized to address each identified problem.

These teams must conduct and/or guide the work to diagnose the causes of the observed problem and develop appropriate remedies for those causes. The effectiveness of these remedies must be proved, cultural resistance to the change must be overcome, and then controls established to hold the gains that have been achieved.

For most of us, the core steps of diagnosing cause before applying a remedy are familiar. We do them in our technical work all the time. What is new here is (1) on insistence that the same rigor be applied to our managerial and operational processes and (2) that those affected be involved in the process.

Many clerical processes have production-line characteristics and are obvious easy targets for improvement projects. Of course, it makes sense to improve these processes. But I would caution strongly that improvement projects should be quickly targeted on professional and design processes as well. This is important for two reasons. First, the clerical staff must not feel that quality improvement is merely an opportunity to pick on them. Second, and even more important, professional and technical processes will often be where your biggest payoffs will be.

Experience is clear in this regard. At least 80 percent of quality problems lie in the design of the process. In my experience, it is also clear that in statistical agencies the design of processes for professionals--statisticians, economists, accountants, system analysts, etc.--require just as much quality improvement and can be addressed just as effectively as the more repetitive operations. It does require more imagination and greater inter-personal skills sometimes, but the results are worth the effort.

This quality improvement process is really the best place to begin your quality management. Improvement comes one project at a time, so it is easy to get started and some results can be achieved carly. Acceleration in quality improvement simply comes by increasing the number of improvement projects that are active at any one time. You will eventually discover that some improvements come best through a quality re-planning of the process. And then you will be in a position to begin serious quality planning for new efforts.

The Juran Trilogy of Ouality Management Processes


This figure illustrates the relationships among the important elements of quality management that we have named The Juran Trilogy. The vertical scale is the cost of poor quality, that is, the costs associated with finding
and fixing work that was not done incorrectly the first time. This chart shows that for our sample process that cost started out at about $20 \%$ of total costs.

This initial $20 \%$ cost of poor quality is the result of the way in which the process was planned, planning being the activity to the left of the vertical axis. When a sporadic spike raises poor quality to $40 \%$, quality control quickly returns the process to the established control zone. Quality control, however, cannot make the process better. Only a quality improvement activity that diagnoses the causes of the high costs of poor quality and develops appropriate remedies can make the process fundamentally better. After the efforts of a quality improvement project team, a new zone of quality control is established in the neighborhood of $3 \%$.

In the course of improving quality, we usually learn some important lessons about the ways in which our processes work and what causes poor quality. If we systematically evaluate these lessons, they form a valuable data base which can be used in the quality planning of the next new process.

## QUALITY AND UPPER MANAGEMENT

A truly prodigious amount of ink has been spilled over the state of quality in national statistics. At this point my perspective may be a bit ethnocentric, but most of these discussions can be grouped into four categories:

1. The budgetary fix. Spend more money.
2. The structural fix. There are many variations on this one, usually reflecting the place of the speaker in the cast of customer characters. Some have included:

- More central direction and control.
- More independence.
- A stronger role of the research function within the agency, and a weaker role for short-sighted policy concerns.

Stronger ties to the policy needs.
3. The data fix.

We need more data on the international sectors. No we need more micro data on small geographic areas.

We must stop destroying historical continuity by changing methods and classifications. Classifications and methods are hopelessly out of date and must be changed at once.

We need more timely and accurate macro-economic indicators. That is a hopelessly misinformed view, everyone knows we need longitudinal data bases.
4. The technical fix.

The defects of the U.S. decennial census would go away if only the Census Bureau would adopt what "every statistician knows" are better methods.

We need a central, common establishment universe file.
We need a CATI system.
Over the years I have found myself in the middle of many of these and other similar debates. And a few of these issues are of considerable importance.

But none of them will do much to improve the quality of national statistics unless the top managers of national statistical agencies lead major revolutions in the way they manage. Quality is only as good as the top management leadership in that direction.

There are specific tasks which upper managers must undertake if data are to become substantially better. We are going to look at a few of those in a moment, but first we need to understand why quality is an upper management responsibility.

## THE ORGANIZATION AND QUALITY

Let's consider the National Statistics Agency for the Federation of Eastern Antarctica--or for just about any place I ever worked or visited. This structure is the evolutionary result of two or more centuries of industrial organization. It has two prominent features. The work flows back and forth, but the authority and much of the communication flow up and down. No wonder we have a quality problem. What we almost invariably find when we look at quality problems is that they are created in the spaces between boxes on the organization chart.

Each box is responsible for its work. The standards for that work are often generated internally. If there are external standards they often relate to "professional" standards for the discipline which that box represents. Those standards rarely relate to the needs of the next box in the process, and even more rarely to the needs of the final customer. Because no one is responsible for all that white space between boxes. No one that is except upper management. That is why quality is upper management's responsibility first and foremost.

In order to address the most important quality problems, which are almost invariably cross-functional, upper management must act as a quality council to direct the effort. They must identify the most important quality problems and charter cross-functional teams to attack each problem. Without the leaders of all of the major organizational functions involved in the quality council, their employees will not bave the support needed to serve on the project team to attack the identified problem.

Vast experience has shown us that only on a project-by-project approach to improving quality can we hope to make significant progress.

These are some of the types of quality improvement projects that have been completed successfully in statistical programs.

- Reduction of the number of errors in published data form 10 per year to zero. That made users happier and saved a quarter million dollars.
- Culting the time it takes to provide the field with clarifications to procedures by $80 \%$. That made the staff happier and reduced the need for later corrections.
- Elimination of interest and penalty payments on benefits for separated employees. This "mere administrative" matter had drained tens of thousands of dollars from programmatic initiatives.

Cut the number and cost of computer reruns by more than $\$ 100,000$.
Reduced the amount of time spent in the field reviewing collected data and also lowered the error rate by improved self-edits and error-resistant instruments.

- The cumulative effect of about $21 / 2$ years of projects on the U.S. Consumer Price Index Program was to eliminate all published errors, release the data 2 days ( $10 \%$ ) faster, and provide data with $20 \%$ lower sampling error--all with $7 \%$ fewer resources.

Quality planning also requires cross-functional teams chartered by upper management councils. All too often we have relegated planning to some set of "experts"--the subject-matter specialists, the statisticians, the program planning office, etc. All generally fail miserably at the task of planning for quality.

First and foremost, they have not been trained to plan for the multiple parameters of quality. They have no appreciation for the role of the customers (both internal and external) nor for the need to establish process capability. Second, they are being asked to plan in a vacuum that which can only be planned successfully with a dedicated, cross-functional team.

Here are samples of quality planning projects for a statistical agency.

- Replacing an old computer system.
- Designing a process for professional review and adjustment of basic data.
- Designing the process used to design multiple different samples for similar surveys.
- Revising an exosting statistical series.
- Introducing a new statistical series.
- Developing a CATI (computer assisted telephone interviewing) system.
- Providing an electronic bulletin board for current data.

Quality control is often thought of as a technical rather than managerial activity. There are certainly many useful technical dimensions to the topic, but one of the differences between organizations that merely talk about quality and those who really do something about it is that the latter groups practice quality control as a management activity.

The key management constructs for quality control are two. The first we have already discussed--namely, that all critical customer needs must be related to a control subject and that all control subjects must be related to a critical customer need.

The second key management principal for quality control is the concept of the state of self control. This is not a psychological state. It is the managerial state of a process that has three conditions:

1. The worker knows exactly what is expected.
2. The worker knows in a timely way exactly how he or she is performing.
3. The worker has the ability to regulate the process, namely:
-- The process is capable of doing what is required
-. The worker has the tools, knowledge, and skill to regulate the process
-- The worker has the authority to regulate it.

## TASKS FOR UPPER MANAGERS

In addition to serving actively on the quality council to identify and charter project teams that make targeted quality improvements or to develop new or replacement products using quality planning, upper managers have some other specific tasks which they alone can perform.

They must approve strategic quality goals that are an integrated part of the agency's strategic plans. These goals might include:

- Reduce the frequency of published corrections by specific amounts.
- Reduce the time lag from reference period to release of the data.
- Lower the costs of poor quality by a specified amount.
- Increase survey response rate to a particular target value.
- Improve the satisfaction ratings that users give our service.
- Become recognized as the preferred source for a specific type of data by a particular user community.
- Increase the scope and precision of the coverage which the press give to our data.

These strategic goals, however, remain just a wish list if the upper managers do not take specific actions to "deploy" these goals to the individual operating units and to the needed cross-functional teams.

Even this assignment of specific actions to meet the goals will be for naught if the assignment is not followed up with active review of progress and recognition of jobs well done. Among the most powerful leadership tools that the upper managers have is to lead by example--serve on some project teams themselves.

And, of course, we all know that upper managers must provide the necessary resources. They must free up individuals' time to work on quality. They need to provide facilitators to these new cross-functional teams that are learning new ways of doing business. They may need to provide the teams with the diagnostic support they need to find the root causes of problems--computer time, special expertise, etc.

Since this is a revolution in the way business is done, everyone, including the head of the agency, will need training. Goals, actions, and progress will need to be communicated effectively throughout the organization.

Progress must be reviewed actively by the top managers and recognition given to those who are providing the change. Because quality now has top priority, there will also need to be changes in the merit rating systems and the ways in which people are evaluated and rewarded.

## A PARTING THOUGHT

With possibly a couple of exceptions, most of the papers that you will consider at this conference will be technical and operational methods that can contribute to improvements in our national statistical systems. Once or twice during the conference, after you have heard a particularly appealing idea, pause and ask yourself some questions about the quality management dimensions of what is being discussed. For example, you might ask:

- Do you know what customer needs the proposal will satisfy and how important those needs are? How do you know this?

How would you demonstrate the capability of the process before implementing it?
How would you address the internal customer needs of those who operate, support, and use the results of the proposal?

- If the proposal is intended to remedy some current quality defects,
-- Is this a vital defect? If so, how do you know?
-- Has the defect been clearly defined?
-- Has the root cause of that defect been demonstrated with data and the participation of all parts of the organization affected by the process?
-- Will the idea being discussed be a remedy for that root cause? Will the cause go away, or at least become a lot smaller?
-- Is this remedy the optimal one for both the statistical agency as a whole and the customers as a whole?
-- What resistance would there be to this change and how would you deal with it?
-- How will you know if the remedy was effective and how would you continue to monitor it?
- Is the proposal consistent with the concept of self-control?

Has upper management established an infrastructure and support to enable us to answer these questions in a way that is beneficial both for the agency and for its customers?

- Are the upper managers providing the resources and leadership for quality that will allow us to make the needed changes?

If you start asking yourself these types of questions, you are starting to think about the management dimensions of quality. If you get good answers to most of them, then your agency is on its way to achieving total quality.

## SESSION 1

The Challenge of Reducing Resources and Improving Data Quality

# THE APPLICATION OF THE TOTAL SURVEY DESIGN APPROACH TO DETERMINING STRATEGIES FOR RESOURCE ALLOCATION IN THE SURVEY OF MOTOR VEHICLE USAGE 

S. Linacre ${ }^{1}$


#### Abstract

Statistical collections are subject to a variety of sources of error and strategies should be adopted and resources allocated to minimise the total survey error resulting for a fixed cost. This paper examines a survey subject to substantial 'respondent error' and outlines alternatives collection strategies and their merits in terms of the resulting total survey error. The paper puts the case for Total Survey Design in the allocation of resources for such a collection.


KEY WORDS: Total Survey Design; Reporting error; Diary approach.

## 1. INTRODUCTION

Statistical collections are subject to errors from various sources. The task of the survey designer is, for a fixed cost and set of output requirements from the survey, to minimise the total error on these outputs. Alternatively the designer might see the task as being to design a survey of minimal cost to provide required outputs with a specified maximum total error. In either case the concept of total error would include not only sampling error, but all aspects of error causing the estimate to deviate from the concept the user is interested in.

Aspects the designer will need to consider to achieve this aim include: the purpose of the survey, how the data will be used, what types of error will affect this type of use, what the sources of these errors are, and, on an operational basis how they might be reduced, and which methods of error reduction might be most cost effective. If the designer is able to take these issues into account a "total survey design" will result. Consideration of such errors as coverage error, or sampling error in isolation will generally result in inefficient use of resources for overall design.

As in any process for which quality management is attempted, there is a cycle the survey designer will follow in evaluating and improving an ongoing survey. Firstly, the process needs to be looked at as a whole, and information from discussions, debriefings, pilot tests and evaluation studies used to identify the major sources of error. Work is then required to obtain quantitative information on these error sources and develop methods to reduce them. Finally, further evaluation studies are needed to obtain quantitative information on the effectiveness of these methods, and the cycle starts again.

The aim of this paper is to highlight the importance of an approach that looks at a collection as a whole, rather than as series of discrete subprocesses, by reference to the design of a specific survey.

## 2. SURVEY OF MOTOR VEHICLE USAGE - BROAD DESCRIPTION

The Australian Bureau of Statistics (ABS) Survey of Motor Vehicle Usage (SMVU) provides information on the characteristics and use of private and commercial vehicles in Australia. The data collected includes:

- type and characteristics of vehicle

[^3]- distance travelled
- purpose of vehicle use
- area of operation of vehicle
- commodities carried
- estimated tonne-kilometres carried
- fuel use
- driver characteristics
- passengers carried (for buses).

It has been conducted triennially since 1976 with the most recent survey having been run in 1988. The main uses of the data include:

- providing the basis to the allocation of commonwealth funds to State Governments for road funding;
- input to the formulation of cost recovery arrangements within States, particularly in relation to setting registration fees and levies;
- input to calculation of accident exposure rates;
- estimates of the use of road transport facilities by industry;
- provision of benchmark data for road policy studies, particularly relating to heavy vehicle truck use and fuel usage patterns;
- air/noise pollution monitoring; and
- monitoring transportation of dangerous materials.

The scope of the survey is all registered vehicles other than caravans, trailers, and tractors and the frame is created by capturing the various State and Territory registry files which are then stratified by vehicle type by size and by year of manufacture. The survey is designed as a single stage, single phase stratified sample, the sampling unit being the individual vehicle. Type of vehicle is also collected as part of the survey and estimates are based on reported vehicle type. Sample allocation is treated as a multi variate problem with accuracy levels being specified at the State $x$ vehicle type level, and expected levels of mistratification are taken into account in the allocation. The data is collected from the owners of selected vehicles (about 68,000 in all) by means of a 12 -month recall mail questionnaire. The reference period of the survey is from October of the year before dispatch to September of the year of dispatch.

From a quality viewpoint, the main areas of concern for the survey are: the difficulty respondents have in providing the recall data; the inefficient stratification using the poor quality auxiliary information on the register frames; and processing problems caused to a large extent by the difficulty the respondents have in providing the required information.

## 3. SOURCES OF ERROR IN SURVEYS

If we look at the major sources of error to which a collection is subject, in the case of the SMVU we can see the following.

Conceptual Error: The data items for the SMVU are relatively straightforward and conceptual problems do not arise. The data is used by State transport authorities as a benchmark measure of road usage and the main items of interest are distance travelled and loads carried. The meaning of these items to the transport authorities are clear and they are also easily reasonably well understood by the respondents. Thus the required items for the SMVU are relatively straightforward and the survey does not suffer from problems of definition, concept etc common to many household surveys.

Definition of Population: The population for the SMVU is also relatively well defined and an appropriate list is provided in the vehicle registration files. However, as used in the SMVU, these files provide a single annual snapshot to cover activity over a full year. Thus there is a problem of vehicles being deregistered during the year or changing ownership during the year. Comparatively however undercoverage and overcoverage are not substantial issues. The main problem arises with the quality of auxiliary information on the files and used in the sample design.

Type of vehicle details given on the file are frequently incorrect and result in mistratification of vehicles. As estimates are required by vehicle type, these are based on reported vehicle type and are subject to significantly higher standard errors than if frame information were more accurate. It is roughly estimated that with high quality 'type of vehicle' information, sample sizes could be reduced by about $20-30 \%$ to achieve the current level of standard errors on the estimates.

Sampling Error: As is always the case, the sampling error on the estimates for SMVU could be improved through an increase in the sample size. Another way of making an impact on the sampling error would be through improving the auxiliary information on the registration files to provide more efficient stratification as discussed above.

Reporting Error: A major area of concern for this survey is the reporting error associated with a number of items in the SMVU. The survey requires a 12 month recall for many questions such as distance travelled, area in which travel occurred, rate of fuel consumption, passengers carried and goods carried. Pilot testing indicates this information is not generally available to respondents in records kept, and some form of estimate is required.

Processing Error: As the SMVU is run on a triennial basis there is a need to train new staff for processing the collection each cycle of the collection. For the 1988 Survey 15 of the 57 officers at the processing centre were temporary staff, the remaining staff had been drawn from other areas of the ABS or elsewhere in the public sector. Very few had any experience in SMVU concepts prior to the processing and a 1 week training course was provided at the beginning of processing. Processing results in large amount of change being made to the data reported by respondents and substantial imputation to fill in the spaces left by respondents unable to provide appropriate figures. The scope for processing error is therefore large. Another problem with the 1988 SMVU, run in Perth, was the communication problem caused by the remote location.

Nonresponse Error: The survey has a $15 \%$ nonresponse rate, although imputation for item non response is higher. At final estimation full nonrespondents are effectively imputed as the stratum mean where strata are defined by vehicle type, age of vehicle and state. Analysis of nonresponse bias has not been undertaken to date.

From the above discussion the main sources of error for the Survey of Motor Vehicle Usage can be seen to be: lack of accuracy of the auxiliary information used in sample design, reporting error problems associated with using a 12 month recall methodology and ensuing processing error complications and further processing problems caused by inexperienced staff and the use of a remote location. The effects of the frame problems are relatively straightforward to measure. To better understand the relative importance and implications of reporting error problems a number of studies were undertaken. These are described briefly in the following section.

## 4. STUDIES UNDERTAKEN TO MEASURE THE EXTENT OF REPORTING ERROR PROBLEMS

### 4.1 Pilot Testing

The SMVU forms were extensively pilot tested prior to the 1988 Survey. As part of this testing respondents were asked to complete a form, then a follow up interview was conducted. This covered the process used by the respondent to complete the form. In addition, respondents were asked to categorise their response as very accurate, fairly accurate, not accurate or no idea how accurate, for a number of items. Table $\mathbf{1}$ below is indicative of the findings from this testing.

From this testing it is clear that there are some problems for respondents in providing 12 month recall data; with the extent of the problem depending on an interpretation of "fairly accurate". In particular it can be seen that only a small percentage of respondents had access to records.

### 4.2 Analysis of Processing of SMVU

Substantial management information was obtained during the processing of the 1988 SMVU. The aim in collecting and analysing this information was twofold. Firstly, there was a need to obtain information on the efficiency of the processing of the survey which was being run for the first time as a centralised operation, in a single processing centre, incorporating both data entry and clerical staff in a multiskilled fashion. Secondly, particularly given the known problems with data reporting, an estimate of the effect of these problems on partial response, and edit and imputation rates was required. This could be fed back to determine appropriate changes to the methodology and form design of the survey.

Table 1 : Percentage of Respondents to the Pilot Test Giving a Very Accurate or Fairly Accurate Rating to their Responses; Percentage of Respondents using records to provide responses

|  | Distance <br> Travelled | Purpose of Travel ${ }^{1}$ | Proportion of Distance Travelled within areas ${ }^{2}$ |
| :---: | :---: | :---: | :---: |
| Private Vehicles |  |  |  |
| Very accurate | 42\% | 36\% | 17\% |
| Fairly Accurate | 36\% | 48\% | 70\% |
| Used records | 30\% | 12\% | 20\% |
| Commercial Vehicles |  |  |  |
| Very accurate | 36\% | 62\% | 41\% |
| Fairly Accurate | 57\% | 31\% | 47\% |
| Used records | 32\% | 30\% | 15\% |
| Buses |  |  |  |
| Very accurate | 62\% | 69\% | 58\% |
| Fairly Accurate | 35\% | 18\% | 37\% |
| Used records | 42\% | 21\% | 38\% |

1 Purpose of travel was categorised by business, professional, firm or government; to and from work; and personal and other use.
2 Areas were categorised by: Capital city and environs; other within State; and each other State and Territory individually.

Some of the management information collected during processing was very interesting. During processing of the 67,000 forms, some 45,000 telephone contacts were made by respondents to the processing centre with a further 50,000 calls being generated from the processing centre. This is despite the survey being a mail based collection with mail reminders. The high level of respondent contact for this survey was felt to be due to a large extent to problems with data availability, causing both respondent initiated and processing centre initiated queries. The final overall response rate for the survey was $85.0 \%$.

To analyse the extent of editing, the degree of recontact with respondents and the effects of edits on the final data, a subsample of 551 vehicles was selected. For each selected vehicle the original reported value and any subsequent change to that value, and the basis for the change (contact with respondent, access to vehicle manuals etc) was recorded. In summary from this sample it was estimated that some editing change was made for $91 \%$ of the forms, with an average of 6.1 changes per form edited or 5.6 changes per form returned. The changes were based on contact with the respondents in $38 \%$ of the cases, on information in reference manuals in $10 \%$ of the cases and on other information (usually on the form) in $52 \%$ of the cases. Of the changes made $66 \%$
were the result of the initial response being blank, the remaining $34 \%$ being changes to a reported value. While an amendment to a initial blank response will inevitably result in an increase in the estimate for this survey, an interesting aspect of amendments to non-blank items was that the weighted aggregate effect for almost all items were negative.

If it is assumed that initial non-response to an item is indicative of insufficient information being available to the respondent to reply, then there is particular concern for a number of items. For items such as distance travelled broken down by purpose and area of travel, between $15-30 \%$ of the estimate was the result of edited or imputed values. On the other hand the percentage of the variable total distance travelled contributed by edited returns (generally items initially left blank) was relatively low at $3 \%$.

## 43 Odometer Survey (OS)

While both the pilot testing and analysis of edits during processing indicate a likely high level of response error in the data, it is not clear whether this will show itself in an increased variability in the estimates, or whether a bias would result, and what the magnitude of the effect would be.

## Table 2

|  | Cars | Rigid Trucks | Articulated Trucks | Total |
| :---: | :---: | :---: | :---: | :---: |
| $\mathrm{D}=\frac{\mathrm{SMVU} \mathrm{U}^{1}-\mathrm{OS}^{1}}{\mathrm{OS}}$ | 14.0\% | 10.3\% | 5.8\% | 13.15\% |
| 95\% CI for D | (8.9\%, 19.1\%) | (3.9\%, 16.6\%) | (1.3\%, 10.4\%) | (9.0\%, 18.0\%) |
| CV for SMVU | . 96 | 1.63 | . 82 |  |
| CV for OS | . 77 | 1.06 | . 92 |  |
| Fraction of SMVU sample required by $O S$ to give same RSE | . 64 | . 42 | 1.26 | . 79 |
| Sample for 1988 SMVU | 12896 | 20469 | 20014 | 53379 |
| RSEs for SMVU | 1\% | 2\% | 1\% | 1\% |

1 SMVU denotes distance travelled from the SMVU. OS denotes the corresponding estimate from the odometer survey.

The odometer survey was designed to provide a quantitative measure of the reporting error effect for one important variable, total distance travelled. Pilot testing and the analysis of processed forms indicates components of distance travelled are likely to be subject to far greater reporting error than distance travelled but the advantage of an odometer survey is that is provides a measure of accuracy on total distance in a relatively straightforward way.

The methodology of the odometer survey test involved selecting a sample of vehicles in one State, New South Wales in parallel to the sample selected for the SMVU. The parallel sample, rather than being asked recall for the full period, were instead approached at the beginning of a corresponding period and asked for an odometer reading and then reapproached at the end of the 12 month period to provide a second odometer reading. The results from the odometer survey and the 12 months recall survey were compared.

Table 2 clearly indicates two things. Firstly there is a marked and consistent positive difference in distance measured by the SMVU compared to the Odometer survey. This is highest for cars and lowest for articulated trucks. (The lower value in the latter case may possibly be the result of better record keeping for this group, although examination of the results shows proportionately more vehicles in the less than $10,000 \mathrm{~km}$ range for the SMVU compared to the odometer survey which may be associated with operators wishing to avoid strict regulations on distance travelled and travelling time within the industry.) In all cases the size of the upward bias is very large compared to the standard errors the survey is designed to achieve.

Secondly, there is evidence of a significant response variance in the figures. The variability of the SMVU data is markedly higher than that of the odometer survey data except for articulated trucks. An implication of this is that the sample size required to provide an estimate of distance travelled based on an odometer survey basis would be substantially lower than that for the SMVU to achieve the same relative standard error. The reduction in sample size for each vehicle type is also given in Table 1.

The empirical probability distribution functions of the distances travelled for the odometer survey and SMVU were compared, see figures 1,2 and 3, and confirm that the Odometer survey figures are far smoother and the spread of distances smaller than in the SMVU.

Substantial rounding is evident in the SMVU data which has spikes occurring at about every $5000 \mathrm{~km}: 34 \%$ of the returns for the SMVU were exactly in units of 5000 miles or kilometres compared to $0.44 \%$ for the odometer survey! This rounding might be taken as another indication of the reliability of the data available to respondents.

Figure 1: Cumulative Distribution Function: Car, Distance Travelled


Figure 2: Cumulative Distribution Function: Rigid Trucks, Distance Travelled


Figure 3: Cumulative Distribution Function: Articulated Trucks,
Distance Travelled


A further point of interest is the heavy tail apparent in the distribution of distance for the SMVU for cars and rigid trucks indicating that the upper $50 \%$ of the respondents tend to over estimate in their recall of distance travelled while the lower $50 \%$ of responses tally reasonably well with the odometer survey.

In summary the odometer survey makes clear there is a large positive bias in the SMVU estimates, as well as an increased level of variability. The bias is substantially greater than the sampling error of the SMVU and in fact it dominates the total mean square error for the survey.

It should be noted that in addition to the odometer survey described above, a similar survey was run in conjunction with the 1985 SMVU. The earlier survey was substantially smaller in size and failed to detect a significant bias although a positive discrepancy between the SMVU and odometer survey estimates was measured.

## 5. ALTERNATIVE METHODOLOGIES

While the odometer survey is a very useful evaluation tool for the SMVU it does not in itself provide an alternative methodology as it cannot provide a number of key items required from the survey, for example a breakdown of travel by purpose and area of travel, passengers carried, freight carried etc. One appropriate methodology for these items is a diary approach. An option that might be considered is a one month diary based survey spread evenly over the year.

The effect of this approach on response load and collection increase with respondents costs is uncertain. On the one hand, respondent load would increase with respondents required to make a number of entries in the diary over a month rather than complete a single recall form. Furthermore, a larger sample would be expected to be required for the same relative standard errors if each questionnaire covers only one month rather than a full twelve month period. (The increase in sample would depend on the coefficient of variations of monthly estimates compared to annual estimates, the between month correlation in distance travelled, and current sampling factors.)

On the other hand the odometer survey has shown that collecting figures more closely related to true distances travelled will not only substantially reduce bias but will provide some gains in variance as well. The sample size required to achieve the same relative mean square error as the current SMVU assuming negligible response variance and bias associated with the diary approach, would therefore be a fraction of that required for the SMVU. Furthermore a diary approach is likely to benefit other poorly answered parts of the form to an even greater extent than the item total distance travelled. In addition, respondents may find it less of a demand to provide readily available data on a diary basis than data they do not have on a 12 month recall basis. These issues will need to be investigated through discussions with users and through pilot testing of the approach.

Other costs and savings for the diary approach compared to the recall approach are set out briefly below:

1. Development of Survey: The twelve month recall based SMVU is already developed, with an associated computer system and processing category. To move to a diary based methodology would involve developing both the methods and the associated systems. This is a major impediment to change
2. Frame creation and sample selection: Under a design with sample spread over time, the frame might be created and the sample selected at the beginning of the year, or periodically eg quarterly, during the year. Compared to the current SMVU the former approach would include that portion of vehicles removed from the register during the year but would miss new registrations during the year. The second option, would be more expensive as it requires 3 additional sets of runs for each of the 8 state and territory files, plus matching of selections to avoid duplication. However it would provide very good coverage of vehicles throughout the year, picking up those vehicles deregistered during the year, which are currently omitted from the SMVU.
3. Despatch and processing: Under the diary approach, despatch and processing would be spread over the year. This has both advantages and disadvantages. A smaller team of permanent staff could handle the more spread workload, reducing the recruitment and training load. For the 1988 SMVU, processing
was spread over 5 months and there were 57 staff required. If processing is spread over 14 months the number of staff involved would be at least halved, reducing recruitment and training costs. It might also be expected that with more experienced staff, higher quality processing would result.

Furthermore, given the reduced difficulty of data availability to the respondent using a one month diary approach, the level of follow up activity might be expected to substantially decrease. On the other hand, given the additional load associated with the diary, further follow up may be needed to achieve acceptable response rates. Without further information from pilot testing the diary approach, it is impossible to predict the extent to which these factors would offset each other.

A reduction in sample size to achieve the same mean square error as for the 1988 survey would provide substantial further cost savings in despatch and processing!
4. Timeliness of data to users: As a diary based survey would be spread over the reference year rather than being conducted following the reference year, an improvement in the timeliness of the resulting data would be expected.
5. Measure of the quality of the data: Assuming a substantial reduction in bias under a diary approach to the survey, the reliability of the estimates would become relatively visible through the RSEs. With a reduced sample, at the small cell level some of these RSEs will be quite high. This may result in some adverse user reaction compared to the current high but relatively hidden level of MSE (the results of the odometer survey were published in a technical note, but RSEs were not adjusted to MSEs).

## 6. CONCLUSION

For the SMVU, the future direction is to make use of a multidisciplinary team approach to pull together the various aspects of the survey design and develop a method that provides the most effective overall strategies for the users. Given the relatively small contribution of sampling error to total error, the sample sizes will be reduced and methodological work will be redirected from issues associated with sampling error to those associated with reducing response and non-response errors. Finally there will be a need to educate some users in the total survey error concept. To avoid fixation on sampling errors we will need to develop ways to provide more regularly, and for more estimates, a corresponding measure of total error.

# AN EXPERIENCE WITH A PRODUCTIVITY INCENTIVE PLAN AT STATISTICS CANADA 

J.-F. Gosselin ${ }^{\prime}$


#### Abstract

A productivity incentive plan involving gainsharing with the participating employees was introduced on an experimental basis in three small units responsible for operational aspects of statistical programs. This paper shows that significant productivily gains have been realized while maintaining or improving quality and timeliness. Reactions of employees have been very positive and the indications are that the introduction of the plan has resulted in less on-the-job stress when compared to control groups.


KEY WORDS: Incentives; Productivity; Quality.

## 1. INTRODUCTION AND BACKGROUND

A major reorganization was initiated at Statistics Canada in the mid-80's to centralize the statistical operations that were located in Program Divisions. The main objective was to achieve efficiencies through a better and more rational use of operational resources. It was realized from the beginning that to maximize the gains of such a massive reorganization we had not only to make a more rational use of resources, but also to make a long term commitment to rethinking the way operations are designed and managed.

Several initiatives were put forward to automate various processes, to successfully introduce new methodologies such as Computer Assisted Telephone Interviewing (CATI) and automated coding, etc. What will be discussed here is an approach that we have successfully tried and tested which is radically different from the way we are used to managing operations in the context of the Public Service.

From the very beginning, because of the need to achieve higher levels of performance, a decision was taken to develop and put in place feedback mechanisms to allow the monitoring of productivity and quality. Accordingly, steps were taken to begin the development of engineered work standards as well as the development of statistical quality control for major applications. Although gains in productivity were observed as a result of the introduction of those measures, it soon became evident that significant and lasting improvements would not be possible without full employee support and the use of incentives. This is when we got the idea of introducing a gainsharing plan.

The principle idea underlying gainsharing plans is to share with employees the gains achieved, when productivity rise above certain benchmark levels. Such plans are very common in private industry and the benefits usually take the form of bonus payments. The benefits we chose to offer was in the form of time off. What is unique in this experiment is that it was applied in the context of the Public Service.

This approach was first introduced on a very small scale in the Building Permit Survey. Time off was very appealing to those employees and they managed to raise their productivity levels from about $60 \%$, which is not an unusual level in areas operating without work standards, to levels well exceeding $100 \%$.

[^4]Based on this very small pilot study which involved only five employees, a decision was taken to proceed with a full scale test in two other units: the International Travel (ITS), and the Survey of Employment, Payrolls and Hours (SEPH) Unit. These Units contained respectively an additional 15 and 25 employees, giving a total of about 45 employees in the test.

Table 1 illustrates the gains in productivity that were achieved.
Table 1: Productivity Incentive Plan - Summary Results

|  | PRODUCTION UNITS |  |  |
| :---: | :---: | :---: | :---: |
|  | Building ${ }^{1}$ Permits | International Travel | SEPH <br> (Survey of Employment Payroll, Hours) |
| Employees | 3-4 | 15 | 25 |
| Productivity <br> (Initial) | 58\% | 69\% | 75\% |
| Productivity $(24 / 11 / 89)$ | 108\% | 118\% | 95\% |
| Bonus Earned Total Days | 62 | 357 | 101 |
| Days/Employee | 18 | 24 | 4 |

1 The Building Permits Plan was terminated at the end of February 89 due to regionalization of the survey.

Depending on the Unit, productivity rose from $60 \%-75 \%$, to the $95 \%-118 \%$ range, a very significant increase, before the Plan was stopped. Depending on the Unit, the number of days off earned by employees varied between 4 and 24 over the period of the test (about 16 months), with an average of about 12 days per employees. As can be seen, time off turned out to be a very strong motivator.

Now, you may ask why if the experiment was so successful, was the Plan stopped? The reason is quite simple: the Federal Court ruled that the Plan contravened the Public Service Staff Relations Act. It is not the purpose of this paper to enter into the legalities of this issue. Rather, the main purpose is to address the following issues:

- What was the impact on productivity?
- Did the increases in productivity have a negative effect on quality and timeliness?
- Did the Plan result in increased stress?

Before these are addressed, the main characteristics of the Plan will be described.

## 2. CHARACTERISTICS OF THE PLAN

One of the fundamental features of the Productivity Incentive program implemented at Statistics Canada is that is a group plan. This is a fundamental difference compared with piece rate scheme used in industry where an individual's pay depends on individual productivity.

In this Plan, time off earned depends on the entire Unit's productivity so that the focus is on team work, i.e. how can we achieve the objectives including quality and timeliness, while making the best possible use of each individual's talents? Internal cohesion and communication become critical factors. This requires a shift in the work culture from doing what one is told, to taking some responsibility for improving the way in which the results are achieved. In this particular example, this was achieved by conducting team building exercises.

Another key feature of the Plan is that it is based on voluntary participation. Motivation and therefore employee support is at the heart of the approach which should exclude any form of coercion. This goes hand in hand with guaranteed employment. No one is interested in doing themselves out of a job, so needless to say this is an essential ingredient. Bonus in the form of time off is earned when the Unit's productivity begins to exceed $80 \%$ of the standard. Originally, this was first set at $100 \%$, however, this would have set a discouraging target for employees. It was decided to lower the threshold and award an increasingly larger proportion of the savings, up to $50 \%$ for a level of productivity exceeding $100 \%$.

To maintain a high spirit, and create a culture of continual improvement through employee participation, the Plan called for no changes in work standards, except in situations where the content of the work had changed for example as a result of a survey re-design, or a major capital investment had been made on the part of management. Finally, in order to promote free flowing communication and to allow employee suggestions to be evaluated, a two-level committee structure was set up with the full involvement of employees.

## 3. EVALUATION

### 3.1 Productivity Analysis

Work standards form the cornerstone of the Productivity Incentive Program. They introduce an important element of objectivity in setting production targets as well as providing a basis for monitoring and providing weekly feedback to employees on the current levels of performance.

Productivity levels for the three Units under study were analyzed for the year preceding, and for the 16 months following the formal introduction of the Plan on July 1, 1988. The following summarizes the major observations.
a) For the Building Permits (B.P.) Unit, prior to the introduction of any form of incentive, productivity was averaging $58 \%$. Within months, the productivity of the Unit reached levels in the $115-130 \%$ range. During the Pilot period, productivity was consistently maintained above $125 \%$ except for the period just before the termination of the Plan as a result of the then pending regionalization.
b) During the Pilot period, the ITS Unit (Table 2) has consistently maintained productivity in the bonus range. Most months have been above $100 \%$ with a high approaching $130 \%$ and a low of about $80 \%$. The latter is due to low volume months. Had this been anticipated by management, new measured work could have been brought into the Unit and most likely productivity could have been maintained. This indicates the importance of careful planning and scheduling in this type of plan.

For the period prior to this, we had seen a gradual upward trend in productivity up to the $90 \%$ level. This resulted from a combined effort of improvement in the organization of the work, the introduction of standards, and formal team building sessions. However, the ITS staff were well aware that discussions were taking place on the possibility of installing a plan in that Unit and were working on the expectation that this would happen. This is especially true of the February-June ' 88 period where productivity increased from the $90 \%$ to $100 \%+$ level. It is our opinion that it would have been very difficult to exceed and maintain levels well above $80 \%$ without the anticipation of a tangible motivator such as time off.

Table 2. International Travel

c) For SEPH (Table 3), we have seen a gradual and steady improvement in productivity from the low of $60 \%$ up to about $90 \%$ at the end of the Pilot period. Again, these are significant increases although not as high as for ITS. This may be due to the fact that this was the last Unit in which standards were introduced. There had been an obvious upward trend since the introduction of the Plan and this might have continued had the Plan been maintained since employees had just started to enjoy the fruits of their efforts.

Table 3. S.E.P.H.

d) The following represents the bonus earned by employees in the Pilot Units.

|  | Bonus Earned |  |  |
| :--- | :---: | :---: | :---: |
| B.P. | L.T.S. | SEPH |  |
| Total Days | 62 | 357 | 101 |
| Days/Employee | 18 | 24 | 4 |
| Period | July 88 <br> Feb 89 | July 88 | Nov 90 |
|  | Nov 98 |  |  |
|  |  |  |  |

e) Overall, significant productivity gains have been achieved in all Units. Had productivity rates continued at their original levels, we would have had to increase the size of the group by about 9 PY 's to accomplish the work which has been done by plan members since the plan was introduced.

### 3.2 Quality and Timeliness

One of the questions which may be raised is whether the increases in productivity has had any impact on the Unit's ability to maintain its quality and target dates.

The evidence suggests that both quality and timeliness have in fact been maintained and in some cases improved:
a) The number of SEPH units left unprocessed at cut-off (Table 4) is well below the corresponding numbers for the previous year. In fact, the incentive initiative combined with changes in Regionalization procedures, has virtually reduced unprocessed records to zero at survey cut-off since January '89.
b) For the ITS program, there are no significant changes in the pattern and behaviour of quality levels of the various outputs for the test period and the previous year. This holds true whether one looks at coding (Table 5) or keying activities (Table 6) for which the Average Outgoing Quality (AOO) was maintained well below the Average Outgoing Quality Limit (AOOL).
c) For SEPH, the test period corresponded roughly to the second year of implementation of this new QC application. It is very interesting to notice (Table 7) that the quality has improved steadily in spite of increases in productivity. We believe this is due to an effective feedback process whereby regular meetings are held to review with the staff, the nature of the errors and how quality can be improved. We would expect the quality level to stabilize once the application has matured.

Table 4. S.E.P.H. Timeliness


- OUERIES LETT

Table 5. ITS - Quality (Coding)


Table 6. ITS - Quality (Keying)


- A00 - AOOL

Table 7. SEPH - Quality (Editing)


## 33 Employee Reaction

Employee reaction to participation in the plan has been assessed in several ways:
a) In-House Questionnaire, administered early in June by the Organizational Development Section of Personnel Division of Statistics Canada showed that $94 \%$ of employees in the plan wanted it to be continued. This survey also showed a strong degree of job satisfaction - $48 \%$ of those who responded indicated a higher degree of satisfaction compared to the previous year. Only $9 \%$ ( 3 respondents) indicated a decreased level of job satisfaction over the previous year. The survey brought forth numerous comments and suggestions.
b) Prof. William Jones of Carleton University was commissioned to examine stress levels found in the incentive groups as compared with other Statistics Canada employees in similar categories and levels. His conclusions, were summarized as follows:

> "Total Work Stress is lower, not higher, among Incentive Plan workers. ... Incentive plan workers experienced significantly less role conflict and significantly higher feelings of job satisfaction and commitment to the organization. There was absolutely no evidence that the operation of the plan had resulted in any increase in interpersonal or social stresses or strains."

### 3.4 Success Factors

The following are what are perceived to be the major success factors in this experiment:
a) Everything starts with an idea. And so often, good ideas are turned down too quickly before they have had a chance to be explored or refined.
b) Management had to be prepared to take some risk. Radical changes can rarely take place without this key ingredient.
c) Employees had an opportunity to participate throughout the process.
d) There were tangible benefits both for employees and management and an incentive to work together to make it work.
e) There was a high degree of openness, support, and flexibility on the part of employees and all levels of management. Perhaps the fact that this was being tried on a small scale in a controlled environment helped in this respect.
f) Job security was a key factor in making this work. Quite simply, nobody is willing to do themselves out of a job.

This can be summarized in three essential characteristics: CONSULTATION; CONCERTATION; SHARING at all levels, i.e. benefits, work, responsibility, decisions, improvements.

## 4. CONCLUSION

This experience with a productivity Incentive Plan at Statistics Canada has demonstrated that such an approach has the potential to be used very successfully in an operational environment to improve productivity while maintaining or improving quality and timeliness. Studies have also indicated that important benefits include significant improvements in job satisfaction, organizational commitment and a reduction in levels of stress.

There is no doubt that the introduction of the Incentive Programs of this type represents a very profound change in the way the Canadian Public Service operates if used on a large scale. On the other hand, it would only apply to areas where the work can be measured and quality monitored.

It will likely entail rethinking some of the aspects that we now take for granted. For example, how does one classify jobs and appraise employees in situations where the work is shared among an entire Unit?

A recommendation was formulated as part of the Public Service 2000 initiative to amend the current legislation to allow for such plans to be implemented. Hopefully, this will create opportunities from which employees and management can benefit.

Proceedings of Statistics Canada Symposium 90
Measurement and Improvement of Data Quality
October 1990

# A SYSTEMATIC APPROACH TO QUANTIFYING THE QUALITY OF REPEATED SURVEYS* 

R.D. Tortora ${ }^{\text {a }}$


#### Abstract

This paper outlines a systematic approach to quantifying the quality of repeated surveys. The approach is based on developing a quality matrix of measurements of the survey process. The elements of the quality matrix are defined by the intersection of the components of quality, the sources of exror in a survey, and the stages of a survey. These measurements, when analyzed over time, can be used to develop an improved survey.


KEY WORDS: Repeated surveys; Quality, Quality improvement.

## 1. INTRODUCTION

This paper develops a systematic method to quantifying the quality of a repeated survey. Once this is done, the survey manager has a management tool to prioritize continuing survey improvement. This management tool focuses quality improvement on the survey process rather than the minimization of relative error per unit cost.

The quality of survey statistics has received considerable attention in the literature. Statisticians have tended to think of quality in terms of precision or accuracy rather than using a broader definition. There are many papers devoted to improving the part of survey quality called precision. Others are broader, addressing the accuracy component of quality. A fundamental paper in this area is Hansen, Hurwitz and Bershad (1961). Platek and Singh (1981) relate survey performance to survey controls by trying to identify the major sources or error and discussing the control activities to assure a certain level of performance. Brooks and Bailar (1978) outline the error structure for a particular estimate and survey, namely employment as measured by the Current Population Survey. Spencer and Mulry (1990) develop a total error model for the Post Enumeration Survey. Others related more than one component of quality. Groves (1989) relates errors and costs. Jabine (1990) develops a quality profile of the Survey of Income and Program Participation. Statistical agencies have developed quality guidelines (Statistics Canada (1987)) and survey performance measures (Energy Information Administration (1989)) to benchmark their products. These two documents are very broad and cover the full scope of survey activities, however again, with special attention to accuracy issues. None of these papers or products pursue a total quality approach to prominent in business and industry today to improving the quality of a survey. This paper attempts to take some of the ideas of Juran (1964) and Deming (1986) and apply them to a repeated survey to foster continuous improvement.

## 2. BACKGROUND

There are several important reasons for developing a systematic method of quantifying the quality of the survey process. First of all, a systematic method allows the survey manager to assess the overall quality of the survey and to identify gaps in the quality measurement of the survey. Second, besides providing a measurement of the quality of the survey at a particular time, the survey manager can use the data to quantify the quality of the

[^5]survey over time. Finally, the systematic approach allows the manager to focus on the entire survey process from initial design and sampling to publication and revision of estimates. Thus the manager is better prepared to address questions of survey performance over time, resource use or allocation, and anticipation of future problems, including justification of future resource needs.

The motivation for this paper comes from the following five assumptions. First, that funds for improving the quality of existing surveys is not going to be readily forthcoming. In fact, quite the opposite may be true and the resource base may decline. Thus the survey manager is faced with the possibility that the quality of the survey will decline unless methods are instituted that allow the manager to better use (a declining) resource base. Second, the survey manager has a lot of data available that helps quantify the quality of the survey. This data, like sampling error, nonresponse rate, cost components, etc., if captured over time can tell the manager a great deal about the performance of the survey by providing indicators of the quality of the survey. Third, it is assumed that quality is much more than accuracy to the survey manager. It also includes timeliness, relevance and resources. Timeliness is more than the timeliness of the published results but also includes the timeliness of the different steps or stages in the survey process. Of course, quality also includes a resource component. Its important is clear. The final component of quality is relevance. Here the emphasis is placed not only on the data users and the relevance of the survey to the users but on the relevance of the products or results along each stage of the survey. For example, the survey manager is interested in the relevance of the questionnaire design not only for the data the questionnaire produces for the data users, but also for the use of the questionnaire by interviewers and respondents. The manager sees both interviewers and respondents as customers in the process of executing the survey. Fourth, this paper uses one of the fundamental ideas of the leaders of quality improvement, namely that a basic characteristic of a system in control is that there is a lot of data available to describe how the system is performing. Then quality improvement can be based on sound analysis of existing data rather than on special studies collecting basic data on system performance which drains scarce resources. Finally, the assumption is made that while direct measure of quality are best, for example, total nonresponse bias, indicators of quality, when examined over time, nonresponse rates, characteristics of nonrespondents, etc. may prove to be just as useful. Note that these assumptions point to the need for a large amount of data analysis. Any system, to be useful, should present data in ways that encourage easy, useful analysis and comparisons mostly time series charts and other graphical representations.

Before discussing the model that describes the systematic approach to quality for repeated surveys, it is worthwhile to consider answers to the question of how will the survey manager know, or be able to say, that a quality improvement has been made to the survey? If additional money has been made available for the survey and that money is used to increase timeliness, increase accuracy, or make the data more relevant, the manager can say the quality of the survey has been improved. However, the more challenging case arises when there has not been an increase in resources for the survey. When has a quality improvement been made? It the survey manager can improve at least one of the components of quality while keeping the others constant then the quality of the survey has been improved.

To help the manager make these quality improvements the notion of pareto analysis seems important. Juran (1964) uses this concept to help an organization focus on its important problems. In manufacturing, if 3 of a company's twenty-five products produce seventy per cent of the revenues then pareto analysis suggests the company concentrate its efforts on quality improvement on the 3 vital few products. For a survey, it may not be clear that anyone step or product in the operation is more important than another. In fact, it seems that it is more the case that when the entire survey operation is done well, that is, with acceptable quality, then the survey has acceptable quality. But the notion of pareto analysis can be reformulated to be useful to the survey manager. Figure 1 illustrates this kind of analysis for evaluating two of the quality components, costs and accuracy. Figure 1a. estimates the cost for each of 3 (say) error components. The components are arranged arbitrarily in decreasing order of cost. Figure 1b. estimates the contribution to total error from each of these components, where the 3 components are arranged in the same order as in Figure 1a.

Figure 1. Pareto Analysis: Cost versus Error


Cost for 3 components of error.


1b. Contribution to total error from same 3 components.

Since the error associated with these 3 components does not follow the same decreasing pattern, the survey manager has the opportunity to start a quality improvement for project the survey. It may be possible to shift for costs component 2 to component 3 to decrease its contribution to total error.

## 3. A SYSTEMATIC APPROACH

This section outlines a systematic approach to quantifying the quality of repeated surveys. The approach focuses on the intersection of the components of quality, the error structure of a survey and the stages of a survey and producing a quality matrix of data each time the survey is conducted. Every element of the quality matrix represents a data point that helps quantify the quality of the survey and every element is formed by considering the intersection of the components of quality, the error structure and the stages of the survey. Some elements will be readily available from the survey, for example, nonresponse rate, others may be more difficult to obtain and require special studies. These latter data points will not be available each time the survey is conducted but as the matrix is augmented over time, special studies can be planned to occur on a regular basis or to fill in data needs about survey quality. Another important piece of data that can often be associated with each cell is a tolerance limit or bound that has been predetermined and helps decide if the process is in control. An example of a tolerance limit is a predetermined relative sampling error for a specific survey estimate. There is no unique quality matrix, differing by survey and survey manager.

The quality components are relevance, accuracy, timeliness and resources. The total quality perspective of this paper emphasizes 1) measuring the other components of quality and 2) placing importance on internal customers. So while relevance may be very difficult to quantify from a data user perspective (especially for government surveys) relevance measures from an internal customer prospective, for example, are the interviewer training manuals easy to use?; do respondents understand the meaning of questions?; are they easier to identify and meaningfully quantify. Timeliness measures of the end product of a survey, the published data, are relatively easy to obtain. But by remembering the internal customer at each step of a survey, measures of timeliness that directly impact the quality of a survey can be produced. An example of the timeliness and the internal customer is the arrival of interviewer training materials in field offices soon enough for trainers to adequately prepare for training sessions. When obtaining measures of accuracy the emphasis will be on obtaining measures that serve as proxies for total survey error or mean square error. This happens for two reasons. First, of course, measuring either quantity can be extremely difficult. Second, the survey system has already available many proxy measures of accuracy that can be used and analyzed over time to provide indications of the accuracy of the survey. Finally, any attempt to improve the quality of a survey must be undertaken with full knowledge of resource implications of the improvement. Often, when approached about a quality improvement, the first question from the survey manager is "how much is it going to cost?" To answer this question completely it is essential to know how much the survey costs now, not only overall, but cost by source of error and stage of survey.

Consideration of the error structure of a survey allows for a more complete quantification of the accuracy of a survey and gives guidance to the measures that need quantified. In this paper the errors or a survey are divided into errors of observation and errors of nonobservation. Errors of observation include measures that occur with
the interaction of the questionnaire, interviewer, and respondent, editing errors, coding errors, etc. Errors of nonobservation include measures of nonresponse, sampling error, coverage, imputation amounts and rates, etc. Some of the measures of quality may be attributable to both sources of error and it may be useful to attribute the components to each source. For example, revisions may be made after the original survey estimates are published. The revisions could be caused by errors of observation, filled in the wrong value, or nonobservation, questionnaires received too late to be used for the original estimate and the imputed values sufficiently different from correct value that a revision is required. This last example leads directly to considering the stages of the survey.

Somewhat arbitrarily the process can be broken down into a sample design and selection stage, a data collection stage, an editing and analysis stage, and an estimation and tabulation stage. For each stage the survey manager can identify important measures of survey quality by considering the components of quality and the error structure of the survey. Figure 2 shows how a part of the quality matrix might look for the data collection stage. ${ }^{\text {. }}$

Figure 2. The Quality Matrix for the Data Collection Stage, Time Period if


Some of the cells in Figure 2 contain data that is readily available from the survey system. Items like quartiles of questionnaire completion dates, per cent of editing for each questionnaire item, nonresponse rate and per cent of the estimate accounted for by imputation. Some items may require extra work to obtain or a special study. an estimate of response bias - to obtain data.

[^6]While the quality matrix provides data on the important elements for a particular survey date, the data is most useful if the data for each cell is presented graphically over time. Figure 3 shows the per cent of imputation for a particular survey estimate from an establishment survey.

Figure 3. Per Cent of Estimate from Imputation for an Establishment Survey


In this time series chart two tolerance levels are identified. If the per cent of imputation is more than 25 per cent then the estimate is suppressed in the publication. If the per cent of the estimate accounted for by imputation is between 10 and 25 per cent the data user is notified about the amount through a footnote in the publication. (The decreasing trend in the amount of the estimate accounted for by imputation represents special efforts by the subject matter experts to get cooperation from respondents.)

By developing the complete quality matrix and analyzing the graphical displays associated with the data the survey manager can 1) designate areas where special studies are needed to learn more about overall survey quality, 2) conduct pareto analysis to optimize the use of resources, 3) report on the current quality of the survey, and 4) setup processes that improve (and quantify the improvement) in the survey.

## 4. FUTURE WORK

While this paper sets up a framework for a total quality approach for repeated surveys, clearly much needs to be done to put it into practice. This is the future direction of the work. Data is now being gathered to build a quality matrix for the Current Industrial Reports of the Census Bureau. A variety of issues are being addressed, including selection of the specific surveys which the quality matrices will be developed, and development of the appropriate elements for the quality matrices.

## REFERENCES

Brooks, C.A., and Bailar, B.A. (1978). An Error Profile: Employment as Measured by the Current Population Survey. Statistical Policy Working Paper Number 3. Washington, D.C.

Deming, W.E. (1986). Out of the Crises. MIT, CAES, Cambridge, MA.
Groves, R.M. (1989). Survey Errors and Survey Costs. New York: John Wiley \& Sons.

Hansen, M.H., Hurwitz, W., and Bershad, M. (1961). Measurement Error in Censuses and Surveys. Bulletin of the Intemational Statistical Institute, 38, 2.

Jabine, T. (1990). SIPP Quality Profile, U.S. Bureau of the Census, Washington, D.C.
Juran, J.M. (1964). Managerial Breakthrough. McGraw Hill, New York.
Mulry, M.H., and Spencer, B. (1990). Total Error in Post Enumeration Survey (PES) Estimates of Population: The Dress Rehearsal Census of 1988. Annual Research Conference Proceedings, U.S. Bureau of the Census, Washington, D.C.

Statistics Canada (1987). Quality Guideliness. Minister of Supply and Services, Ottawa
U.S. Energy Information Administration (1989). Standards Manual. Washington, D.C.

## SESSION 2

Data Integration

# INTEGRATION OF ECONOMIC DATA: BENEFITS AND PROBLEMS 

M. Colledge ${ }^{1}$


#### Abstract

" A fundamental question is whether integrated censuses and surveys are a powerful and flexible instrument or whether the "integrated approach" has cemented the Bureau into complexities and refinements that together mean that data can never be collected and processed quickly or cheaply.", R. Cameron, Australian Statistician, 1981.

This paper explores the benefits and problems of developing and using an integrated approach for economic statistics. In brief the benefits accrue from the ability to juxtapose data from different sources and thus enhance their value and test their quality. The problems arise from the possible over-simplification and distortion of economic realities which may occur as a result of attempting to standardize concepts and procedures across data sources, processing systems and client requirements. This theme is illustrated by examples drawn from Statistics Canada and from the Australian Burcau of Statistics.


KEY WORDS: Economic data; Integration.

## 1. INTRODUCTION

Integration of data, and in particular of economic data, is an objective widely promoted by national statistical agencies. For example, the Methods and Standards Committee at Statistics Canada (1990) stated:
"Integration improves the information content and quality of the data and provides a means, quite often the only means of discovering gaps in our statistical series and frameworks".

In a recent review which culminated in the ongoing redevelopment of the Australian economic statistics program G. Sarossy (1987) wrote:

The objective of integration is and will continue to be a prime objective of the Australian Bureau of Statistics. The statistical system required to meet that objective is an integrated system of censuses and surveys which encompasses all industries and economic activities".

In a review of the U.S. statistical system, the Bonnen (1981) report opened its discussion of integration with the following remarks.
"Historically, statistics have come into being through endeavours to satisfy demands in individual subject-matter fields, e.g., agriculture, population, health, labor, manufacturing and trade. The increased importance of the national economic accounts and, more recently, social indicators bave helped develop a more comprehensive approach to statistical programs but, in the main, there is still a strong pull to serve the detailed needs in specific fields as they arise. However, increasingly, statistical series are not used in isolation -- they are used jointly with other statistics in a wide variety

[^7]of cross-cutting analyses. Hence an integrated product is essential, in terms of concepts, measurement, classifications, and analytical methods."

The report went on to state that: "Integration can, among other things, result in improvements in the quality of data, in its relevance for policy purposes, and be an aid in reducing respondent burden".

In the words of Goldberg, when Assistant Dominion Statistician at Statistics Canada, integration should be: "... a guiding philosophy which permeates the actions and policies of a statistical office". Subsequently, as Director of the U.N. statistical office, Goldberg stated that the integration function:
"... should provide a corporate consciousness, maintain checks and balances in the face of differing pressures from the individual sections, foster interdisciplinary project planning and execution and overcome barriers, real or imagined, between the various parts of the organization. It should ensure that common concepts, definitions, classifications and methods are not only available but actually implemented in the various divisions and sections so that statistical series represent elements of an integrated framework and are as consistent and comparable as possible."

As these quotations illustrate, the thrust for integration is by no means new. In many agencies it has a history going back many years. It was, for example the main theme of the Australian 1968-69 Integrated Economic Censuses which constituted: "... a major reorganization designed to increase substantially the usefulness and comparability of the kinds of statistics already being collected and published (ABS, 1970)*. At Statistics Canada, it was considered as "... a basic approach to the organization and storage of statistics" (Gigantes et al, 1970).

In summary, integration can enhance the value and increase the relevance of statistical outputs, improve the accuracy of data collected, and reduce respondent burden and processing costs. Thus it is a very powerful mechanism for improving quality, hence the inclusion of this paper in Symposium 90. However, as its long history illustrates, integration is an elusive goal. Implementation is far from easy. The results can actually be additional costs, loss of timeliness, and forcefitting of data to a conceptual straightjacket. Australian Statistician, R. Cameron alluded to the problems when he asked: "... whether integrated censuses and surveys are a powerful and flexible instrument or whether the integrated approach has cemented the Bureau into complexities and refinements that together mean that data can never be collected and processed quickly or cheaply".

This paper examines the advantages and problems of developing and using an integrated approach for economic statistics. Various types of integration, and potential benefits associated with each, are detailed in Section 2. The generic activities associated with integration are oullined in Section 3. Some actual developments involving integration are described in Section 4, and Section 5 summarizes the lessons learned.

## 2. TYPES AND BENEFITS OF INTEGRATION

### 2.1 Introductory Remarks

One of the problems in discussing "integration" is that the term itself is not precisely defined. Even within the relatively narrow context of collection and production of economic statistics, integration is multi-facetted, having a variety of meanings. The following paragraphs address this issue.

Any statistics program may be viewed as a system which obtains data inputs from a variety of sources, and transforms them into data outputs to serve a range of clients, and which operates within some, more or less well defined, conceptual framework. In terms of this simple model, four types of integration may be identified and described: integration of concepts; of data inputs; of data processing; and of data outputs. Although these aspects of integration are interdependent it is convenient to consider them separately. The logical starting point is integration of concepts, as this underpins all other aspects, but the main thrust for integration centres on outputs, thus the description begins there.

### 2.2 Integration of Data Outputs

Integration of data outputs implies more than simply bringing together separately collected data sets into a single output data base or dissemination vehicle. It implies that the data sets can be related, and that information can be obtained from them jointly which could not be provided by any one of them alone. Data integration may thus be characterized and quantified as the increase in information content of combined data sets relative to the sum of their individual information contents.

It is not difficult to illustrate why integration may produce synergy. Consider the information content of a data set of, say, n fields where, for simplicity, it is assumed that each individual field is equally informative. The overall information content obviously increases with n . The increase is likely to be more than linear because of the information embedded in the fields taken as groups. For example, if all individual data fields and pairs of fields are assumed equally valuable, then their information content is proportional to $n(n+1) / 2$. It follows that if two data sets with m and n fields are integrated then, under the same assumptions, the combined information content will be about $(\mathrm{m}+\mathrm{n})^{2} /\left(\mathrm{m}^{2}+\mathrm{n}^{2}\right)$ times the sum of the individual information contents. If larger groups of fields, say groups of 3 and 4 , are considered informative, then the gains due to integration will be still greater.

The above example assumes that the fields in the two data sets can be related pairwise, or in larger groups. To the extent that they can not, the potential for gain of information is correspondingly reduced. In essence, the scope for integration is determined by the degree to which the data sets respect a common conceptual framework, as will be further elaborated in the Section 2.3.

Integration can take place at micro (unit) or macro (aggregate) level. The lower the level the greater the potential for gain of information. However integration at micro level is often not feasible, for example when the data are derived from different sets of sampled units from the same population, or from different populations.

The benefits to be expected from integration of data outputs can be broadly summarized in three categories as follows.

## Servicing Cross-Cutting Data Needs

Policy is seldom made in consideration of a single variable alone. It is likely to involve several dimensions, and hence to require integration of data.

The more complex the society and/or the greater the extent of government intervention, the wider the range of statistical data requirements. One way of meeting new needs is to establish additional surveys. In some measure, this is how programs of economic statistics have grown piecemeal. However, in an era when additional resources are scarce, the development of new collection vehicles may not be economically feasible. The alternative is to squeeze more information from the existing collection processes. Integration of data outputs is a very important mechanism for doing so. It is a tool by means of which the agency can shift the emphasis of its outputs from being process oriented to being product/client oriented. The gains to be expected are in servicing needs which cut across existing data sets.

For example, integration of production data and financing data may result in information which neither data set alone could provide - knowledge of the profitability of individual investment activities is essential for prudent allocation of resources. For analysis and policy making it may be desirable to integrate data about imports with domestic production data, dividends paid with profits, foreign investment with capital expenditure, etc. Integration over time, i.e., the creation of longitudinal data sets, is another specific class of integration.

Viewed as an internal client of economic surveys, the System of National Accounts requires data across the full range of outputs. Data not integrated elsewhere within the economic statistics program must, de facto, be combined within the National Accounts. From the perspective of maximizing the information gain, integration is best done as an integral part of data production process rather than subsequently within the National Accounts where there is less scope for combination and comparison at micro level. Furthermore, integration at broad aggregate level within the National Accounts is not sufficient to service more detailed cross-cutting requirements.

## Improving Coherence and Accuracy

Only in attempting to integrate data outputs will their coherence be tested. Identification of inconsistency is a step in the direction of improving data relevance and accuracy. Furthermore, if the agency does not detect incoherence its clients may well do so. Lack of coherence may arise from differences in the units, classification systems or data item definitions, i.e. in the underlying conceptual basis, or it may point to errors in the individual data sets. This latter possibility indicates the potential for integration to uncover errors which would otherwise go undetected. It is the basis for macro editing.

## Identifying Data Gaps and Duplication

Another possible consequence of integration is the exposure of gaps or duplication in the data outputs. Such information is of assistance in determining where and when to redirect program resources.

## 23 Integration of Concepts

As previously noted, bringing together data sets does not necessarily imply integration. The amount of integration which can occur depends upon the extent to which the data sets are constructed from the same "information building blocks", i.e., within a common conceptual framework.

A conceptual framework appropriate for economic surveys can be defined in terms of four basic components:

- a model, i.e., a collection of economic data items and their relationships;
- sets of standard statistical units, about which data are to be collected from respondents, and for which data are required by clients;
- standard systems for classifying units by industry, geography, and size;
- standard data item definitions and adjustments, expressed in terms understood by respondents and acceptable to clients.


## Economic Model

The System of National Accounts provides a comprehensive conceptual framework for economic statistics that needs little further expansion (Statistics Canada, 1989). In addition, generally accepted accounting principles with respect to revenue and expense statements, balance sheets, etc., provide a loose, partial framework for business bookkeeping practices. Within this framework a business may maintain a number of accounts which are not integrated and may not be fully consistent with one another or with the National Accounts model. This must be taken into consideration when integrating data, as will be discussed shortly.

## Standard Statistical Units

In establishing standard units there are three major requirements to be taken into account. First, the units should provide complete coverage of all the activities considered to be in scope for economic statistics. In particular, the units should facilitate integration of data across businesses and institutions engaged in different industries (horizontal integration). Second, the units should be appropriate for collection of economic data of various types - employment, production, financial, etc. Different types of data are kept by operational units at different levels within large organizations. For example, wages and salaries are available from individual pay centres, whereas investment intentions may only be available from divisional or head offices. Thus, more than one set of standard statistical units is required; there must be a set for each level from which significant blocks of data are to be obtained. Third, it should be possible to integrate data collected at different levels (vertical integration). Thus, ideally, the sets of units should be hierarchical to enable roll up of lower level data to a higher level.

## Standard Classlfication System

Classification by industry is a requirement for virtually all cconomic statistics, though the desired level of industrial detail may vary from one client to another. Thus, a standard industrial classification system (or
systems) is required for each set of standard statistical units. It should be hierarchically arranged to allow for various degrees of aggregation.

Similarly a system for geographic classification of units is required to satisfy client demands for regional, provincial, and small area statistics. Again a hierarchical scheme is desirable to facilitate roll up of data from smaller to larger geographic areas.

## Standard Data Item Definitions

Ideally there should be a simple standard definition for each economic concept = employment, gross income, value added, etc. In practice, there may be compelling reasons for allowing and even standardizing some variants. Also, as businesses may not maintain integrated accounts in accordance with the National Accounts framework, there should be a set of standard micro-data adjustments, such as exist in macro-data form in the National Accounts, to enable integration of data.

Integration of concepts into a standard framework brings some direct benefits in terms of servicing clients who themselves require a conceptual basis for collection, classification or analysis of economic data. However the benefits to be expected are primarily indirect, accruing from the fact that a common conceptual framework is a prerequisite for all other forms of integration, and the benefits they bring.

### 2.4 Integration of Data Inputs

Traditionally, data requirements have been partitioned by statistical agencies into distinct groups of data items, e.g., production, financing, etc. Data are collected separately for each group. The particular partitions which exist, and which define separate "surveys", are a product of business accounting practices and the historical development of the economic statistics program.

Integration of data inputs presents an alternative approach. Data requirements are grouped by respondent rather than by data type, i.e., the focus is respondents rather than surveys. Thus, in place of, say, a dozen surveys each separately contacting businesses, data which meet the needs of all these surveys are obtained from each respondent through an integrated collection vehicle. In a sense this approach parallels the integration of data outputs by shifting the agency from being process oriented to being input (respondent) oriented. It is likely to be most beneficial for large businesses with complex organizational structures.

Integration of data inputs encourages establishment and use of collection arrangements that are well matched to respondent bookkeeping practices, reflecting availability of data and ease of access. This may produce a number of benefits. First, the agency will be more likely to ensure that the data demands imposed by its surveys are coherent and not duplicated. Second, the data collection arrangements can be chosen to place minimal burden on respondents. Third, the various data reported by a respondent are more likely to be consistent and coherent, in this respect limited only by the degree of integration and coherence of each respondent's own accounts.

### 2.5 Integration of Data Processing

Integration of data processing implies use of generic methods, and generalized systems. Examples of generic methods include the definition and use of standard sample designs from which the particular design for any new survey would be chosen, the definition and use of a standard pattern of treatment for non-response, etc. Generalized systems may be meta-systems from which a specific system appropriate for any particular survey can be generated, or they may comprise multi purpose programs with sufficient options to provide the variations required for a range of surveys.

There are three obvious benefits to integration of processing. First, if standard methods and systems are constantly used they become well tried and tested and less prone to error. Second, the existence of a body of standard methods and systems reduces the costs and time required to develop new surveys. Third, use of standard procedures facilitates integration of data inputs and outputs.

## 3. ACTIVITIES ASSOCLATED WITH INTEGRATION

There are a number of activities required for, and indicative of, integration in its various forms. They are briefly summarized in the following paragraphs. The more a statistical agency engages in these activities the greater its commitment to integration.

Conceptual framework. As previously noted, an integrated framework of units, classification schemes, and data items definitions is a prerequisite for integration. Such a framework must be constantly enhanced to keep abreast of economic and technological charges.

Business Register. A business register which provides classified sets of units and contact information to all economic surveys is the mechanism for putting the conceptual framework into operational form. In view of the constant creation and dissolution of economic units, comprehensive procedures for register maintenance are vital. In particular, large businesses must be regularly "profiled" to ensure that the agency's understanding of the appropriate units and collection arrangements is kept current.

Coordinated Data Collection Arrangements. Coordination of data collection arrangements across surveys, at least for the large units which are in all survey samples, is the basis for integration of data inputs.

Gemeric Methods and Systems. Development and maintenance of generic methods and systems is the basis for integration of processing.

Cross-Cutting Analyses. By definition, cross-cutting analyses imply integration of data outputs. The virtue of the statistical agency engaging in such analyses is that they help in assessing the relevance, coherence and accuracy of the outputs.

## 4. DEVELOPMENTS INVOLVING INTEGRATION

### 4.1 Introductory Remarks

The objective of this section is to illustrate the difficulties encountered in aiming for integration. Three developments involving integration at Statistics Canada (STC) are described. Three further examples are drawn from the Australian Bureau of Statistics (ABS) to indicate both the similarities, and the differences, in approach, relative to Statistics Canada.

In both agencies, the 1960's marked the start of a major thrust towards integration. Prior to that time the economic statistics programs had expanded largely as a result of new surveys being added, often independently of one another, in response to particular demands. The System of National Accounts was an integrating force within both agencies, but virtually the only one. There was little overall coordination. The focus was on survey processes rather than products. For example, there was intensive editing of the data at micro level, whereas macro editing, and the concern for compatibility across related data sets was very limited. Only in the National Accounts were individual data outputs likely to be brought together.

The following paragraphs describe attempts to move in the direction of integration over the last twenty years or so. Not all developments involving integration during the period are included, and the descriptions cover only those aspects of the developments which relate specifically to integration, not the full details.

### 4.2 Business Register Redevelopment: STC (1970-75)

Prior to the redevelopment, frame data for economic surveys at Statistics Canada were provided by survey master files which tended to be independent of one to another and to make varying degrees of use of data available from a centrally maintained list of businesses. The integration related objectives of the redevelopment were to replace the central list and most of the functions of the existing survey master files by a business register which would provide complete and unduplicated coverage (Sunter, 1971). The new register was to service the frame data requirements of all economic surveys by providing sets of statistical units, classified according to the 1970

Standard Industrial Classification, together with initial contact information for each unit. Maintenance of the new register was to be substantially based on central processing of administrative data from the Revenue Canada payroll deduction system. In summary, the objectives involved the introduction of standard sets of units and a standard industrial classification scheme, and the integration of frame data processing methods and systems.

The redevelopment was successful to the extent of creating a new business register of standard units with classification and contact information, which could be maintained using payroll deduction data. Use of data from this register resulted in substantial improvements to the coverage provided by survey frames. However, most survey operations retained their master files, in some cases with amalgamation into "divisional" master files. Over the next ten years the business register became utilized as a source of new units for surveys but not the primary generator of frames. Frames for some surveys continued to be based directly on income tax files. Thus, the full extent of the integration envisaged was never achieved.

The problems encountered, and the reasons cited for not fully adopting the new register, related to quality in all its aspects - relevance, accuracy, timeliness and cost. The new register contained fairly large numbers of inactive or out of scope units; many units were misclassified at detailed industry level; the register system was slow and ponderous to update; and the costs of adopting or replacing existing survey systems were high.

With hindsight, there are perhaps three lessons to be learned from this experience. First, use of integrated concepts, methods and systems has to be enforced. Without such legislation and constant monitoring, the multitude of specific, parochial difficulties, and the sacrifices of local optima which are needed, will outweigh the global thrust for integration. Second, integration can be expected to take a long time. The numerous parochial difficulties do have to be addressed, though not allowed to dominate. Third, fast turnaround by computer systems is vital, even if it means sacrificing some functionality.

## 43 Integration Project: STC (1970's)

The objective of the project was to integrate all data collection for the largest businesses in the country, accounting in total for about $50 \%$ of all economic production.

Notwithstanding a lengthy period of commitment, the objective was never close to being achieved, and the project was abandoned. With hindsight it is evident that the aim of integrating all data collection was too ambitious, given the limited accounting expertise, clerical resources and computing power made available to the project, and the fact that businesses themselves may not have integrated accounts. Substantive progress would have been more probable if the number of business and the range of surveys to be integrated had been smaller, or if the commitment of expert resources had been greater.

### 4.4 Business Register Redesign Project Phase 1: STC (1985-90)

The project was a major undertaking, affecting virtually all the major economic surveys. It was originally targetted for completion in three years, but was extended to six years. It was motivated initially by the need to address some serious data problems that were apparent when outputs from the economic survey program were integrated within the system of National Accounts.

The principal thrust of the project was to standardize and integrate both systems and data as reported by Colledge (1987):
"At the core of the project is the redesign of concepts, procedures and systems for provision of frame data and for the use of income tax data. In addition the project presents an opportunity to review the whole program of economic surveys and to develop general purpose systems and procedures".

The redesign started where previous integration efforts had left off. It envisaged (Cain et al, 1984) the creation of a new business register function and data base. The function was to:
"... support a full range of services, including the provision of classified lists of statistical and reporting units, the maintenance of these data using administrative and other sources, the provision of standards,
guidelines and procedures, and the sampling of survey universes and generation of mailing lists for surveys; and to co-ordinate sampling of income tax data, and acquisition, processing and use of such data to replace survey data."

The new data base was to store, maintain and provide access to frame data relevant to economic surveys using three types of records:
"... source records containing edited administrative or contact data which must be further processed to derive standard statistical records. The sources in question are the income tax and payroll deduction universes received regularly from Revenue Canada;
standard statistical records, which are the basis of survey frames. These records form a hierarchy of enterprises, companies, establishments, and locations;
reporting unit records, developed by negotiation between survey managers and respondents, and related to the statistical units appropriate to the survey".

A feature of the new frame data base, which reflected a compromise in terms of integration, was to rationalise use of maintenance resources by division of the base into two parts - an "integrated portion" and a "nonintegrated" portion (Cain et al, 1984).

The integrated portion of the frame will provide complete and unduplicated coverage of all large, complex, or otherwise significant statistical units. Priority in terms of both frequency and precision of maintenance will be given to these units. Additions or changes to the statistical units will be based primarily on direct contact, i.e. by survey or specific frame data enquiry.

The non-integrated portion of the CFDB (Central Frame Data Base) will consist of all of the other units which, because of their small size, do not warrant the cost of detailed support. These units, although in the majority in terms of numbers, account for only a small fraction of the economic activity represented by the integrated portion. The maintenance for this group will be almost entirely automated, using enhanced versions of systems currently in place. The non-integrated portion of the CFDB will be divided into two overlapping groups of statistical units: those derived from income tax source records, and those derived from payroll source records".

Facilities were to be provided to track units in both parts through time, thus providing the functionality for integration over time, and the scope for longitudinal analysis.

In summary, the integration objectives were:

- to create a new system of statistical units and collection arrangements better matched to business bookkeeping practices;
- to create an integrated business register, based on the new conceptual framework, with comprehensive facilities for storage and maintenance of frame data for large businesses and the capacity to track administrative and statistical units over time;
- to acquire and load the business register with up to date frame data obtained by direct contact ("a profiling") of large businesses and from administrative data sources for small businesses;
- to use the new business register as the source of frame data for each economic survey, with appropriate versions being available at any stage, from a preliminary version at the time of initial sample selection to a final version at the time of final estimation;
- to rationalize and extend the use of income tax data to supplement or replace direct annual data collection from small businesses.

These objectives have not yet been fully achieved. A comprehensive frame data model was developed (Statistics Canada, 1985). A new business register based on this model, with facilities for storage, longitudinal tracking and provision of frame data was designed, loaded and put into production (Cuthill, 1990). Facilities for sampling and utilizing financial data from income tax return, were redeveloped. However, as of today, the largest businesses
have not been profiled, the strategy for income tax data has not been implemented, and some surveys are not using the new register as the source of their frames. Furthermore, there are some doubts about the operational capacity to maintain frame data for the large businesses in the way originally intended. Also register operating costs are high given that full advantage has yet to be taken of the increased functionality, e.g. for longitudinal tracking. These problems will be tackled in a second phase of development.

It is evident that the time required to specify and implement the complex system and procedures implicit in the project strategy was vastly underestimated. With hindsight, some simplifications should have been considered. For example, the conceptual model may be more elaborate then is strictly necessary. Also the requirement for comprehensive longitudinal tracking facilities has certainly increased the costs of both implementation, and operations, without producing any substantive benefits to date.

It should be added that this description of the project focuses entirely on those objections related to integration, to illustrate the benefits being sought and the problems encountered in the quest for integration. There were other substantial objectives and achievements, for example the redesign of two business surveys, and the introduction of an automated register task scheduler, which are not discussed here.

### 4.5 Integrated Economic Censuses: ABS (1962-72)

A substantive thrust towards integration at the Australian Bureau of Statistics began in 1962 when work was first initiated on integrating the economic censuses and surveys. This led ultimately to a massive redevelopment and to integration of the censuses for 1968-69 reference period.

The basic rationate for the integrated approach was to: "... increase substantially the usefulness and comparability of the kinds of statistics already being collected and published, for purposes of general economic analysis and market research" (ABS, 1970), the objectives of the redevelopment were to put the collection of annual production on an integrated basis, i.e., in accordance with a common framework of reporting units, data concepts and industrial classification. The statistics for industrics covered by the censuses were to be produced without overlapping gaps in coverage, and in such a way that selected important data items such as value added, employment, salary and wages, fixed capital expenditures and stock were consistent across the industries. Vertical integration was to be facilitated by appropriate definition of statistical units. In addition the census data were to allow compilation of the Natural Accounts production account, and were to provide a basis for designing or adjusting sample particularly capital expenditures and stocks (important components of the quarterly national income and expenditure estimates).

The integration activities required to need these objective included (ABS, 1970):

- defining business units at standard level to strata in the business structure for which various types of economic statistics were required and could be collected; and devising standard rules for identifying such business units;
- identifying the standard units for all businesses and recording them in an integrated register to be used in the running of censuses and surveys;
- adopting a common system of industrial classification suitable for all censuses and surveys;
- defining in common terms the basic data items for which statistics were required across all industries;
- revising the questionnaires in accordance with the common definitions;
- revising the collection arrangements such that data were obtained through head offices, each of which was to be asked to report in a consistent way for each of the establishments in scope for the censuses and for the enterprisc.

The integration activities were substantially completed and a round of data collection under the new framework took place for the reference year ending June 1969 when the censuses of mining, manufacturing, utilities, trade and selected services were conducted for the first time in an integrated basis. However, although well planned and supported throughout the Bureau, the development fell short of its goals in a number of areas.

The integrated register was never brought fully up to date, even for the industry divisions in scope for the first round of collections. Though data availability was checked, the costs of extraction were not fully taken into
account for the smaller units. The results were higher respondent load and non-response rates than expected, and a requirement to divert ABS resources into editing and imputing for missing and poor quality data. The problems were compounded by a set of stringent edits and by cumbersome processing procedures. As a consequence there were long delays in publishing and a loss of rapport with both respondents and users. The data were too incomplete for their intended uses in the National Accounts.

The overall effects were a decision to abandon full scale integration as originally envisaged, and a long lasting reluctance by the organisation to undertake further developments along the same lines and on the same scale.

With the benefit of hindsight, the project was too ambitious on several counts. First, the base of operational knowledge was insufficient because the concepts were largely new. Second, the Bureau was relatively inexperienced in processing collections of the size and complexity involved. Running the censuses simultaneously and using a two tier (enterprise-establishment) collection approach involved substantial changes to previous practice. Third, the computing technology of the time - hardware and software - imposed far more limitations than it does nowadays.

The achievements of the project in terms of developing integration concepts and procedure were significant. They are detailed in an excellent paper published in the Australian Year Book (ABS, 1970). Apart from minor revisions required to take into account subsequent changes in business organizational structures, and in data processing technology, this paper is as relevant today as when it was written. It provides a very comprehensive and lucid summary of the elements and benefits of integration.

### 4.6 Rotating Program of Economic Census and Surveys: ABS (1971-87)

Following abandonment of the thrust towards fully integrated economic censuses, the Australian Bureau of Statistics set up a smaller scale project. Its objectives were to devise an alternative, more supportable approach to collection of annual reference period data, and to specify an appropriate processing system.

By 1973 the team had developed a new strategy. In place of the previous, periodic, economic wide snapshot approach, the term proposed a rotating program of censuses and surveys. Manufacturing, mining, utilities and agriculture were to be covered every year; other industry divisions were to be surveyed on a rotating basis. The first rotating collection under the new program was retail trade for 1973-74 reference year. Coverage was expanded gradually over the following 15 years. Agriculture was integrated within the program for reference year 1974-75. The first economic surveys of the construction, wholesale, and the transport industries took place for 1978-79, 1981-82 and 1983-84 reference years respectively. Two groups of selected service industries were included in 1986-87 and 1987-88 respectively.

Specifications for a new processing system resulted ultimately in the Integrated Economic Statistics Information System (IESIS) which is still operational today. IESIS makes use of the Generalised Interrogation System, which, although an Bureau product of 20 years standing, has many of the features of a fourth generation language and data base. A system specific to each census or survey is built within the IESIS standard framework. The IESIS approach is explicitly designed to handle the enterprise-establishment hierarchy, using a phased approach, establishment data being processed first.

In terms of integration objectives, the introduction of the rotating program was a step backwards, forced by practical necessity. Data for different industries collected in different years cannot be easily related. Rotation is a barrier to integration. On the other hand, industrial coverage has been expanded using a standard set of concepts and units. Furthermore, the IESIS system has been a great success in the integration context. It has been operational for twelve years, and instrumental in putting annual censuses and surveys on a standard footing.

### 4.7 Economic Statistics Strategy: ABS (1987-92)

The main theme of the Economic Statistics Strategy development, which began in 1987, is integration. In a sense, the development has picked up where the integrated Censuses development left off. The principal components of the strategy in relation to integration are:

- revision of the framework of statistical units and their classifications to produce closer alignment with business organisational structures and bookkeeping practices and a more clearly defined relationship to income tax units;
- revision and enhancement of the business register systems and procedures, first, to support the new statistical units framework, second, to establish and maintain more accurate structures and better reporting arrangements for large businesses, and third, to make more effective use of administrative data from the taxation authorities for maintenance of small business units,
- development and introduction of annual economy wide collection and dissemination of key structural data items leading at least to the derivation of gross operating surplus,
- development and introduction of quarterly economy wide collection of capital expenditure, stocks and operating surplus data,
- development and introduction of systems and procedures to produce economy wide small area data from the business register,
- development and introduction of systems and procedures to process income tax data to replace or supplement direct data collection.

The development is a substantial undertaking which is still ongoing. Progress to date may be summarized as follows.

A new statistical units model has been developed which though it is simpler, resembles quite closely the model developed during the Business Survey Redesign Project at Statistics Canada. It provides a good basis for integration, though not all the details of the higher level units for financial data have been fully thought out.

The business register system bas been upgraded to reflect the new units and improved processing procedures. The modifications took considerably longer than originally estimated. All large businesses have been profiled in accordance with the new units and the data register.

An economy wide annual survey of enterprises has been developed and is in its second year of testing. Blending the survey with the ongoing rotating establishment surveys posed many operational problems which have been more or less satisfactorily addressed. In particular, the data collection arrangements for large businesses were integrated. Procedures for integrated quarterly collection of data, production of small area data, and use of income tax data are still in the early stages of development.

### 4.8 Remark

The preceding examples refer to the most significant developments involving integration at Statistics Canada and at the Australian Bureau of Statistics over the last twenty years. However, there have been a number of other projects and proposals, some still active, with a strong flavour of integration. At Statistics Canada these include the Longitudinal Employment Analysis Program, the General Survey Function Development Project and the integration of survey operations. The current electronic data processing strategy stresses integration of computing facilities. At the Australian Bureau, the business register was overhauled and strengthened over 197984, a general purpose dispatch and collection control system was developed with a specialized version for annual economic surveys. Facilities for use of centralized data dictionaries have been installed on the main frame, and the EDP (Electronic Data Processing) strategy also emphasizes integration.

## 5. CONCLUSIONS

If quality is broadly defined in terms of four components - relevance, accuracy, timeliness and cost - then integration is a means of addressing each one of them. For example, relevance can be enhanced by integration of data outputs to increase the scope for cross-cutting analyses. Accuracy can be improved by integration of data
inputs from respondents to ensure micro level consistency across surveys, and/or by integration of outputs to check on their mutual compatibility at micro or macro level. The time required to satisfy a new demand can be vastly reduced if it can be achieved by integration of existing data outputs rather than creation of a new collection vehicle. Development costs can be trimmed if there is an integrated set of general purpose systems and procedures from which to draw.

Expressing the same ideas from a slightly different perspective, integration is a valuable tool for achieving stated agency objectives. In developing its 1991-92 plan, Statistics Canada defined several management and technical priorities which integration can help address, including: product and data quality; output enhancements; search for efficiency, program balance; and respondent issues.

In summary, there are enormous potential benefits associated with integration in its various forms. However, as the examples in Section 4 demonstrate, the road to integration is littered with abandoned objectives, time overruns, and corresponding losses of confidence. The quest for integration can result in attempts to build systems and introduce procedures which exceed the expertise and computing power available. There is always the danger that imposition of standard concepts or general purpose systems will lead to unacceptable loss of flexibility or timeliness for specific processes or outputs.

In conclusion, although integration is an elusive goal, it can yield such rich benefits that there is no option but to keep aiming for it. There is bound to be a tension between the framework imposed by integration and the requirements to satisfy specific needs or to optimize specific processes. In the absence of any general guidelines, local short term preferences which work against integration, will outweigh global longer term objectives. Thus integration has to be actively promoted by policy statement, if not by a functional unit. "Integration consciousness". i.e., staff awareness of the potential benefits of integration is vital.

Perhaps progress is most likely to be achieved in a series of small steps, taking advantage of opportunities as they arrive, with a major thrust from time to time. As an example of making incremental progress, the results of data integration in the National Accounts or during cross-cutting analyses should be systematically fed back to the surveys providing the data, with a view to identifying and correcting the sources of discrepancies. In planning and implementing major integration initiatives, it is important to balance objectives and practical limitations. Deviations and exceptions have to be recognized and permitted. The integrating framework must leave room for subsequent evolution.

## ACKNOWLEDGEMENT AND NOTE

The author wishes to thank Gordon Brackstone, Jacob Ryten, Nanjamma Chinnappa and George Sarossy for their helpful comments.

Although this paper contains examples drawn from Statistics Canada and from the Australian Bureau of Statistics, the opinions presented are those of the author and not necessarily of the agencies.

## REFERENCES

ABS (1970). Australian Year Book, Chapter 31, 1970, Canberra.
Bonnen, J. (1979). (Project Director), Improving the Federal Statistical System: Issues and Options, The President's Federal Reorganization Project for the Statistical System, February 1981, Washington, U.S.

Cain et al (1984). Infrastructure Development: Objectives, Policy and Strategy, Working Paper, Business Survey Redesign Project, November 1984, Statistics Canada, Ottawa.

Colledge (1987). The Business Survey Redesign Project: Implementation of a New Strategy at Statistics Canada, presented to the U.S. Bureau of the Census Third Annual Research Conference, March 1987.

Cuthill (1990). The Statistics Canada Business Register, Working Paper, Business Survey Redesign Project, August 1990, Statistics Canada, Ottawa.

Gigantes, T., Fellegi, I., Goldberg, S., and Podoluk, J. (1970). Micro Data Sets, Simulation and Statistical systems, Presented to the Workshop on Micro Data Sets, Washington, U.S., Statistics Canada, Ottawa.

Sarossy, G. (1987). Review of Economic Statistics, Working Paper, Industry Division, Australian Bureau of Statistics, Canberra.

Statistics Canada (1985). Business Statistics Survey Frame Model, Working Paper, December 1985, Business Survey Redesign Project.

Statistics Canada (1990). Minutes of Methods and Standards Committee, January 1990, Ottawa
Sunter, A. (1971). On the Construction and Maintenance of a Central Register for Business Surveys, Working Paper, Statistics Canada, Ottawa.

## SESSION 3

Coverage Measurement in Censuses of Population

# DIFFERENTIAL COVERAGE IN THE UNITED STATES CENSUS OF POPULATION: A HISTORIC REVIEW 

J. G. Robinson and H. Hogan ${ }^{1}$


#### Abstract

This paper discusses the findings from available estimates on differential coverage in the United States censuses. Estimates of the net undercount are available from demographic studies for censuses since 1880. They demonstrate a decline in net undercount, with the exception of the undercount of Black children and Black adult men. Other studies point to a differential undercount for other groups including Hispanics and for geographic areas. Results from the 1980 Census highlight the fact that the decline in net coverage is caused not only by the decline in Census omissions but the probable rise in census gross overcoverage. Because of the nature and distribution of coverage errors, even the exceedingly low undercount rate for the 1980 Census cannot be ignored by data users.


KEY WORDS: Undercount; Differential coverage

## 1. INTRODUCTION

The undercount of persons in the decennial Censuses of the United States has been recognized since the first Census in 1790. Between 1790 and the 1960 's, the issue of the Census undercount as a national concern lay essentially dormant, with two exceptions being the concern over the alleged large undercounts in the 1870 and 1920 Censuses.

Beginning with the years leading up to the 1970 Census, however, the topic of the undercount in the decennial Census has taken on a life of its own. Concerns about possible large undercounts in the 1970 Census were associated with the social unrest of those times and growing distrust of the government (U.S. Bureau of the Census, 1976: Chapter 7). The implications of the Supreme Court's "one-man-one-vote" decisions for state redistricting were well understood. The federal government was distributing money to local areas based, in part, on population through programs such as General Revenue Sharing, Urban Block Grants and Urban Mass Transit. By 1980, the Census undercount had become a political and legal issue, as well as a statistical one - the undercount debate has been centered on whether the official Census counts should be adjusted for purposes of apportionment, redistricting and other uses. Lawsuits and Congressional bills calling for adjustments have firmly brought the undercount issue into the political and judicial arena.

A casual observer of the activity surrounding the Census "undercount" might get the impression that coverage in the Decennial Census has worsened in recent times, thus prompting the need for adjustment. The fact is the net national undercount is at an all time low. In 1980 the net undercount was about one and a half percent, compared to an estimated net undercount of just under three percent in 1970 and over five percent in 1940. To understand what is happening, one cannot focus simply on the net national rate. Rather, one must remember that the net undercount is different across groups and it is made up on latger gross errors. For example, even though the net undercount in 1980 was at a record low, the relative undercount for Blacks (differential undercount) persists. What is also missing is the historical dimension in order to place the relative level of Census coverage in perspective. Is coverage really much better than it was 20,50 or 100 years ago? Have Blacks always been differentially undercounted? Has the relative coverage of race, sex and age groups changed over

[^8]time? Are the gains evenly distributed geographically? What roles do gross undercount and gross overcount play? Finally, with net overcount being so small, can data users comfortably ignore this factor in their work?

## 2. ESTIMATES OF COVERAGE OF THE U.S. CENSUSES SINCE 1880

Like the census itself, the evaluation of census coverage has improved with time. ${ }^{2}$ No analytic estimates of the coverage of population in the U.S. exist before 1880. For the Census from 1880 to 1940, separate estimates of coverage for the White and Black population have been prepared by Ansley Coale and coauthors. The U.S. Census Bureau has prepared consistent sets of estimates of coverage, i.e. estimates based on the same methodology, for all censuses beginning with 1940. These estimates are available by race, sex and age. These estimates of net undercount are based on demographic analysis.

In general, the demographic method of estimating coverage involves developing estimates for the population at the Census date by the analysis of various types of demographic data essentially independent of the Census, such as birth, death, and immigration statistics, as well as emigration estimates, and Medicare data. The difference between the estimated population and the Census count measures the net coverage in the Census.

The foundation of the demographic method is the logical consistency and relation of the underlying demographic data. With the use of components of change (births, deaths, net immigration) the estimated population of one Census for any age group (e.g., age $65-69$ in 1980) can be carried backward or forward to derive estimates of coverage in another Census (e.g., age 45-49 in 1960). In this way, the new estimates of coverage for 1940 to 1980 based on demographic analysis are internally consistent. We are now in the process of developing consistent demographic estimates of coverage for the 1990 Census.

Different methods have been used to develop estimates of coverage for specific race groups before 1940. Coale and Rives (1973) developed estimates of coverage of the Black population in the Censuses from 1880-1970. Coale and Zelnik (1963) prepared estimates of coverage of the native White population for 1880 to 1950. Each method involved the complex reconstruction of key populations. Robinson (1988) extended the historical scope of the series of coverage estimates to periods before 1940, by combining the Coale-Zelnik and Coale-Rives estimates to form composite estimates of coverage of the total population for the decennial years of 1880 to 1930.

The demographic estimates of coverage are obviously subject to error. Because they are based on better data and incorporate improved methods of estimation, the demographic estimates for the 1940-1980 period are believed to be of superior quality to those for the $1880-1940$ period. ${ }^{3}$

## 3. TRENDS AND DIFFERENTIALS IN COVERAGE, 1880 TO 1980

Estimates of coverage for the eleven decennial Censuses from 1880 to 1980 are shown in Table 1. Two general historical trends are depicted by these estimates: (1) a period of small, unsteady improvement in overall coverage between 1880 and 1940, with differing trends for Whites and Blacks, and (2) the emergence of large and persistent improvements in coverage since 1940 for both race groups, yet with a differential undercount in each census.

[^9]The rates in Table 1 demonstrate that net undercount levels for both Blacks and Whites were high in the Censuses before 1940 relative to levels of the most recent Censuses. First, the Coale-Rives estimates for Blacks indicate that the net undercount exceeded 10 percent in every Census from 1890 to 1940 with relatively high rates occurring in 1890 and 1920. Interestingly, the coverage levels for Blacks do not reveal any distinct trend of improving Census coverage up to 1940. This contrasts with the record since 1940. Next, the Coale-Zelnick estimates of net undercount for Whites reveal relatively constant coverage levels of about six percent up to 1920 , then a trend toward improving coverage beginning with the 1930 Census. A comparison of the undercount estimates by race in each Census (see last column of Table 1) reveals that the differential undercount of Blacks and Whites has long historical roots."

The demographic estimates of coverage for 1940 to 1980 for the total population and sex-race groups show that improvements in net coverage have occurred with each successive Census since 1940. By 1980, the net undercount of the total population had been reduced to less than two percent (1.4), compared to an overall undercount of over five percent in 1940 (5.6). The reduction in net undercounts was especially noteworthy between the 1970 and 1980 Censuses.

The Black population has continued to experience a much higher rate of undercount than the White population. In 1980, 5.9 percent of Blacks were missed as compared to only 0.9 percent of Whites; in 1940, the respective rates were 10.3 and 5.1 percent. As shown in Table 1 (last column), this differential undercount of five to six percentage points has persisted in each of the last five Censuses. Unlike the White population, where each successive Census has missed fewer persons, the net number of Blacks missed has remained roughly consistent (between 1.5 and 2.0 million) in each of the last five Censuses.

Table 1: Estimates of Percent Net Undercount by Race and Sex: 1880 to 1980
(Base of percent undercount is the estimated population. See text for description of the different estimates.)

| Year and source of estimate | Total | Male | Female | Black | White | Percentage point difference |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  |  |  |  | Male/ <br> Female | Black/ White |
| Census Bureau |  |  |  |  |  |  |  |
| 1980 | 1.4 | 2.4 | 0.4 | 5.9 | 0.9 | 2.0 | 5.0 |
| 1970 | 2.9 | 3.7 | 2.2 | 8.0 | 2.2 | 1.5 | 5.8 |
| 1960 | 3.3 | 3.8 | 2.8 | 8.3 | 2.7 | 1.0 | 5.6 |
| 1950 | 4.4 | 4.8 | 4.1 | 9.6 | 3.8 | 0.7 | 5.8 |
| 1940 | 5.6 | 6.1 | 5.2 | 10.3 | 5.1 | 0.9 | 5.2 |
| Coale \& Zelnick (White) |  |  |  |  |  |  |  |
| Coale \& Rives (Black) |  |  |  |  |  |  |  |
| 1940 | 5.0 | 5.4 | 4.5 | 12.7 | 4.1 | 0.9 | 8.6 |
| 1930 | 5.3 | 5.4 | 5.1 | 12.5 | 4.4 | 0.3 | 8.1 |
| 1920 | 6.7 | 6.2 | 7.2 | 15.1 | 5.6 | -1.0 | 9.5 |
| 1910 | 6.5 | 5.9 | 7.1 | 12.1 | 5.7 | -1.2 | 6.4 |
| 1900 | 6.7 | 5.9 | 7.7 | 10.9 | 6.2 | -1.8 | 4.7 |
| 1890 | 7.4 | 6.4 | 8.4 | 14.6 | 6.3 | -2.0 | 8.3 |
| 1880 | 6.5 | 5.5 | 7.4 | 9.2 | 6.1 | -1.9 | 3.1 |

Source: J. Gregory Robinson, "Perspectives on the Completeness of Coverage of Population in the United States Decennial Censuses", paper presented at the Population Association of America, New Orleans, 1988.

[^10]The historical trends in age patterns of coverage that underlie the trends described for race and sex groups are portrayed in the estimates of Table 2. In general, the undercount of the decennial Censuses has evolved from relatively high rates at most ages to relatively low undercount rates with the exception to two groups - Black children and Black adult males.

Table 2: Estimates of Percent Net Undercount for Selected Age Groups, by Race and Sex: 1890, 1910, and 1950-1980
(Base of percents is the estimated population)

| Race, sex and year | Under 5 | 5-19 | Age <br> 20-44 | 45-64 | 65 and over |
| :---: | :---: | :---: | :---: | :---: | :---: |
| White male |  |  |  |  |  |
| 1980 | 0.2 | 0.3 | 2.7 | 2.5 | 0.3 |
| 1970 | 2.5 | 1.7 | 3.6 | 3.5 | 2.3 |
| 1960 | 2.1 | 2.5 | 4.3 | 2.7 | 1.3 |
| 1950 | 4.4 | 3.2 | 4.5 | 4.1 | 2.8 |
| 1930 | 7.1 | 2.9 | 5.5 | 4.4 | 4.1 |
| 1910 | 7.3 | 4.3 | 5.4 | 5.3 | 3.6 |
| 1890 | 4.6 | 4.9 | 6.9 | 6.1 | 3.6 |
| White female |  |  |  |  |  |
| 1980 | 0.1 | 0.1 | 0.2 | -0.3 | 0.7 |
| 1970 | 2.1 | 1.3 | 1.2 | 1.8 | 3.6 |
| 1960 | 1.3 | 1.5 | 1.8 | 3.9 | 4.4 |
| 1950 | 3.8 | 2.6 | 2.4 | 6.1 | 4.9 |
| 1930 | 6.2 | 1.6 | 2.6 | 9.2 | 12.1 |
| 1910 | 6.9 | 2.8 | 4.7 | 14.8 | 10.2 |
| 1890 | 4.8 | 3.4 | 8.6 | 15.4 | 10.2 |
| Black male |  |  |  |  |  |
| 1980 | 9.6 | 2.6 | 13.2 | 13.6 | -1.4 |
| 1970 | 10.4 | 5.3 | 17.6 | 12.1 | -0.7 |
| 1960 | 7.1 | 6.6 | 17.4 | 11.7 | -6.6 |
| 1950 | 9.8 | 11.0 | 14.7 | 13.9 | -14.4 |
| 1930 | 23.4 | 11.7 | 12.9 | 7.6 | 10.4 |
| 1910 | 20.8 | 11.2 | 12.7 | 1.4 | -22.4 |
| 1890 | 28.5 | 9.7 | 15.5 | 6.9 | -27.9 |
| Black female |  |  |  |  |  |
| 1980 | 9.0 | 2.3 | 3.5 | 2.4 | -0.3 |
| 1970 | 9.5 | 4.6 | 5.9 | 6.0 | 2.0 |
| 1960 | 5.4 | 5.1 | 6.7 | 11.2 | -3.8 |
| 1950 | 9.0 | 8.5 | 5.2 | 16.5 | -17.4 |
| 1930 | 22.4 | 8.7 | 4.9 | 25.3 | 27.0 |
| 1910 | 21.0 | 9.7 | 9.5 | 22.9 | 13.2 |
| 1890 | 28.9 | 8.2 | 14.3 | 24.3 | 4.3 |

Source: Robinson (1988)
The estimates in Table 2 show that for White males and White females at all ages, for Black females over the age of 5 , and for Black males aged 5-19 and 65 and over, the undercount rates in 1980 are much lower than the undercount levels estimated for the early Censuses. By historical standards, the net undercount for these groups
is essentially eliminated. For Black children (under age 5), the undercount rates have dropped in half from high levels of over 20 percent to around 10 percent in 1980 -- a level that is still high in comparison to most other groups. Black adult men present the sole exception to the significant improvements in coverage that have been made over the early to modern Census era. For Black men aged 20 to 44 , coverage has remained essentially unchanged at net undercount rates of around 15 percent from 1890 to 1980 (Table 2). For Black men aged 45 to 64 , the estimates indicate that coverage has decreased in the later half of this century compared to earlier years, with the net undercount exceeding ten percent in every Census since 1950. The role that these two groups play in contributing to the much cited differential undercount of Whites and Blacks should be emphasized -- if the undercount rates of Black adult males (aged 20-64) and Black children under 5 were the same as the remaining Black population, the differential undercount in 1980 would be reduced from 5.0 to 1.5 percentage points.

## 4. SUBNATIONAL AND SOCIOECONOMIC DIFFERENTIALS IN COVERAGE

Less is known about the trends and variations in coverage for geographic areas within the country or for socioeconomic groups than for the levels and trends in coverage for the nation as a whole. In this section, we examine some of the available evidence.

In nearly all evaluation studies for the 1970 Census and earlier, the net undercount was higher in the South than in any other region (Table 3). Part of this differential was attributable to the fact that Blacks-who have been undercounted at higher rates than Whites-were concentrated in the South. It also resulted from the higher-race-specific undercount rates of Whites and Blacks in the South. In general, the studies indicate that undercount rates in the West tended to be intermediate between the South and North, and undercoverage was lowest in the North. In 1980 a new regional pattern of coverage emerged-the net undercount rate was highest in the West and relatively lower in the North and South. The significant drop in net undercount in the South (for both Whites and Blacks) was the main reason for the regional shift in coverage patterns between 1970 and 1980. Differential patterns in gross overcoverage may also have played a role. (See Section 4).

Table 3: Undercoverage Rates Shown by Evaluation Studies, For Regions: 1970-1980 (Rates for population represent the percentage missed based on corrected population)

| Group and year <br> States | Northwest | North <br> Central | South | West |  |
| :--- | :---: | :---: | :---: | :---: | :---: |
| 1980 Demographic Analysis |  |  |  |  |  |
| All Classes | 1.4 | 1.1 | 0.4 | 1.0 | 3.5 |
| White | 0.9 | 0.6 | 0.3 | 0.5 | 3.0 |
| Black and other races | 4.3 | 5.6 | 4.7 | 3.3 | 6.8 |
| 1980 Post-Enumeration Program |  |  |  |  |  |
| Net Undercount - Set 2-9 | 1.6 | 1.1 | 1.3 | 1.2 | 3.1 |
| All Classes | 0.7 | 0.1 | 0.5 | 0.1 | 2.5 |
| White | 7.2 | 9.5 | 7.7 | 5.7 | 10.3 |
| Black | 2.6 | 1.4 | 1.4 | 4.0 | 3.3 |
| 1970 Demographic Analysis | 1.9 | 0.9 | 1.1 | 3.0 | 2.8 |
| All Classes | 6.9 | 5.5 | 5.1 | 7.8 | 7.9 |
| White |  |  |  |  |  |
| Black and other races |  |  |  |  |  |

[^11]Table 4: Estimated Differential Undercount as Estimated by 1980 Post-Enumeration Program

| Series | Black | Non-Black <br> Hispanic |
| :--- | ---: | :---: |
| $2-8$ | 5.0 | 3.6 |
| $3-8$ | 4.7 | 3.5 |
| $2-9$ | 5.8 | 4.3 |
| $3-9$ | 5.5 | 4.2 |
| $14-9$ | 2.8 | 1.7 |
| $2-20$ | 5.9 | 4.2 |
| $3-20$ | 5.7 | 4.2 |
| $14-20$ | 3.0 | 1.7 |
| $5-8$ | 2.8 | 4.9 |
| $10-8$ | 2.5 | 3.4 |
| $5-9$ | 3.6 | 5.7 |
| $14-8$ | 2.1 | $1.0^{*}$ |
| Approx. s.e. | 0.6 | 0.8 |
| (Sets 2, 3, 14) | 0.6 | 1.0 |
| (Sets 5, 10) |  |  |

- Not significant at $90 \%$ confidence level

Findings on the association of coverage and socioeconomic status and other variables are summarized in U.S. Bureau of the Census (1975), pp. 8-10 and in Citro and Cohen (1985), Appendix 5.1. As an example, the 1970 Census CPS Match Study found that the "missed" rate of persons in families earning less than $\$ 7,500$ in 1969 (4.4 percent) was double the missed rate in families earning more than $\$ 7,500$ ( 2.1 percent); the missed rate for persons who had completed only eight years of schooling or less ( 3.8 percent was significantly bigher than for persons who had some high school education ( 2.5 percent (U.S. Bureau of the Census, 1975, p. 8-9).

Undercount may be higher for some ethnic groups, such as Hispanics. Table 4 gives the estimated differential undercounts for Hispanics as well as Blacks from the 1980 PEP (Post Enumeration Program) estimators. There were several sets of estimates based on a range of assumptions. The Census Bureau was quite concerned about the bias of the estimates. The twelve sets are based on different data and different assumptions in an attempt to show the sensitivity of estimates to possible violations of assumptions. ${ }^{5}$ We can look at the differential undercount implied for each set of estimates. Subtracting the estimated national undercount tends to remove any uniform bias from the sets, but will leave any bias that is differential between areas. Again, a differential undercount is shown. This finding demonstrates that, although we lack the historic data, the problem of differential undercount is not confined to Blacks. Indeed, according to the 1980 PEP, the measured differential undercount for Hispanics appears to be similar to that of Blacks. (See also U.S. Bureau of the Census, 1979).

Studies of coverage that deal with the type of residence (such as urban/rural) have fairly consistently found that the omission rate of occupied housing units is higher in rural areas than urban areas, though within urban areas the findings on variations in coverage by size of place is mixed. The 1980 PEP was the first study where the coverage of individual cities could be evaluated. The 1980 coverage evaluation program produced estimates of the undercount for 16 central cities.

Table 5 gives the difference between the cities' estimated undercount and the national undercount for each city and set of estimates. Many of the individual estimates are not significant at the 90 percent level. However, if cities as a group have the same undercount as a nation as a whole, one would expect that by chance the estimates of only two of the sixteen cities would be significant at the 90 percent level $(0.1 \times 16<2)$. This outcome is only

[^12]seen for the " $14-8$ " estimates. For the other series, four or more cities exhibit a statistically significant undercount. ${ }^{6}$

Table 5: Estimates of Differential Undercount for 16 Cities from 1980 Post-Enumeration Program

|  | $2-8$ | 2-9 | $3-8$ | 3-9 | Series |  |  | 10-8 | 14-8 | 149 | Stimdard Errors |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  |  |  | 5-8 | 5-9 | 3-20 |  |  |  | 2,3,14 | 5,10 |
| Ballimore | 4.7 | 4.6 | 4.4 | 4.3 | 2.9 | 2.8 | 4.8 | 2.4 | 4.2 | 4.0 | 1.7 | 1.7 |
| Boston | -1.9 | 0.8 | -2.0 | 0.7 | -0.3 | 2.6 | 0.7 | 0.1 | 4.3 | -15 | 4.1 | 7.2 |
| Chicago | 25 | 3.9 | 3.4 | 4.8 | 2.1 | 3.6 | 4.9 | 1.8 | -0.6 | 0.8 | 15 | 1.8 |
| Cleveland | 3.6 | 35 | 3.0 | 3.8 | 5.5 | 5.4 | 3.5 | 5.0 | - 2.7 | 2.5 | 2.2 | 2.5 |
| Dallas | 5.9 | 6.7 | 4.9 | 5.7 | 3.0 | 3.9 | 5.8 | 3.0 | -0.6 | 0.2 | 20 | 1.8 |
| Detroit | 2.2 | 2.8 | 2.1 | 2.7 | 1.9 | 25 | 3.1 | 0.4 | 1.4 | 2.0 | 2.3 | 1.8 |
| Houston | 3.7 | 4.3 | 3.6 | 4.2 | 6.4 | 7.0 | 4.7 | 5.2 | -25 | -1.8 | 2.6 | 2.8 |
| Indianapolis | -0.8 | -1.2 | -1.2 | -1.6 | 4.1 | 3.7 | -1.8 | 2.9 | -15 | $-2.0$ | 2.1 | 2.7 |
| Los Angeles | 4.2 | 6.0 | 3.6 | 5.4 | 1.6 | 3.4 | 5.3 | 1.6 | 1.8 | 3.5 | 1.4 | 1.2 |
| Mihaukee | 2.1 | 1.7 | 2.1 | 1.7 | 0.3 | -0.1 | 15 | 1.0 | 2.2 | 1.7 | 1.4 | 1.4 |
| New York | 5.3 | 5.9 | 5.0 | 5.6 | 1.5 | 2.1 | 5.8 | 1.7 | 1.8 | 2.3 | 1.2 | 1.0 |
| Philadelphia | 4.8 | 4.9 | 3.7 | 3.8 | 1.1 | 1.3 | 3.8 | 1.2 | 2.5 | 2.5 | 1.9 | 1.5 |
| St. Louis | 2.0 | 2.1 | 2.1 | 2.2 | 3.0 | 3.1 | 3.2 | 1.0 | 1.3 | 1.3 | 2.1 | 2.8 |
| San Diego | -2.1 | -1.0 | $-2.0$ | -0.9 | -1.7 | -0.6 | -0.3 | -0.9 | -1.0 | 0.0 | 2.3 | 2.4 |
| San Francisco | 3.2 | 3.7 | 3.6 | 4.1 | 0.6 | 1.1 | 4.6 | -0.8 | 1.0 | 1.4 | 2.9 | 1.9 |
| Washington, DC | 2.9 | 4.4 | 2.6 | 4.2 | 1.2 | 0.4 | 5.0 | -1.1 | 2.4 | 3.9 | 1.6 | 2.2 |

Another approach is to compare estimates from the different series for each city using boxplots (Figure 1). One must remember that the April estimates (Sets 2, 3, 14) are highly correlated with each other, as are the August estimates $(5,10)$, while the correlation between April and August is weak, with a correlation coefficient of only .2 or .3. Leaving aside the biases, certain patterns tend to emerge. The 1980 PES showed no differential undercount for Boston, Indianapolis, or San Diego. The pattern for such cities as Milwaukee, San Francisco, St. Louis, Detroit and Washington is weak but suggestive of a differential undercount. The pattern for Baltimore, Chicago, Cleveland, Dallas, Houston, Los Angeles, New York and Philadelphia is largely consistent with the national average. So while caution must be used, both in terms of bias and variance, there is evidence in the 1980 PES to support the idea of a differential undercount for central cities of some large metropolitan areas.

There is evidence that the undercount is higher for the poor, for the single, and for the unemployed. Undercount is higher for those who rent than for those who own their home (Isaki, et al. 1987). Differences have been shown with many other social variables as well. See Fein and West (1988) for a discussion with respect to the 1986 Census of East Central Los Angeles County.

[^13]Figure 1. Eetimates of Differential Undercount for 16 Chies


## 5. GROSS COVERAGE ERROR

Although this paper has focused on the net undercount of the population, the net rates are actually the product of larger gross omissions that are partially offset by gross erroneous inclusions. So, beneath the pattern of net differential undercount is a pattern of omissions and a pattern of duplications (people counted more than once, falsifications by Census enumerators, and other errors). While the trends in the net undercount are well documented, the size and trends in the separate components of omissions and duplicates are much less clear.

In 1980, the net undercount was on the order of 1.4 percent or 3.2 million people. According to the 1980 evaluation, at least 2.7 million people were counted more than one time (Cowan and Fay, 1984; Jones, 1986). ${ }^{\text {² }}$ This implies that at least 5.9 million people were omitted. The result is to omit millions of real people from the data, and add in many enumerations that should not be there. If the erroncous enumerations were distributed differently, either geographically or demographically, than the omitted people, the implications for the data user are more serious than implied by a net undercount of 1.4 percent.

The problem of gross overcount and gross undercount is related to the problem of differential net undercount. For example, most of the duplicated housing units occurred in prelist areas (Bureau of the Census, 1985), that are outside the central cities of metropolitan areas. The overenumeration rate of occupied housing units in areas outside MSA's (1.20\%) was higher than that for areas inside MSA's ( $0.78 \%$ ). In those areas of the country enumerated conventionally (i.c., not mail-out-mail-back), the overenumeration rate was extremely low ( $0.11 \%$ ). Indeed, unless the gross erroneous enumerations are distributed identically by age, race, sex and geography as are the gross omissions, then a study of differential coverage must include analysis of gross coverage errors.

[^14]Data on the coverage of units sheds important light on the differential coverage by region. Table 6 gives the gross omissions and gross overcount by region. The gross undercoverage in the South was about the same as the West. However, the South also had the highest level of gross overenumeration, with the West's rate among the lowest. The Census Bureau concluded "Although equally thorough evaluations of duplications had not been conducted for earlier censuses, the evidence implies much lower levels of duplications then. Thus, regrettably, duplication receives dubious credit for part of the improvement in 1980 in net census coverage." (U.S. Bureau of the Census (1988) p.10.)

Table 6: Estimated Gross Coverage Error by Region: 1980

|  | United <br> States | Northeast | North <br> Central | South | West |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Gross <br> omissions | 5.6 | 5.1 | 4.3 | 6.5 | 6.1 |
| Gross <br> overcounts | 2.9 | 3.3 | 2.1 | 3.6 | 2.0 |

Charles D. Cowan and Robert E. Fay, (1984); (Set 2-9)

## 6. IMPLICATIONS OF THE DIFFERENTIAL UNDERCOUNT

One role of coverage evaluation is to inform data users of the implications. What are the effects of the undercount on data users? We might roughly divide the users of Census data into four categories: political, statutory, social science and planning.

Many political uses focus only on the count. Apportionment is a prime example. Differential undercount plays a role only if it is reflected geographically. However, because of the Voting Rights Act, differential coverage by race, ethnicity and even age, can play a direct role in the use of Census data for redistricting. Many government programs use population as part of the formula to distribute federal money. For most of these, only the net undercount plays a direct role.

There are other legal uses of population counts besides funds allocation. The number of people in an employment area for different race groups can plan an important part in court cases over job discrimination. Decennial census data are commonly used to address whether the jury selection system in a particular jurisdiction is representative. Race is the most common characteristic, but some challenges have specified sex, age, occupation, education, and/or poverty level. (See Rolph, 1986) Differential coverage by race and age can distort the estimated 'job pool' or 'jury pool's.' Inclusions of erroneous enumerations can distort both the size and distribution of the estimated populations.

Census data play an important role, directly or indirectly, in much social science research. Looking at just a few uses gives an idea of the importance of understanding Census undercount.

If we look at the ratio of observed Black male crude death rate (CDR) in 1980 to the observed white male CDR, then we get a $5 \%$ excess (1.05). The true ratio will equal coverage rate for Black males (.912) to the rate for White males (.983) or:

$$
.912 / .983=.928
$$

The ratio of the true rates will be 1.05 * $.928=0.9744$
The National Center for Health Statistics has, at times, included instructions in the vital statistics report on how to correct the rates for Census undercoverage and provided the necessary factors (U.S. Department of Health and Human Services, 1981).

A common measure of the extent of social problems in the prevalence rate: that is, the number of events per person. The analysis here is similar to that of crude death rates except it is no longer always safe to assume accurate reporting of the event. The event may be more poorly reported than the population. AIDS cases may be badly underreported. Still, a researcher interested in the prevalence of AIDS in certain inner city populations would do well to take Census coverage into account. [See Shai and Aptekar (1990)].

Expectation of life is a more complex measure of mortality than crude death rate. The 1980 undercount does not greatly influence either the expectation of life at birth or expectation of life at age 65. This was not true for previous censuses. The 1970 Census undercount lowered the observed expectation of life at birth by 1.5 years for Black males. In 1980 correcting for Black male undercoverage raises expectation of life at age 20 by more than one year.

Census sex ratios can be misleading. The Census misses more males than females. This is especially true among Black adults. If we look at the Census and estimated true sex ratios, we can see the effects. The observed 1980 sex ratio for Blacks, aged 35-54, was 13 percent too low. What is the effect on our perceptions of Black problems, Black social patterns, Black living quarters? One small ethnographic study of a Black neighborhood in 1970 found that while Census Bureau data indicated that 72 percent of households were headed by females, only 12 percent actually were. On the basis of this finding, the ethnograhers argued that biased census data create and support a distorted image of the female-headed Black household. (See Hainer et af., 1988).

In an article on Black marriage pattern Goldman, et al. (1984) were forced to deal with this issue.
"Because of the severity of the problem among Black males, we have adjusted the whole unmarried population, specific for race, age, and sex ... Since the undercount is probably greater in the unmarried population, there may be further underrepresentation of men, especially Black men, in our estimates of the marriage market."

Their results would have been different if they had ignored the issue.
Measured intercensal growth rates are affected by the undercount. The resident population of the United States increased between 1970 and 1980 by 11.4 percent based on official Census counts. When the comparison is based on the estimated true population, the growth rate is only 9.7 percent for this decade. The same distortion is no doubt occurring at the local and state level as well.

Differential undercoverage can distort other data series as well. The Census is used as a sampling frame for other surveys, such as the Current Population Survey, CPS. Since only the housing units are sample, the undercount of people in the Census does not directly affect the sample. However, these surveys have undercoverage problems of their own. Overall, CPS coverage is about seven percent lower than the Census coverage. Undercoverage of Black males is 17 percent lower than the Census and Black males $20-24$ are 27 percent worse. (Hainer, et al., 1988). One of the uses of the Census is to correct for coverage problems in other surveys. The Bureau statistically controls the CPS to agree with the projected data. So the undercoverage of the Census is carried into the CPS.

The private sector uses of Census data is extensive. The impact of undercount there has not gone unnoticed. To quote Mayor Richard Berkeley of Kansas City (1988) "It can happen in the private sector as well, where in fact, people make advertising buys and things of that nature based on the population in a community or in a region." The undercount can affect decisions of advertisers to buy time on, for example, Spanish language stations. The effect of the differential undercount in the private sector represents an untapped research area.

## 7. CONCLUSIONS

Net census undercount is a complex subject, the result of both census misses and erroneous inclusions. Its occurrence is differential across groups and among areas. Similarly, its impact is differential for different uses. The methods we use to measure it are complex and subject to errors of their own. In planning for 1990 , we have put together a multifaceted program. Demographic analysis and the post-enumeration survey are the twin pillars. In addition, we have developed a program to evaluate our coverage measurement program itself, and a program of participant observation to give us insight into the twin pillars. Together, these studies should provide botb the data user and the student of Census methods more accurate measures of Census coverage.

## REFERENCES

Berkeley, R. (1988). Interview on Moming Edition, National Public Radio, January 10, 1988.
Citro, C., and Cohen, M.L., eds. (1985). The Bicentennial Census: New Directions for Academy Press, Washington, D.C.

Coale, AJ., and Zelnik, M. (1963). New Estimates of Fertility and Population in the United States, Princeton University, Press, Princeton, New Jersey.

Coale, A.J., and Rives, N.W. (1973). A Statistical Reconstruction of the Black Population of the United States, 1880-1970: Estimates of True Numbers by Age and Sex, Birth Rates, and Total Fertility, Population Index 39(1):3-36.

Cowan, C., and Fay, R. (1984). Estimates of Undercount in the 1980 Census, Proceedings of the American Statistical Association Annual Meetings.

Cowan, C., and Fay, R. (1984). Estimates of Undercount in the 1980 Census, Pnoceedings of the American Statistical Association Annual Meeting.

Fein, DJ., and West, K. (1988). The Sources of Census Undercount: Findings from the 1986 Los Angeles Test, Survey Methodology, 14, 2, 223-240.

Goldman, N., Westoff, C.F., and Hammerslough, C. (1984). Demography of the Marriage Market in the United States, Population Index, 50, 1.

Hainer, P.C. (1987). A Brief and Qualitative Anthropological Study Exploring the Reasons for Census Coverage Error Among Low Income Black Households, unpublished.

Hainer, P., et al. (1988). Research on Improving Coverage in Household Surveys, Fourth Annual Research Conference Proceedings.

Isaki, C.T., Schultz, L.K., Smith, P.J., and Diffendal, G.J. (1987). Small Area Estimation Research for Census Undercount: Progress Report, Small Area Statistics, An Intemational Symposium, New York: John Wiley \& Sons.

Jones, C.D. (1986). Memorandum for Barbara Bailar, Duplicate Enumerations, October 7, 1986.
Robinson, J.G. (1988). Perspectives on the Completeness of Coverage of Population in the United States Censuses, Paper presented at the Population Association of America, New Orleans, Louisiana.

Robinson, J.G. (1986). Temporal and Regional Variations of the Coverage of Population in the 1980 and 1970 Censuses, paper presented at the annual meeting of the Southern Regional Demographic Group, 1986.

Rolph, J. Discussion. Proceedings of the Second Annual Research Conference, 344346.

Schaefer, J. (1989). Missing Data Procedures in the 1980 Post Enumeration Program: Why PEP Series 14 Cannot Be Trusted, unpublished manuscript, Harvard University, Cambridge, Massachusetts.

Shai, D., and Aptekar, L. (1990). Factors in Mortality by Drug Dependence Among Puerto Ricans in New York City; American Joumal of Drug and Alcohol Abuse, 97-107.
U.S. Bureau of the Census (1973). 1970 Census of Population and Housing Research and Evaluation Reseanch Program, PHC(E)-5, The Coverage of Housing in the 1970 Census, U.S. Government Printing Office: Washington, D.C.
U.S. Bureau of the Census (1975). U.S. Census of Population and Housing: 1970, Procedural History PHC(R)-1, U.S. Government Printing Office, Washington, D.C.
U.S. Bureau of the Census (1979). Coverage of the Hispanic Population of the United States in the United States in the 1970 Census, Current Population Reports, Series P-23, No. 82, U.S. Government Printing Office: Washington, D.C.
U.S. Bureau of the Census (1985). 1980 Census of Population and Housing, Evaluation, and Reseanch Reports, PHC80-E1, The Coverage of Housing in the 1980 Census, U.S. Government Printing Office: Washington, D.C.
U.S. Bureau of the Census (1988). 1980 Census of Population and Housing Evaluation and Research Reports, PHC80-E4, The Coverage of Population in the 1980 Census, U.S. Government Printing Office: Washington, D.C.
U.S. Department of Health and Human Services (1981). Vital Statistics of the United States 1977, Volume II, Part A, Technical Appendix and Table 6-4.

## NOTE

The views expressed are attributed to the authors and do not necessarily reflect those of the Census Bureau.

# THE MEASUREMENT OF NET COVERAGE ERROR IN CANADIAN CENSUSES 

R.G. Carter ${ }^{\text {t }}$


#### Abstract

The coverage of censuses of population can be evaluated either by demographic methods or survey-based estimates of coverage errors, or by a combination of the two. Although most countries that attempt direct measurement of coverage errors conduct a Post-Enumeration Survey (PES), Canada has adopted a somewhat different approach. Since 1966, the coverage of Canadian censuses has been assessed by means of a Reverse Record Check (RRC), in which a sample of eligible persons, selected in advance, is matched to census documents to determine whether or not the selected persons were enumerated. While the RRC yields estimates of undercoverage, no direct estimate of net coverage error (that is, undercoverage minus overcoverage) has been available. The assumption has always been that overcoverage is small, but until the 1986 Census, no attempt had been made to measure it. In 1986, however, an experimental overcoverage study was conducted. Although the results were limited, the 1986 study provided valuable experience in measuring overcoverage, and a larger and improved Overcoverage Study (OS) is planned for the 1991 Census. Based on the RRC and OS, it is intended that national and provincial estimates of 1991 net coverage error will be produced and disseminated.


In this paper, an outline of the 1991 coverage measurement programme is presented. The reasons for opting for a separate overcoverage study rather than a PES are also discussed. In addition, demographically-derived estimates of coverage error are considered as a means of evaluating the results of the RRC and OS and of estimating net coverage error for censuses prior to 1991.

KEY WORDS: Undercoverage; Overcoverage; Reverse Record Check; Error of closure.

## 1. INTRODUCTION

Traditionally, the primary purpose of a Census has been to count the number of people and households resident in a country and its various administrative sub-divisions. Consequently, one of the most important considerations for evaluating the quality of Census data is the coverage achieved: that is the proportion of the population actually counted in the Census. Coverage errors can arise at various stages of collecting and processing Census data. Most such errors result in undercoverage, that is the omission of houscholds or of some of the houschold members. Some coverage errors, however, such as the inclusion of out-of-scope or fictitious people and the enumeration of some people more than once, tend to increase the Census count relative to the true population and are referred to as overcoverage. For most purposes, interest centres on the difference between the actual Census count and the true population, this difference being the net effect of undercoverage and overcoverage. However, to the extent that missed or double counted persons or households have characteristics which differ from one another and from those correctly enumerated, erroneous conclusions may be drawn even when net coverage error is zero. Hence, in attempting to evaluate coverage error, it is useful to measure both undercoverage and overcoverage and to identify the characteristics of the persons or households concerned.

[^15]Various methods have been developed to measure coverage errors. These may generally be categorized as either "micro" or "macro", and either based on sample survey or administrative data. "Micro" methods involve the matching of individual records to identify cases missed or wrongly enumerated, whereas "macro" methods rely on comparison of aggregates. Many countries employ more than one approach to evaluate the quality of their Census. For example, the United States Bureau of the Census uses both a micro/sample survey (PostEnumeration Survey) methodology and a macro/administrative record (Demographic Analysis) approach.

Statistics Canada also uses two basic approaches to the evaluation of coverage errors. First, demographic (macro/administrative record) methods are used to estimate the change in the population from one Census to the next. In conjunction with the corresponding Census counts, this yields an estimate of the change in net coverage error (see Section 3). Secondly, a micro approach based on samples from the previous Census and administrative files is used to measure undercoverage (Reverse Record Check). A micro/ sample survey method for measuring overcoverage was tested in the 1986 Census and will be implemented fully in 1991.

In the following sections the two basic approaches used by Statistics Canada will be described in more detail. Section 2 will outline the Reverse Record Check methodology and its limitations and present some results. The 1986 test of an overcoverage study will also be briefly described. The modifications to these studies planned for 1991 will be discussed together with the rationale for choosing this approach over a Post-Enumeration Survey methodology. The demographic method will be described in Section 3 and the potential use of the demographic approach to estimate net coverage error for Censuses prior to 1991 will be discussed in Section 4.

## 2. SURVEY-BASED ESTIMATES OF UNDERCOVERAGE AND OVERCOVERAGE

### 2.1 The Reverse Record Check

## Methodology

The basic idea behind the Reverse Record Check (RRC) is to select a sample of individuals who should be enumerated in the Census, trace them to their Census Day address, and then verify whether or not they were enumerated. Traditionally, the sample has been drawn from four frames:-
(1) Persons enumerated in the previous Census; and
(2) Persons missed in the previous Census.

Together these frames cover the population five years previously, and to these are added:-
(3) Inter-censal births; and
(4) Inter-censal immigrants.

Conceptually, these four frames cover the target population for the current Census (plus persons who have died or emigrated, and who must therefore be identified during tracing as no longer part of the Canadian population). With the exception of the "Missed" frame, a stratified sample is selected from each frame. The samples are matched to an administrative file to obtain more up-to date addresses. After the Census, the samples are then matched to Census documents, and non-matches are sent to a field tracing operation to establish the correct Census Day address, and then re-matched. At the end of this process, each selected person is classified as either enumerated, missed, deceased, emigrated or not traced. Final estimates are obtained by reweighting the traced records to compensate for those not traced, and adjusting the weights to arrive at known frame totals.

## Results

Figure 1 presents the overall undercoverage rates estimated from the RRC for the Censuses 1966 to 1986. (The RRC was initiated in 1961 but the results are incomplete since the sample was limited to the Census frame). As can be seen, after settling around $2 \%$ in the three Censuses from 1971 to 1981, undercoverage increased significantly in 1986. As Figure 2 shows, there are considerable variations in undercoverage across the country. British Columbia's rate is consistently above the national rate, while the rates for the Atlantic and Prairie
provinces tend to be below the national rate. (This pattern has held through all five Censuses since 1966). There are also significant variations by age and sex, as illustrated in Figure 3. In particular, rates for $20-24$ year olds are much higher than for any other group, no doubt reflecting their lower degree of attachment to a usual residence. (Again this pattern is consistent over time).

Figure 1. Undercoverage rates, Canada 1966 to 1986


Figure 2. Undercoverage rates by province
1981 \& 1986



More detailed results of the 1986 RRC are given in Statistics Canada (1990).

## Limitations

While the RRC has produced results which are plausible, reasonably compatible with other sources of information, and display consistent patterns over time, this approach does have limitations. First, there are some conceptual weaknesses:-
(i) the Census frame contains multiple enumerations;
(ii) there is, of course, no complete list of persons missed in the previous Census and the sample from this conceptual frame is simply those persons classified as missed in the previous RRC - those persons who have never been enumerated in a Census are therefore unrepresented in the RRC sample;
(iii) the immigrant frame is restricted to landed immigrants - Canadian citizens returning after an extended stay abroad are not represented in any frame;
(iv) up to now, except on an experimental basis, no sample has been drawn for the Yukon and Northwest Territories, and therefore the results for individual provinces as well as for Canada as a whole are not complete.

Being based on a sample (around 36000 persons in 1986), the results are subject to sampling error, and this imposes constraints on the amount of detail for which results can be presented. For the smaller provinces, even overall undercoverage rates have relatively high sampling errors. (While the national coefficient of variation (CV) was around $4 \%$ in 1986, provincial CVs varied from $6 \%$ for Ontario to $37 \%$ for PEI).

Perhaps more serious, however, are the potential non-sampling errors, such as non-response bias (due to the approximately $4 \%$ no-trace rate), and classification error. Very little is known about the magnitude of these errors. Conventional wisdom, however, is that non-response tends to bias undercoverage estimates downwards (since persons who are hard to find during tracing were more likely missed in the Census); on the other hand, classification errors tend to result in an upward bias (since persons can only be classified as enumerated if they
are found on a Census form, while enumerated persons may be incorrectly classified as missed if the traced address information is incorrect).

A more complete discussion of the limitations of the RRC is given in Burgess (1988).

### 2.2 The 1986 Overcoverage Study

## Methodology

Despite its limitations, the RRC has become the principal tool for evaluating the coverage of the Canadian Census. It does, however, look at only one side of the coin: that is it attempts to measure gross undercoverage rather than net coverage error. It has always been the assumption in Canada that overcoverage is small in relation to undercoverage and therefore coverage measurement efforts have been concentrated on the latter. Coverage evaluations in other countries have, however, suggested that overcoverage should perhaps not be ignored, although admittedly there are considerable differences among countries both in the way the Census itself is conducted and in the methodology of coverage evaluations. For example, there is arguably more risk of duplication in the US mail-out Census, particularly given the many coverage enhancement procedures in place. (See Ericksen, 1988). Also, the ratio of over- to undercoverage as measured in the US Post-Enumeration Survey tends to be greater because coverage is measured relative to a limited geographical area. In the RRC, on the other hand, someone is classified as missed only if they were not enumerated anywhere in the country. (A person found on a Census questionnaire is counted as enumerated even if they were not enumerated where they should have been). Nevertheless, in recent years there has been more concern over the potential level of overcoverage in the Canadian Census, and in 1986 an experimental programme to investigate duplicate enumerations was carried out.

This programme comprised four studies:-
(1) a re-interview in a sample of private dwellings to identify all other addresses at which household members may have been enumerated, followed by a search of the appropriate Census questionnaires to identify persons enumerated more than once;
(2) visits to a small sample of collective dwellings to obtain "home" or previous addresses for persons enumerated in the Census as usual residents of the institution, again with a search of the corresponding Census forms;
(3) a computerized match of a sample of Census records to identify duplications within the same Enumeration Area (EA); and
(4) a computerized match of records in a sample of contiguous EAs to identify duplication by two neighbouring enumerators. (In the event, no such duplication was detected in the small sample studied).

## Results

The results of the 1986 Overcoverage study are given in Table 1. Overall, the study identified an estimated 45600 duplications, almost half of which (22200) were identified through the re-interview component. (Note that this figure excludes those cases that could have been detected by the automated match component of the study). The automated matching within EAs detected 16300 duplicates. Despite the restricted coverage of the collective dwelling component, $15 \%$ of the duplications were found in collectives. It should be noted that the standard errors of all these estimates are quite large, and for this reason no further breakdown of results has been attempted.

Table 1
1986 Overcoverage Study - Results

| Component | Estimate | Standard Error |
| :--- | :---: | :---: |
| Private dwellings <br> -re-interview | 22200 | 6050 |
| - automated match | 16300 | 3200 |
| Collective dwellings | 7100 | 1350 |
| Total | 45600 | 6950 |

Source: Statistics Canada (1990).

## Limitations

The 1986 overcoverage studies were experimental and as such extremely limited. The sample size for the private dwelling re-interview was only 11000 households; only three types of collective dwelling were included in the study and the sample was very small ( 39 collectives and 1392 persons). The computerized matches each involved samples of 400 EAs. The match with neighbouring households involved a sample of 50 households in each EA. Furthermore, the methodology of the re-interview survey was less than ideal. Non-response was high, and there was a greater than expected number of persons enumerated in the Census but not in the re-interview for whom it was impossible to determine whether or not they had been correctly included in the Census.

While much useful experience in detecting overcoverage was gained, the results cannot be considered definitive. They should perhaps be viewed as providing a lower bound.

### 2.3 Options considered for 1991

Given the considerable experience Statistics Canada has with the Reverse Record Check and the risks involved in switching to another methodology, it was decided at a relatively early stage of planning to continue with the RRC as the cornerstone of the 1991 coverage measurement programme. It was, however, also decided not unnaturally to investigate some improvements to the RRC, notably its extension to the Yukon and Northwest Territories, an overall increase in sample size, more efficient matching and tracing procedures, and a follow-up of apparently missed persons to verify the traced address and possibly identify the reasons for being missed by the Census.

The main research effort went into investigating the extension to the territories. The problem here is threefold:-
(1) Field tracing is more difficult and expensive in these remote and sparsely populated areas;
(2) Both territories, but particularly the NWT, have a high proportion of aboriginals in the population, many of whom present matching difficulties because of variations and errors in recording names;
(3) A relatively large proportion of the territories' population consists of recent in-migrants from the provinces; no feasible sampling fraction in the provinces could yield an adequate sample of this part of the territorial population.

While very little can be done about the second of these, recent improvements in the territorial health care files offer hope of a reasonably complete and current sampling frame which would overcome the third problem and minimize the first by reducing the need for field tracing. With the co-operation of the Yukon government, a copy
of the 1986 health care file was obtained, and on a sample basis matched with the 1986 Census to determine the feasibility of measuring undercoverage in 1991.

The other major issue was how best to estimate net coverage error. The options considered were an improved programme of overcoverage studies on the one hand, and a US-style PES on the other. The advantage of the former is that it would build on the experience gained in 1986. The arguments in favour of the PES are that it is the method used by most other countries that conduct a micro/survey coverage study, its methodology has been considerably refined as a result of the research recently carried out at the US Bureau of the Census, and it measures net coverage error directly and thereby minimizes the problem of integrating results from studies with different error structures.

After careful consideration, however, it was decided that a PES would not be feasible in the 1991 Census environment. The PES approach is based on two overlapping samples of geographic blocks. An independent listing of dwellings and persons is conducted in a sample of these blocks, and a two-way match of survey respondents and persons enumerated in the Census is conducted. If necessary, a follow-up interview is carried out to obtain missing information and to reconcile non-matches. In the US, there is a comprehensive geographic structure based on the "block". This block which contains on average $40-75$ houscholds is the sampling unit for the PES. In Canada, the smallest Census area unit is the EA, with an average of 350 households. Such an area would be too large to serve as a PES sampling unit. Subsampling of areas within EAs would be very difficult and error prone, especially in rural areas, without considerable development and testing, which was not deemed feasible in time for 1991. Another problem with this approach was the integration of matching and follow-up procedures into the Census production cycle. Because of the constraints imposed by the Census processing cycle, the follow-up interviews could not be carried out until six to nine months after Census Day, and this was considered unacceptable given the number of movers and the potential for respondent recall errors. Furthermore, the cost of a US-style PES, given that we were also committed to a RRC, would have been prohibitive.

### 2.4 The 1991 coverage measurement surveys

## Objectives

The basic objective in 1991 is to obtain as accurate a national estimate of net undercoverage as possible, while at the same time achieving acceptable estimates at the province level. In terms of sampling error, an acceptable estimate would be one with a CV no greater than $25 \%$. At the same time, some of the major limitations and non-sampling errors should be dealt with. To achieve this objective, a number of improvements are being introduced.

## Undercoverage

The regular RRC sample is being increased from 36000 to 50000 persons. Most of the increase is in the smaller provinces to ensure that the CVs are within $\mathbf{2 0 \%}$. Additional frames are being sampled to represent the new groups included in the 1991 target population, namely refugee claimants, and persons staying in Canada by virtue of a student visa, work permit or Minister's permit. On the basis of the research with the Yukon Health care file, it has been decided to attempt a multiple frame extension of the RRC to both the Yukon and Northwest Territories. Samples will be selected for the territories from the regular RRC frames and matched to the territorial health care files. Non-matches will be retained in the sample, together with independent samples from the health care files. The resulting samples should provide good coverage of the territorial populations, including inter-censal migrants and persons not covered by the health plans. Improvements are being introduced into the RRC procedures, including modifications to the RRC questionnaire administered to persons traced in the field, automated matching to Census records based on exact day, month, and year of birth, and follow-up of potentially missed cases to verify address information.

## Overcoverage

The Overcoverage Study has been redesigned based on the experience gained in 1986. As in 1986, however, there will be a re-interview component, a collective dwelling component and an automated match.

For the re-interview, a two-stage sample of about 30000 households ( 15 households in 2000 EAs) will be selected. On the assumption of a gross overcoverage rate of $0.5 \%$, this is expected to yield CV's of $20-40 \%$ provincially, and $10 \%$ at the national level. The sample will be selected from the list of dwellings created by the Census Representative in the selected EAs. Names, addresses, telephone numbers and basic demographic information will be transcribed onto OS questionnaires. A telephone interview will establish whether the individual is in the target population, the usual place of residence on Census Day, and any other addresses at which he or she may have been enumerated.

Movers will be traced and there will be a field follow-up of non-respondents and households without telephones. Where there is contact and the individuals are in scope for the Census, the person will be considered correctly enumerated and no further matching is required unless alternative addresses have been reported, in which case a search of the appropriate Census forms will be carried out to identify duplication.

The collective dwelling component will be similar to that for 1986 but will cover all types of collective dwelling and will involve a larger sample (about 500 institutions and 1200 persons in non-institutional collectives).

The automated matching of persons in neighbouring households conducted in the 1986 study resulted in an estimate of 16000 duplications, despite the fact that only eight neighbouring households were checked for each selected household. The reason for this restriction was that only month and year of birth, sex and marital status were available as matching variables (names are of course not captured in the Census). In 1991, day of birth will also be captured from the Census form and this will make it feasible to match all households in the EA. We are currently carrying out developmental work to determine the most efficient matching strategy, given this more precise matching variable. We shall use the same sample of EAs for the automated matching component as for the re-interview sample, thus facilitating the integration of the two components.

Further details of the 1991 OS are given in Dibbs and Royce, (1990).

## Limitations

While we believe that this strategy is the best approach for 1991 given the time and resources available and the risks involved in implementing new methodologies without comprehensive testing, we acknowledge that it does have some limitations. Many of the limitations of the RRC mentioned earlier, such as duplications in the Census frame, the lack of coverage of returning Canadians, and the roll-over of the missed sample, remain. Sampling error will be somewhat reduced, and we hope to also reduce classification error. Extension of undercoverage measurement to the territories will eliminate another lacuna in the RRC. Overcoverage study procedures are also being refined and the sample sizes increased. These studies are also being extended to detect out-of-scope enumerations as well as duplications. Nevertheless, errors and in particular biases will remain. One weakness of this strategy is the fact that it is difficult, if not impossible, to "balance" the errors involved in the various studies. For example, misclassification is likely to bias undercoverage estimates upwards and overcoverage estimates downwards. Estimates of net undercoverage may therefore be biased upwards (although there are of course other components of error which may work in the opposite direction). The USBC is designing their PES so as to ensure that biases in under- and overcoverage estimates tend to offset one another.

## 3. DEMOGRAPHICALLY-DERIVED ESTIMATES OF COVERAGE ERROR

As mentioned in the introduction, demographic methods can also be used to derive measures of coverage error. Demographic methods essentially rely on administrative sources of data such as registration of vital events, immigration documents and possibly taxation, health and social security records. Such sources may be used either to obtain estimates of population totals at a point in time or estimates of the change in population over a period. In Canada, although some administrative records provide high quality population estimates for certain subgroups of the population - for example, Family Allowance records yield good estimates of the population up to 14 years of age (Fortier and Raby, 1989) - there is no administrative record system or systems from which estimates of the total population can be derived. For complete coverage, it is necessary to rely on the Census; administrative records are then used to derive estimates of the change in population during the interval between Censuses. These estimates are built up from the so-called "components of change"; that is, births, deaths,
immigration, emigration and internal migration. (See Statistics Canada, 1987a). Birth and death registrations are virtually complete in Canada, and immigration records are good as far as "landed immigrants" are concerned. (Up to and including the 1986 Census, only Canadian citizens and landed immigrants were included in the target population. In 1991 other immigrants with temporary but relatively long-term status in Canada, such as refugec claimants, foreign students and work permit holders will also be included). Canadians returning from an extended stay in another country, however, are not well documented. (See Fortier, 1990). Emigration estimates are less accurate than immigration statistics, being based partly on other countries' immigration records and partly on model-based inferences from family allowance and taxation records. The national estimates of population change based on these components are considered to be reasonably accurate. Below the national level, however, an additional component - internal migration - must be incorporated. Internal migration estimates are again model-based inferences from family allowance and taxation records. Provincial estimates of inter-censal population change are considerably less reliable than national estimates, particularly for those provinces with high migration rates. It should be noted that highly mobile groups such as young adult males who tend to be missed in the Census are also likely to be under-represented in family allowance and taxation records, and hence the inter-provincial migration of these groups is likely to be under-estimated.

A demographically-derived measure of coverage error can be obtained by calculating an "estimated population" in the current Census year and comparing this with the actual Census count. The estimated population is obtained by adding the components of population growth to the previous Census count. The difference between the estimated population and the Census count is known as the "error of closure", and provides an estimate of the change in net undercoverage from one Census to the next. The estimate is subject to an error equal to the error in the demographic estimate of population growth. (See also Romaniuc, 1988, for an alternative demographic estimate of coverage error).

The values for the error of closure for each of the ten Canadian provinces are shown in Table 2 for 1981 and 1986. As previously indicated, it is generally assumed that, at the national level (ie the total of the ten provinces), the crror of closure can probably be accepted as reasonably accurate. Table 2 indicates that net undercoverage decreased between 1976 and 1981, and increased in 1986. Provincial estimates are less reliable because of errors in estimating inter-provincial migration. This is particularly cvident in regard to Alberta in 1981 and Newfoundland in 1986. As previously mentioned, a significant part of the error in estimating inter-provincial migration may arise from the under-representation of young adults in both family allowance and taxation records, although there is no firm evidence to support this speculation.

Table 2
Error of Closure for 1981 and 1986

| Province | 1981 |  | 1986 |  |
| :--- | ---: | ---: | ---: | ---: |
|  |  |  |  |  |
| Nnd | 7100 | 1.25 | 11500 | 2.02 |
| PEI | -400 | -0.31 | 1300 | 1.06 |
| NS | -200 | -0.03 | 11100 | 1.28 |
| NB | -2000 | -0.28 | 11200 | 1.57 |
| Quc | -37600 | -0.58 | 87700 | 1.34 |
| Ont | 31500 | 0.37 | 66900 | 0.73 |
| Man | 8600 | 0.83 | 6100 | 0.57 |
| Sask | -5000 | -0.52 | 10700 | 1.06 |
| Alta | -53900 | -2.41 | 19400 | 0.81 |
| BC | -6000 | -0.22 | 16800 | 0.58 |
| Total | -58000 | -0.24 | 242700 | 0.96 |

Source: Statistics Canada (1987b, 1988)

* Total for the ten provinces (i.e. Canada cxcluding the Yukon and Northwest Territories). Owing to rounding error, provincial figures may not add to total.
* Demographic estimate - Census count


## 4. ESTIMATING NET UNDERCOVERAGE FOR CENSUSES PRIOR TO 1991

Given the errors involved, why the interest in demographically-derived estimates of coverage error? We shall, after all, be measuring undercoverage and overcoverage directly in 1991 and expect to obtain results sufficiently accurate to estimate net undercoverage at the provincial level. There are several reasons for considering alternative approaches such as demographically-derived estimates. First, they are available earlier than the survey-based estimates, and can provide an early warning of a potential problem. Secondly, they provide a check on the survey-based estimates. The demographically-derived estimates of undercoverage were certainly useful in 1986 in confirming the significant increase in the RRC estimate. Also, as demonstrated below, comparison of survey-based and demographically-derived estimates can point to problems in one or both sources. A further reason is that, although we expect to have reliable survey-based estimates of net undercoverage for the 1991 Census, we do not have comparable estimates for previous Censuses. For many purposes, trends in coverage error are as important as absolute levels, and it would be useful to have estimates of net undercoverage for earlier Censuses. We do, of course, have RRC estimates for Censuses back to 1966. The sample sizes are smaller than for 1991, but the main problems as far as undercoverage estimates are concerned are lack of estimates for the Yukon and Northwest Territories, and the change in Census target population in 1991. The more difficult problem, however, is to estimate overcoverage in the Censuses prior to 1991. The 1986 study was small and is not considered sufficiently reliable to be used to estimate net undercoverage in 1986, and there were no overcoverage studies for earlier Censuses. In the absence of direct survey estimates, indirect methods have to be considered.

There are in fact several possible approaches:-
a) If overcoverage in 1991 is negligible, we could assume overcoverage in earlier Censuses was also negligible.
b) Even if non-negligible, we could assume overcoverage has remained constant.
c) We could assume overcoverage is constant in relation to some other variable such as undercoverage, the Census count, or the true population.
d) We could attempt to develop a model of overcoverage in terms of known population characteristics - this approach has not, however, proved successful in the case of undercoverage.
e) We could use a demographically-derived estimate.

The latter would involve subtracting the components of growth from the estimated 1991 population (ie the 1991 Census count plus the net undercoverage as estimated by the RRC and OS). The resulting estimate of the population would then be compared with the corresponding Census count to obtain the implied net undercoverage. Such an estimate would be subject to two components of error: the error in the estimate of net undercoverage in 1991 and the error in the estimate of population growth. Although this approach yields an estimate of net undercoverage, an implied estimate of overcoverage can be obtained by subtracting net undercoverage from the RRC estimate of undercoverage for the appropriate Census.

This is equivalent to estimating the change in overcoverage by subtracting the error of closure from the change in undercoverage as estimated by the RRC. In order to assess the usefulness of this approach, let us look at the corresponding estimates of the change in overcoverage between 1976 and 1981 and between 1981 and 1986 as shown in Table 3. (The RRC estimates of undercoverage for 1976, 1981 and 1986 are given in Table 4). These estimates of overcoverage are subject to errors in the RRC estimates of the change in undercoverage as well as the error in the demographic estimate of growth. Estimates of the standard error of the change in RRC undercoverage estimates are, therefore, also given in Table 3. This component of error clearly cannot be ignored. In 1986, for example, only Newfoundland, Ontario and the ten provinces together have estimates of the change in overcoverage greater than two standard errors.

Table 3
Implied Changes in Overcoverage between 1976 \& 1981 and between 1981 \& 1986 (together with approximate standard errors for the change in RRC estimates of undercoverage)

| Province | 1976-1981 |  |  | 1981-1986 |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Nfld | -3300 |  | ( 3 400) | -9800 | \# (3 200) |
| PEI | 1500 |  | 7 600) | -200 | (1 300) |
| NS | 2000 |  | (4 100) | 3500 | (4500) |
| NB | -100 |  | (3 400) | -3400 | (3500) |
| Que | -26900 |  | $(21300)$ | - 6800 | (24500) |
| Ont | 12300 |  | (19000) | 83000 | * (22 100) |
| Man | -9 500 |  | $(5000)$ | 8100 | (5700) |
| Sask | 2300 |  | (2 400) | 5600 | ( 5 300) |
| Alta | 84400 | \# | $(9800)$ | -10600 | (11900) |
| BC | 1560 |  | (12 300) | 29800 | (15 500) |
| Total* | 786 |  | (33 700) | 99300 | * (40900) |

* Total for the ten provinces (i.e. Canada excluding the Yukon and Northwest Territories). Owing to rounding error, provincial figures may not add to total.

Statistically significant at the $5 \%$ level.

Table 4
RRC Estimates of Undercoverage In the 1976, 1981 and 1986 Censuses (standard errors in parenthesis)

| Province | 1976 |  | 1981 |  | 1986 |  |
| :--- | ---: | ---: | ---: | ---: | ---: | ---: |
| Nfld | 6200 | $(2200)$ | 10000 | $(2600)$ | 11700 | $(1900)$ |
| PEI | 400 | $(300)$ | 1500 | $(700)$ | 2800 | $(1800)$ |
| NS | 7200 | $(2900)$ | 9000 | $(2900)$ | 23600 | $(3500)$ |
| NB | 15000 | $(2600)$ | 12900 | $(2200)$ | 20700 | $(2700)$ |
| Que | 189700 | $(16200)$ | 125200 | $(13800)$ | 206100 | $(20200)$ |
| Ont | 127200 | $(14200)$ | 171000 | $(12600)$ | 320900 | $(18200)$ |
| Man | 11800 | $(3500)$ | 10200 | $(3600)$ | 24400 | $(4400)$ |
| Sask | 12400 | $(3200)$ | 9700 | $(3700)$ | 26000 | $(3800)$ |
| Alta | 27800 | $(4900)$ | 58300 | $(8500)$ | 67100 | $(8300)$ |
| BC | 79800 | $(7800)$ | 89400 | $(9500)$ | 136000 | $(12300)$ |
| Total | 476700 | $(23900)$ | 497300 | $(23800)$ | 839300 | $(33300)$ |

* Total for the ten provinces (i.e. Canada excluding the Yukon and Northwest Territories). Owing to rounding error, provincial figures may not add to total.

Despite the problems with these estimators, particularly at the provincial level, these results are instructive and can help in the evaluation of the survey-based estimates of coverage error. For example, if we can consider the error in the components of growth at the national level to be negligible, net undercoverage decreased by 58,000 between 1976 and 1981, and increased by 242,700 between 1981 and 1986. The RRC estimates of undercoverage were $476,700(2.04 \%)$ in $1976,497,300(2.01 \%)$ in 1981 and $839,300((3.21 \%)$ in 1986 . Thus, the implied
increases in overcoverage were 78600 between 1976 and 1981, and 99,300 between 1981 and 1986. Since overcoverage cannot be less than zero, this implies an overcoverage of at least $177900(0.7 \%$ of the Census count) in 1986. The experimental overcoverage study carried out in 1986 was only able to detect 45600 persons ( $0.2 \%$ ) double-counted (with a standard error of 6950 ). Either overcoverage was significantly underestimated by the study, or there is significant error at the national level in the components of change, or the increase in undercoverage was over-estimated by the RRC. (It should be noted, however, that the $95 \%$ confidence interval for the increase in undercoverage between 1976 and 1986 is $362600 \pm 82,000$. If we take the boltom of the range, 280600 , the implied increase in overcoverage between 1976 and 1986 would only be 95900 . Between 1981 and 1986, the $95 \%$ confidence interval for the increase in undercoverage is $342,000 \pm 81800$, the bottom of the range implying an increase in overcoverage of only 17500 ).

At the provincial level, only three of the estimates in Table 3 are statistically significant. Again, however, they imply either relatively large values for overcoverage, or substantial errors in the estimates of population growth. As we have seen, for Alberta in 1981 and for Newfoundland in 1986, errors in estimating net inter-provincial migration probably account for the apparently high magnitude of the estimated change in overcoverage.

Hence, it would appear that the implicit estimates of the provincial changes in overcoverage are not reliable. Assuming the RRC estimates of undercoverage in Censuses prior to 1991 are to stand, the demographic approach to measuring net undercoverage in these Censuses is therefore not viable at the province level. It may, however, be possible to use this approach at the national level, and disaggregate the estimate of overcoverage among the provinces in proportion to their Census count or estimated undercoverage (as in option c) above).

## 5. CONCLUSIONS

The intention is to have for the first time in Canada reliable estimates of net coverage error. We believe that, if everything goes according to plan, the programme we have put in place will achieve this goal. The question is what will we do with these estimates? It is well known that the adjustment of the Census for net coverage error is a major issue in the US. Although less politicized in Canada, there is here too a debate concerning the wisdom of adjusting Census counts. However, given the fact that we are far from having perfected our procedures for measuring net coverage error, certainly below the provincial level, and given the time required to complete and evaluate the coverage studies, Statistics Canada does not intend to adjust the 1991 Census counts on the basis of the estimates of net coverage error described above. Adjustment may result in better estimates for some purposes but not others. We prefer to release the Census counts resulting from traditional procedures and to make available to users the information we obtain on coverage errors so that they can make their own informed interpretation of Census data. Having said that, we are examining the possibility of incorporating information on net coverage error into the official population estimates. At present, these estimates are based on the Census counts, and in the periods between Censuses rely on the components of growth described above. Research is underway to explore the feasiblity of removing the constraint that the population estimate for June of a Census year should be the published Census count. Instead, an allowance might be made for coverage error. The implications of such a change are being investigated, together with the question of the geographic level at which such estimates would be reliable. A final decision will have to await the results of the coverage measurement studies and an assessment of their quality.

## ACKNOWLEDGEMENTS

I would like to thank the following colleagues at Statistics Canada for their helpful comments on this paper: Céline Fortier, Dr. Edward Pryor, Ronald Raby, Don Royce, and Ravi Verma.

## REFERENCES

Burgess, R. (1988). Evaluation of Reverse Record Check Estimates of Undercoverage in the Canadian Census of Population, Survey Methodology, 14, 2, 137-156.

Dibbs, R., and Royce, D. (1990). Measuring Overcoverage in the 1991 Census of Canada, presented at the meeting of the American Statistical Association, Government Statistics Section, Anaheim, Ca.

Ericksen, E. (1988). Bayes Estimates of Population Undercount for Local Areas, Proceedings of the Section on Survey Research Methods, American Statistical Association.

Fortier, C. (1990). Canadiens de retour: nombre (1981-1989) et caractéristiques (1988 et 1989), internal report, Demography Division, Statistics Canada.

Fortier, C., and Raby, R. (1989). Evaluation de la qualité des données par âge et sexe du recensement canadien de 1986, Cahiers québécois de démographie, 18, 2.

Romaniuc, A. (1988). A Demographic Approach to the Evaluation of the 1986 Census and the Estimates of Canada's Population, Survey Methodology, 14, 2, 157-171.

Statistics Canada (1987a). Population Estimation Methods, Catalogue No. 91-528.
Statistics Canada (1987b). Postcensal Annual Estimates of Population by Marital Status, Age, Sex and Components of Growth for Canada, Provinces and Territories, June 1, 1986, Vol. 4 Fourth Issue, Catalogue No. 91-210.

Statistics Canada (1988). Postcensal Annual Estimates of Population by Marital Status, Age, Sex and Components of Growth for Canada, Provinces and Territorics, June 1, 1988, Vol. 6 Sixth Issue, Catalogue No. 91-210.

Statistics Canada (1990). User's Guide to the Quality of 1986 Census Data: Coverage, Catalogue No. 99-135.

# SMALL-AREA PREDICTION OF UNDERCOUNT USING THE GENERAL LINEAR MODEL 

N. Cressie ${ }^{1}$


#### Abstract

Suppose the nation is divided into $n$ small areas, each of which has an (unobservable) undercount. Assume the undercount variable is a linear combination of explanatory variables, plus a zero mean error vector whose variance matrix $\Gamma(Y)$ depends on parameters $Y$. However, undercount data have an additional error component, with variance matrix $\Delta$, that is known from sampling (and nonsampling) considerations. From known $\Gamma(\underline{Y})$ and $\Delta$, optimal linear small-area predictors can be derived. In this paper, a spatial model for $\Gamma(\gamma)$ is fit to data from the 1980 U.S. Census and the 1980 Post Enumeration Survey, allowing the small-area predictors to be calculated.


KEY WORDS: Adjustment factors; Heteroskedasticity; Maximum likelihood estimation; Mean squared prediction errors; Spatial dependence.

## 1. INTRODUCTION

This paper will first review results on small-area prediction (of undercount), based on a general linear model posited in Cressie (1990). The model is general enough so that both heteroskedasticity and spatial dependence can be considered as special cases. The effect of introducing spatial dependence, for prediction of 1980 U. S. undercount, will be investigated.

Currently, there are two closely related approaches to predicting undercount. One is a stratification approach (Cressie, 1988, 1989), and the other is a regression approach (Ericksen and Kadane, 1985; National Academy of Sciences, 1985; Diffendal, 1988; Cressie, 1990). Suppose that the nation is divided into $i=1, \ldots, n$ areas (e.g., states, including Washington, D.C.). The goal is to predict undercount in each of these areas based on a sample obtained from dual-system estimation (e.g., Wolter, 1986). Assume further that each area is divided into (as many as) J strata, where it is thought that the undercount "mechanism" is homogeneous within strata. Let the $(j, i)$-th subarea have census count $C_{j i}$ and true count $T_{j i}$; then the proportion undercounted is,

$$
\begin{equation*}
U_{f i} \equiv\left(T_{j i}-C_{j i}\right) / T_{j i}, \tag{1.1}
\end{equation*}
$$

and the adjustment factor is defined as,

$$
\begin{equation*}
F_{j f}=T_{j i} / C_{j i} . \tag{1.2}
\end{equation*}
$$

Clearly, $F_{j i}=1 /\left[1-U_{j i}\right)$, showing one quantity to be a monotonic function of the other. It is usually the undercount variable that is quoted (as a percentage) in press releases and final reports, but for model-consistency reasons the statistical analysis is often carried out on the adjustment factor.

[^16]Define:

$$
\begin{equation*}
T_{i} \equiv \sum_{j=1}^{j} T_{j n} ; C_{i} \equiv \sum_{j=1}^{j} C_{j} ; F_{i} \equiv T_{i} / C_{i} ; i=1, \ldots, n . \tag{1.3}
\end{equation*}
$$

Notice that $T_{1}=\sum_{j=1}^{J} F_{\mu} C_{H}$, which explains why $F_{j 1}$ is called an adjustment factor. Suppose that statistical inference based on post-enumeration survey (PES) data (or any other source of data) produce predictors $\left\{F_{j l}^{\text {md }}\right\}$; then a predictor of the $i$-th area's true population is

$$
\begin{equation*}
T_{i}^{p r d} \equiv \sum_{j=1}^{J} F_{j i}^{p r d} C_{j i} \tag{1.4}
\end{equation*}
$$

The stratification approach assumes that $\left\{F_{j i}: i=1, \ldots n\right\}$ are random variables each with mean $\mu_{j}$, which is a parameter to be estimated; $j=1, \ldots, J$. The usual regression approach assumes that $F_{i}$ defined by (1.3) is a random variable with mean, $\sum_{k=1}^{p} \beta_{k} z_{k, i} ; i=1, \ldots, n$, where $\left\{z_{k i}: k=1, \ldots, p\right\}$ are given explanatory variables and $\left\{\beta_{k}: k=1, \ldots, p\right\}$ are parameters to be estimated. The truly general model, of which each approach is a special case, assumes that $F_{j 1}$ has mean $\sum_{k=1}^{p_{j}} \beta_{k, j} z_{k, j 1} ; j=1, \ldots, J, i=1, \ldots, n$. That is, each stratum is allowed to have its own regression model. For example, blacks, whites, nonblack Hispanics, and so on have different undercount characteristics that should require different regression models. The special case $p_{j}=1$ and $z_{1, j l}=1$ yields the stratification approach, and the special case $p_{j}=p$ and $\beta_{k j}=\beta_{k}$ (alternatively, the special case $j=1$ ) yields the regression approach.

An important part of the (stratification) model proposed by Cressie $(1988,1989)$ is the assumption,

$$
\begin{equation*}
\operatorname{var}\left(F_{j i}\right)=\tau_{j}^{2} / C_{j i} . \tag{1.5}
\end{equation*}
$$

for which both a Bayesian and a frequentist justification (for weighting inversely proportional to census counts) was given. Intuitively, (1.5) is sensible, since adjustment factors are likely to be less variable from more populous areas. The consequence of $(1.5)$ on variances of the regression model is,

$$
\begin{equation*}
\operatorname{var}\left(F_{i}\right)=v_{i}^{2}=\left\{\sum_{j=1}^{J}\left(C_{k} / C_{i}\right) \varepsilon_{j}^{2}\right\} / C_{i} \tag{1.6}
\end{equation*}
$$

Provided $\sum_{j=1}^{J}\left(C_{j} / C_{i}\right) \tau_{j}^{2}$ does not vary greatly over the $n$ areas, it follows that, approximately,

$$
\begin{equation*}
\operatorname{var}\left(F_{i}\right)=\tau^{2} / C_{i} . \tag{1.7}
\end{equation*}
$$

Cressie (1990) shows that, for the 1980 data, the difference between assuming (1.6) or (1.7) is of very small consequence. Of more consequence, is the difference between (1.7) and the homoskedasticity assumption,

$$
\begin{equation*}
\operatorname{var}\left(F_{i}\right)=\tau^{2} \tag{1.8}
\end{equation*}
$$

Notice that, in (1.6), (1.7), and (1.8), there is an implicit assumption that $\operatorname{cov}\left(F_{i}, F_{i^{\prime}}\right)=0$, for $i * i$. However, it is possible to make considerable progress by assuming only that $\operatorname{var}\left\{\left(F_{1}, \ldots, F_{n}\right)^{\prime}\right\}$ is nonnegative-definite.

Section 2 reviews the results obtained by Cressic (1990). Section 3 applies these results to a spatial linear model for $\left(F_{i}, \ldots, F_{n}\right)^{\prime}$, using the method of maximum likelihood to estimate unknown parameters. Section 4 gives
optimal predictors based on a (heteroskedastic) spatial model for 1980 undercount data. Section 5 contains some concluding remarks.

## 2. EMPIRICAL BAYES PREDICTION WITH THE GENERAL LINEAR MODEL

I consider the true population in any state-stratum combination to be unknown. After observing the corresponding census population, the uncertainties about the true population are updated. Therefore, the models I shall build, and the subsequent inferences, are conditional on the observed census counts.

In the rest of this paper, I shall adopt the regression approach to modeling undercount, referred to in Section 1. Contributors to that approach, and its understanding, are Ericksen and Kadane (1985), Freedman and Navidi (1986), Diffendal (1988), Ericksen, Kadane, and Tukey (1989), and Cressie (1990). The general linear model adopted here includes all these authors' models as special cases.

Recall that the adjusiment factor for the $i$-th small area is,

$$
\begin{equation*}
F_{i}=T_{i} / C_{i} \tag{2.1}
\end{equation*}
$$

Then the undercount for that area is,

$$
\begin{align*}
U_{i} & =\left(T_{i}-C_{i}\right) / T_{i}  \tag{2.2}\\
& =1-F_{i}^{-1} ; i=1, \ldots, n . \tag{2.3}
\end{align*}
$$

Since the undercount variable is a monotonic one-to-one function of the adjustment-factor variable, mathematically it matters little which is modeled. Statistically, there are advantages to modeling and predicting the adjustment factors.

Let $\underset{\sim}{F} \equiv\left(F_{1}, \ldots, F_{n}\right)^{\prime}$ be the nxl vector of adjustment factors to be predicted. Notice that if two small areas $i$ and $i^{\prime}$ are aggregated into $i v i^{\prime}$, then,

$$
\begin{equation*}
F_{i U_{i}}=\left\{C_{i} F_{i}+C_{i^{\prime}} F_{i^{\prime}}\right\} /\left\{C_{i}+C_{i^{\prime}}\right\} \tag{2.4}
\end{equation*}
$$

Thus, linear models are preserved under aggregation, and (less importantly) so are Gaussian distributional assumptions.

The notation, $\underset{\sim}{\sim} \sim \operatorname{Gou}(\underset{\sim}{\mu}, \Sigma)$, means that the column vector $\underset{\sim}{\underset{\sim}{Z}}$ has a multivariate Gaussian distribution with mean $\underset{\sim}{\mu}$ and variance matrix $\Sigma$. Assume that the (prior) distribution of adjustment factors is given by,

$$
\begin{equation*}
\underset{\sim}{F} \sim \operatorname{Gau}(X \hat{\beta}, \Gamma(\gamma)), \tag{2.5}
\end{equation*}
$$

where $N$ is an nxp matrix of explanatory variables, $\beta$ is a px1 vector of unknown coefficients, and $\Gamma(\gamma)$ is an nxn variance matrix depending on a kxl vector of unknown parameters $\boldsymbol{Y}$. Examples of explanatory variables might be percentage minorities, percentage of households rented, measure of poverty, etc.

Since $\underset{\sim}{F}$ is not available as data, imperfect observations on adjustment factors have to be used. For example, dual-system estimators obtained from a post-enumeration survey (PES) yield the observation vector $\underset{\sim}{Y}$. Assume that, conditional on $\underset{\sim}{\underset{\sim}{F}}$,

$$
\begin{equation*}
\underset{\sim}{Y} \mid \underset{\sim}{F} \sim \operatorname{Gau}(\underset{\sim}{F}, \Delta), \tag{2.6}
\end{equation*}
$$

where $\Delta$ is an $n \times n$ matrix of known sampling variances and covariances.
Let $\underset{\sim}{p}(\underset{\sim}{Y})$ and $\underset{\sim}{p}(\underset{\sim}{Y})$ be two predictors of $\underset{\sim}{F}$. Then $\underset{\sim}{p}(\underset{\sim}{Y})$ is said to be as $\operatorname{good}$ as $\underset{\sim}{p}(\underset{\sim}{Y})$ if,

$$
E\left\{\left(\underset{\sim}{F}-{\underset{\sim}{p}}_{2}(\underset{\sim}{Y})\right)\left(\underset{\sim}{F}-{\underset{\sim}{p}}_{2}(\underset{\sim}{Y})\right)^{\prime}\right\}-E\left(\left(\underset{\sim}{F}-{\underset{\sim}{1}}_{1}(\underset{\sim}{Y})\right)\left(\underset{\sim}{F}-{\underset{\sim}{p}}_{1}(\underset{\sim}{Y})\right)^{\prime}\right\}
$$

is nonnegative-definite. Now, the optimal predictor is $\underset{\sim}{F} \underset{\sim}{F} \mid \underset{\sim}{Y}$ (Cressie, 1990), which for the Gaussian model is,

$$
{\underset{\sim}{p}}^{*}(\underset{\sim}{Y})=\left(\Gamma(\underset{\sim}{\gamma})(\Delta+\Gamma(\underset{\sim}{\gamma}))^{-1} \mid \underset{\sim}{Y}+\left(I-\Gamma(\underset{\sim}{\gamma})(\Delta+\Gamma(\underset{\sim}{\gamma}))^{-1}\right) X \underset{\beta}{\beta},\right.
$$

and,

$$
\begin{equation*}
E\left\{\left(\underset{\sim}{F}-{\underset{\sim}{p}}^{0}(\underset{\sim}{Y})\right)\left(\underset{\sim}{F}-{\underset{\sim}{p}}^{0}(\underset{\sim}{Y})\right)^{\prime}\right\}=\left(I-\Gamma(\underset{\sim}{\gamma})(\Delta+\Gamma(\underset{\sim}{\gamma}))^{-1}\right\} \Gamma(\underset{\sim}{\gamma}) . \tag{2.7}
\end{equation*}
$$

From a Bayesian point of view, $\underset{\sim}{\boldsymbol{p}} \underset{\sim}{\bullet}(\underset{\sim}{\boldsymbol{Y}})$ is a Bayes estimator of $\underset{\sim}{\boldsymbol{F}}$, for the loss matrix $L \underset{\sim}{\boldsymbol{F}}, \underset{\sim}{\boldsymbol{p}})=(\underset{\sim}{\boldsymbol{F}}-\underset{\sim}{\boldsymbol{p}})(\underset{\sim}{\boldsymbol{F}}-\underset{\sim}{\boldsymbol{p}})^{\prime}$.

Assume now that $\underset{\sim}{\beta}$ is unknown, but that $\underset{\sim}{\gamma}$ is known. Then the optimal linear unbiased predictor of $\underset{\sim}{F}$ is (Cressie, 1990),

$$
\begin{align*}
& \underset{\sim}{p}(\underset{\sim}{\eta})=\left[\left\{\Gamma(\underset{\gamma}{\gamma})(\Delta+\Gamma(\underline{\gamma}))^{-1}\right\}+\left(I-\Gamma(\underset{\gamma}{ })(\Delta+\Gamma(\underline{\gamma}))^{-1}\right\}\right. \\
& X\left\{X^{\prime}\left(\boldsymbol{\gamma}(\Delta+\Gamma(\underset{\sim}{\gamma}))^{-1} X\right\}^{-1} X^{\prime}(\Delta+\tilde{\Gamma}(\underset{\sim}{\gamma}))^{-1}\right] \underset{\sim}{\mathcal{Y}} . \tag{2.8}
\end{align*}
$$

When $\underset{\sim}{\boldsymbol{\gamma}}$ is unknown (which it usually is), replace it with $\underset{\sim}{\hat{\gamma}}$, an estimator based on $\underset{\sim}{\boldsymbol{Y}}$. Then (2.8) (with estimator $\underset{\sim}{\hat{\gamma}}$ ) has been called an empirical Bayes estimator (Ericksen and Kadane, 1985).

In Cressic (1990), two approaches to the estimation of $\underset{\sim}{\underset{\sim}{y}}$ were taken: maximum likelihood and method-of-moments. For spatial models, method-of-moments estimators may not be readily available. In the rest of this paper I shall use the method of (Gaussian) maximum likelihood for estimation of $\underset{\sim}{\gamma}$.

## 3. SPATIAL MODELS OF UNDERCOUNT

In this section, a spatial model for the distribution of the (prior) errors, $\underset{\sim}{F}-X \underset{\sim}{x}$, is assumed. (Spatial dependence may be defined according to a Euclidean distance, or a city-block distance, or a "sociologic/ethnographic" distance, etc.). Introducing spatial dependence through one or two unknown parameters is low-premium insurance against the concern (expressed by Freedman and Navidi, 1986) that an important explanatory variable may have been missed, or that the linear functional relationship is in fact more complicated.

In 1934, the statistician F. F. Stephan wrote "...Data of geographic units are tied together, like bunches of grapes, not separate, like balls in an urn. Of course, mere contiguity in time and space does not of itself indicate lack of independence between units in a relevant variable or attribute, but in dealing with social data, we know that by virtue of their very social character, persons, groups and their characteristics are interrelated and not independent. Sampling error formulas may yet be developed which are applicable to these data, but until then
the older formulas must be used with great caution. Likewise, other statistical measures must be carefully scrutinized when applied to these data..." (Stephan, 1934).

### 3.1 Spatial linear model

In the model given by (2.5) and (2.6), spatial dependence can arise in either component of the variance matrix. First, in (2.5), suppose that

$$
\begin{equation*}
\underset{\sim}{8} \equiv \underset{\sim}{F}-X \underset{\sim}{\beta}, \tag{3.1}
\end{equation*}
$$

forms a spatially-dependent Gaussian process. For example, consider the conditionally-specified Gaussian Markov random field (Besag, 1974). In this case, the conditional distribution of $\delta_{i}$ given $\left\{\delta_{j}: j * i\right\}$ is Gaussian and depends only on $\left\{\delta_{j}: j \in N_{i}\right\}$, where $N_{i}$ represents a set of "neighboring" areas of the $i$-th area; $i=1, \ldots, n$. Specifically,

$$
\begin{equation*}
\delta_{i} \mid\left\{\delta_{j} ; j=i\right\} \sim \operatorname{Gau}\left(\sum_{j \in N_{t}} q_{\psi} \delta_{j} \tau_{i}^{2}\right) ; i=1, \ldots, n, \tag{3.2}
\end{equation*}
$$

which implies that,

$$
\begin{equation*}
\underset{\sim}{F} \sim \operatorname{Gau}\left(X \underset{\sim}{\beta},(I-Q)^{-1} M\right), \tag{3.3}
\end{equation*}
$$

where $M=\operatorname{diag}\left(\tau_{1}^{2} \ldots, \tau_{n}^{2}\right)$, and $Q \equiv\left(q_{i j}\right)$ with $q_{i j}=0$ whenever $j \& N_{i}$ (including $q_{i j}=0$ ). Of course, the matrix I-Q must be invertible, and $q_{i j} \tau_{j}^{2}=q_{i j} \tau_{j}^{2}(i<j)$ guarantees that $(I-Q)^{-1} M$ is symmetric; it is further required that $(I-Q)^{-1} M$ be positive-definite. These conditions define the parameter space for any conditionally-specified Gaussian spatial model. For example, let $Q=\lambda H$ and $M=\tau^{2} D$, where $H$ and $D$ are known. Then $\Gamma(\underset{\sim}{\boldsymbol{\gamma}})=\tau^{2}(I-\lambda H)^{-1} D$, where $\underset{\sim}{\boldsymbol{y}}=\left(\lambda, \tau^{2}\right)^{\prime}$; see, e.g., Cressic and Chan (1989). For Markov random fields, $\underset{\sim}{Y}$ appears nonlinearly in $\Gamma(\underset{\sim}{\boldsymbol{Y}}$ ), and can be estimated straightforwardly by maximum likelihood.

Another example would be to use a model from geostatistics, such as $\Gamma(\underset{\sim}{\gamma})=\left\{\gamma_{1} I+\gamma_{2}\left(\exp \left(-d_{v}\right)\right)\right\} M$, where $\left(\exp \left(-d_{i j}\right)\right)$ represents a matrix whose $(i, j)$-th element is $\exp \left(-d_{i j}\right)$, and $d_{\psi}$ is a "distance" (not necessarily Euclidean) between the $i$-th small area and the $j$-th small area. Here, the two parameters appear linearly and either maximum likelihood or method-of-moments estimation for $\underset{\sim}{\boldsymbol{Y}}$ would be straightforward.

Other geostatistical models, such as $\Gamma(\underset{\sim}{\gamma})=\left\{\gamma_{1} I+\gamma_{2}\left(\gamma_{3}^{d}\right)\right\} M$, involve $\underset{\sim}{\gamma}$ nonlinearly, for which maximum likelihood estimation is straightforward.

Notice that the sociologist's/ethnographer's map of the small areas may be quite different from the geographer's map. For example, Philadelphia, New York City, Detroit, Chicago, and Los Angeles may be neighbors for the purpose of studying undercount in small areas that include states, cities, and rest of states. Also, New York City and the rest of New York State may not be neighbors. Demographers, sociologists, and ethnographers could together construct a small-area neighborhood structure based on their knowledge of the undercount mechanisms (e.g., see Hainer et al., 1988, for a study of some of these mechanisms).

The second source of spatial dependence can be in (2.6), through $\boldsymbol{\Delta}$. In dual-system estimation of the true $\underset{\sim}{F}$, certain census blocks within the small areas are sampled. Calculation of sampling variances and covariances yields $\Delta$, which may (or may not) be modeled and estimated using the geographic location of the blocks; it is here that geostatistical methods would be particularly appropriate. The U.S. Census Bureau's TIGER system is potentially a powerful tool for any spatial modeling.

Apart from the extra generality and elegance of a spatial linear model, it has a pragmatic role. It addresses the concern that an important explanatory variable in $E(\underset{\sim}{F})=X \beta$ may have been missed, or that the linear functional relationship is in fact more complicated. Ideally, all important variables are chosen, or (less ideally) proxies for them appear in the regression relation. These variables, missed variables, and the dependent variable (adjustment factor) are all varying spatially. The benefit obtained from fitting one or two spatial-dependence parameters is considerabic. Local bias (in individual small areas) is reduced and more efficient estimators of regression parameters $\underset{\sim}{\beta}$ are obtained (e.g., Dorcian, 1980; Dow et al., 1982; Anselin and Griffith, 1988; Dubin, 1988).

### 3.2 Estimation of variance-matrix parameters

Suppose,

$$
\begin{equation*}
\underset{\sim}{Y}-\operatorname{Gau}(X \underset{\sim}{\beta}, \Sigma(\underset{\sim}{Y})) \tag{3.4}
\end{equation*}
$$

where recall that $\underset{\sim}{Y} \equiv\left(Y_{1}, \ldots, Y_{n}\right)^{\prime}$ is an $n \times 1$ vector of data, $X$ is an nxp matrix of explanatory variables, $\underset{\sim}{\beta} \equiv\left(\beta_{1}, \ldots, \beta_{p}\right)$ is a $p x 1$ vector of unknown regression coefficients, $\Sigma(\underset{\sim}{\gamma})$ is an non symmetric positive-definite matrix and $\underset{\sim}{Y} \equiv\left(\gamma_{1}, \ldots, Y_{k}\right)^{\prime}$ is a $k x 1$ vector of unknown variance-matrix parameters. Thus, the negative loglikelihood of $\underset{\sim}{\beta}$ and $\underset{\sim}{Y}$ is:

Minimization of this function with respect to $\underset{\sim}{\beta}$ and $\underset{\sim}{\boldsymbol{\gamma}}$ yields maximum likelihood (m.l.) estimates $\underset{\sim}{\hat{\beta}}$ and $\underset{\sim}{\hat{\gamma}}$. In practice, the solutions usually have to be found using numerical optimization; the most common algorithm employed is iterative Gauss-Newton (or scoring).

Define the px1 vector:

$$
\begin{equation*}
\underset{\sim}{L_{\beta}} \overline{=} X \Sigma(\underset{\sim}{\gamma})^{-1} X \underset{\sim}{\beta}-X \Sigma(\underset{\sim}{\gamma})^{-1} \underset{\sim}{\gamma} \tag{3.6}
\end{equation*}
$$

and the kxI vector $\underset{\sim}{\underset{\gamma}{\gamma}}$ to have $i$ th element:

$$
\begin{equation*}
\left(\underset{\sim}{L_{\gamma}}\right)_{1} \equiv(1 / 2) \operatorname{tr}\left(\Sigma(\underset{\sim}{\gamma})^{-1} \Sigma_{\sim}^{\gamma}(\underset{\sim}{\gamma})+(1 / 2) \underset{\sim}{8} / \Sigma(\underset{\sim}{\gamma}) \underset{\sim}{\delta}\right. \tag{3.7}
\end{equation*}
$$

where,

$$
\begin{gather*}
\underset{\sim}{\delta} \equiv \underset{\sim}{Y}-X \beta  \tag{3.8}\\
\Sigma_{1}(\underset{\sim}{\gamma}) \equiv \partial \Sigma(\underset{\sim}{\gamma}) / \partial \gamma_{i}  \tag{3.9}\\
\Sigma^{\prime}(\underset{\sim}{\gamma}) \equiv \partial \Sigma(\underset{\sim}{\gamma})^{-1} / \partial \gamma_{i}=-\Sigma(\underset{\sim}{\gamma})^{-1} \Sigma_{i}(\underset{\sim}{\gamma}) \Sigma(\underset{\sim}{\gamma})^{-1} ; i=1, \ldots, k . \tag{3.10}
\end{gather*}
$$

(The derivative of a matrix is the matrix of elementwise derivatives, and the matrix operator $\operatorname{tr}(G)$, called the trace, sums the diagonal elements of the square matrix $G$.) Define the pxp matrix:

$$
\begin{equation*}
J_{B}=X \Sigma(\underset{\sim}{\gamma})^{-1} X, \tag{3.11}
\end{equation*}
$$

and the kxk matrix $J_{\gamma}$ to have $(i, j)$-th element:

$$
\begin{equation*}
\left(J_{y}\right)_{i j} \equiv(1 / 2) t_{i j} \equiv(1 / 2) \operatorname{tr}\left(\Sigma_{\sim}\left(\underset{\gamma}{-1} \Sigma_{l}(\gamma) \Sigma(\underset{\sim}{\gamma})^{-1} \Sigma_{j}(\gamma)\right) .\right. \tag{3.12}
\end{equation*}
$$

Then, the scoring algorithm is:
where

$$
A=\left[\begin{array}{ll}
J_{\beta} & O  \tag{3.14}\\
O & J_{\gamma}
\end{array}\right]
$$

and $A^{(n}, \underset{\sim}{L_{\beta}^{(1)}}$, and $\underset{\sim}{{\underset{\sim}{\gamma}}^{(1)}}$ denote the quantities $\boldsymbol{A}, \underset{\sim}{\boldsymbol{L}}$, and $\underset{\sim}{L_{\gamma}}$, respectively, evaluated at $\underset{\sim}{\beta}={\underset{\sim}{\beta}}^{(1)}$ and $\underset{\sim}{\boldsymbol{Y}}={\underset{\sim}{\gamma}}^{(1)}$. Equation (3.13) reduces to:

$$
\begin{align*}
& {\underset{\sim}{\beta}}^{(0)}=\left(X^{\prime} \Sigma\left({\underset{\sim}{10}}^{(0)}\right)^{-1} X\right)^{-1} X^{\prime} \Sigma\left(\underline{\sim}^{(0}\right)^{-1} \underset{\sim}{\boldsymbol{Y}}  \tag{3.15}\\
& \underline{\sim}^{(1+1)}={\underset{\sim}{r}}^{(0)}-\left(J_{\gamma}^{(0)}\right)^{-1}{\underset{\sim}{r}}^{(\theta)}, \tag{3.16}
\end{align*}
$$

where $J_{\gamma}^{(n)}$ is $J_{\gamma}$ evaluated at $\underset{\sim}{\beta}={\underset{ }{\beta}}^{(1)}$ and $\underset{\sim}{\gamma}={\underset{\sim}{\gamma}}^{(1)}$. Further details on implementation of these algorithms can be found in Mardia and Marshall (1984) and Kitanidis and Lane (1985).

In the case where there is only one spatial dependence parameter, it is easy to obtain a direct assessment of the maximum likelihood estimate of that parameter from its profile likelihood. For example, consider the spatial-dependence model (3.3), where

$$
\begin{equation*}
Q=\lambda H \text { and } M=\tau^{2} D \tag{3.17}
\end{equation*}
$$

for known H and D . Then, in (3.4),

$$
\begin{equation*}
\Sigma(\underline{\gamma})=\tau^{2}(I-\lambda H)^{-1} D+\Delta ; \underset{\sim}{\gamma}=\left(\lambda, \tau^{2}\right)^{t} \tag{3.18}
\end{equation*}
$$

Assume, for the moment, that $\lambda$ is known. Maximizing (3.5) over $\beta$ and $\tau^{2}$ yields maximum likelihood estimates, $\hat{\mathbb{R}}(\lambda)$ and $\hat{\tau}^{2}(\lambda)$. Substituting these values back into (3.5) results in the profile negative loglikelihood for spatial dependence parameter $\lambda$ :

$$
\begin{equation*}
L^{\prime}(\lambda) \equiv L\left(\hat{\varrho}(\lambda), \hat{\tau}^{2}(\lambda), \lambda\right), \tag{3.19}
\end{equation*}
$$

where $\left.L \beta, \tau^{2}, \lambda\right)$ is given by (3.5) and (3.18).

The maximum likelihood estimate $\hat{\lambda}$ is obtained by minimizing $L^{\prime}$ with respect to $\lambda$. Moreover, an approximate (based on asymptotic considerations) $100(1-\alpha) \%$ confidence region for $\lambda$, due to Whittle (1954), is given by:

$$
\begin{equation*}
\left(\lambda: L^{*}(\lambda) \leq L^{*}(\hat{\lambda})+(n /(n-p-2)) x_{1}^{2}(\alpha) / 2\right\}, \tag{3.20}
\end{equation*}
$$

where $x_{1}^{2}(\alpha)$ is the upper $100(1-\alpha) \%$ point of the chi-squared distribution on one degree of freedom.
Typically, as $n \rightarrow \infty$, maximum likelihood estimators are consistent, asymptotically Gaussian, and asymptotically efficient. The asymptotic variance matrix of $n^{1 / 2}\left(\hat{\beta}^{\prime}, \hat{\lambda}^{\prime}\right)^{\prime}$ is $\hat{A}^{-1}$, where A is given by (3.14). Notice that $\hat{\hat{\beta}}$ and $\hat{\sim}$ are asymptotically independent. In practice, standard errors are obtained by computing the square roots of the diagonal elements of $\boldsymbol{A}^{-1}$, and evaluating them at $\hat{\beta}$ and $\hat{\sim}$.

## 4. A SPATIAL ANALYSIS OF THE PEP 3-8 DATA FROM THE 1980 PES

The PEP 3-8 data for then $=51$ states of the USA (including Washington, DC) are used to illustrate spatial modeling and fitting. These data are presented in Cressie (1988). The eight explanatory variables, given by Ericksen, Kadane, and Tukey (1989), were collapsed to the 51 states (from 66 small areas that included cities, rests of states, and states). They are:

1. Minority percentage
2. Crime rate
3. Poverty percentage
4. Percentage with language difficulty
5. Education
6. Housing
7. Proportion of population in any of 16 prespecified central cities
8. Percentage conventionally counted in the census.

To find a subset of these variables that provides a good model for undercount, I used the selection method of Ericksen, Kadane, and Tukey (1989), but weighted the data proportionally to the square roots of the small areas' census counts. The variables selected were 1 (minority) and 5 (education), as well as the constant term. Henceforth, in this paper, these three variables will be the only ones considered in the general linear model; i.e., only regression coefficients $\boldsymbol{B}_{0}, \boldsymbol{B}_{1}$, and $\boldsymbol{B}_{5}$ will be fit.

Spatial dependence will be modeled via the Gaussian Markov random field given by (3.3) and (3.17). In (3.17),

$$
H=\left\{\begin{array}{l}
1 ; d_{i v} \leqslant 700 \text { miles, } i * j  \tag{4.1}\\
0 ; \text { elsewhere }
\end{array}\right.
$$

and

$$
\begin{equation*}
D=\operatorname{diag}\left(1 / C_{1}, \ldots, 1 / C_{51}\right) \tag{4.2}
\end{equation*}
$$

The form of the diagonal matrix D is suggested by (1.7). The form of H was obtained by exploratory spatial data analysis; Cressic and Chan (1989) give a more detailed discussion of the choice of H . In (4.1), $\boldsymbol{d}_{i j}$ is the distance between the centers of gravity of the i -th and j -th states. Figure 1 shows the centers of gravity of the 49 mainland states (including Washington, D. C.), and, for illustration, a circle of 700 mile radius centered on Iowa.


Figure 1: Map of the mainland states of the USA, showing their centers of gravity and the 700 -mile neighborhood of the state of Iowa.

It should be emphasized that Figure 1 shows just one way of assigning nonzero $h_{i j}$ 's. Any other assignments, not necessarily distance-based, could be chosen: the only restriction is that $h_{i j} / C_{j}=h_{i j} / C_{i}(i<j)$.

To see the effect of spatial modeling, the following configurations were fit to the model (2.6), (3.3), and (3.17):
(1,1): $\boldsymbol{\beta}_{s}=0, \lambda=0$
(1,2): $\lambda=0$
$(2,1): \boldsymbol{\beta}_{s}=\mathbf{0}$
$(2,2)$ : no restrictions.

Thus, the variable 5 is deliberately omitted (or not), and spatial dependence is set equal to zero (or not). One could use $\hat{t}^{2}$, for each model, to summarize its success at capturing the important features; small values are preferred. Using obvious notation, the values are: $\hat{\tau}_{11}^{2}=109.1, \hat{\tau}_{12}^{2}=47.32, \hat{\tau}_{21}^{2}=18.39$, and $\hat{\tau}_{22}^{2}=0$. Notice that a spatial model with omitted variable has a smaller $\hat{\tau}^{2}$ than a nonspatial model with the variable included, which reinforces my earlier assertion that spatial modeling can compensate for omitted variables.

For the (spatial) configuration (2,1), the following maximum likelihood estimates (with estimated standard errors in parentheses) were obtained.

$$
\begin{array}{ll}
\hat{\mathbf{\beta}}_{0}: 1.00703(0.00254) & \hat{\beta}_{1}: 0.0004453(0.0001209) \\
\hat{\lambda}: 0.04950(0.00016) & \hat{\imath}^{2}: 18.39(20.65) .
\end{array}
$$

The profile negative loglikelihood of $\lambda$, given by (3.19), is presented in Figure 2.


Figure 2: $\quad$ The profile negative loglikelihood of $\lambda$ given by (3.19). The model fit is (2.6), (3.3), and (3.17), with $\beta_{5}=0$ (i.e., $\beta_{\infty} \beta_{j}, \gamma^{2}$, and $\lambda$ are fit). The range of $\lambda$ is determined by the requirement that ( $I-\lambda H)$ be positive definite.

What is important for the study of undercount is the model-based, empirical Bayes predictions given by (2.8) (with $\underset{\sim}{\gamma}$ replaced by $\underset{\underset{\gamma}{\gamma}}{ }$ ). In a like manner to the approach taken by Cressie (1990), the differences $\left\{Y_{i}-1: i=1, \ldots, 51\right\}$ can be decomposed into differences that correspond to the effect of smoothing, the effect of spatial modeling, the effect of omitting an explanatory variable, and the effect of adjusting. In obvious notation,

$$
\begin{equation*}
Y_{i}-1=\left(Y_{i}-F_{11, j}^{p n d}\right)+\left(F_{11, i}^{\text {prd }}-F_{21, j}^{p r d}\right)+\left(F_{21, i}^{p r d}-F_{22, j}^{p r d}\right)+\left(F_{22, i}^{p r d}-1\right) \tag{4.3}
\end{equation*}
$$

These values are given in Table 1, along with the values obtained after multiplying respective differences by $C_{i j}$. $i=1, \ldots, 51$. The last row of Table 1 shows the sum of squares (SS) and weighted sum of squares (WSS) of these differences; e.g., for the difference between raw adjustment factors and no adjustment, SS is $\sum_{i=1}^{51}\left(Y_{i}-1\right)^{2}$, and WSS is $\sum_{i=1}^{51} C_{i}\left(Y_{i}-1\right)^{2}$.

The interpretation is clear: It matters most whether the raw adjusted counts are smoothed or not, and whether the census counts are adjusted or not. The omission of a variable or the presence of spatial dependence has relatively little effect on the predicted adjustments.

| State | $\checkmark-1$ | Y-Fil | F11-F21 | F21-F22 | F22-1 |
| :---: | :---: | :---: | :---: | :---: | :---: |
| ala | -0.003528 | -0.012484 | -0.001182 | 0.005967 | 0.004172 |
| aka | 0.028754 | 0.009413 | 0.005502 | -0.002486 | 0.016326 |
| arz | 0.020377 | 0.009467 | -0.004609 | 0.000464 | 0.015054 |
| ark | -0.010508 | -0.014640 | -0.000425 | 0.007382 | -0.002824 |
| cal | 0.030674 | 0.015399 | -0.003956 | -0.002078 | 0.021308 |
| col | 0.003251 | -0.003132 | -0.005277 | -0.007055 | 0.018714 |
| con | -0.011418 | -0.014126 | -0.000396 | -0.004594 | 0.007698 |
| de 1 | -0.006211 | -0.005859 | 0.009368 | -0.020133 | 0.010413 |
| fla | O 014384 | 0.002854 | -0.002720 | 0.002561 | 0.011689 |
| g9a | -0.004519 | -0.013227 | -0.002086 | 0.005739 | 0.005055 |
| hat | 0.011061 | 0.003561 | -0.003582 | 0.001063 | 0.010019 |
| lah | 0.012545 | 0.002909 | -0.000092 | 0.003012 | 0.006716 |
| 111 | 0.021051 | 0.008822 | 0.001811 | 0.000351 | 0.010067 |
| ind | -0.006387 | -0.008655 | 0.000879 | -0.001389 | 0.002778 |
| low | -0.006784 | -0.006468 | -0.001265 | -0.002367 | 0.003317 |
| kan | 0.005581 | 0.000785 | -0.0000 14 | -0.003934 | 0.008744 |
| kty | -0.015506 | -0.014840 | 0.001105 | 0.009423 | -0.011194 |
| lou | 0.023384 | 0.007688 | -0.000584 | 0.007465 | 0.008815 |
| mne | 0.020065 | 0.009081 | 0.009147 | 0.002312 | -0.000475 |
| mla | 0.024187 | 0.010805 | 0.005306 | -0.005404 | 0.013480 |
| mas | -0.011820 | -0.010630 | -0.003878 | -0.003863 | 0.006551 |
| men | 0.007866 | 0.000278 | 0.000951 | -0.001315 | 0.007953 |
| min | 0.011081 | 0.005456 | -0.000024 | 0.000895 | 0.004755 |
| mis | 0.009652 | -0.006049 | 0.002150 | 0.004703 | 0.008848 |
| mou | 0.008028 | 0.001473 | 0.000874 | 0.004227 | 0.001455 |
| mon | 0.014405 | 0.006089 | 0.000396 | 0.002283 | 0.005637 |
| neb | 0.000761 | -0.000737 | -0.002372 | -0.002945 | 0.006815 |
| nev | 0.026546 | 0.011352 | 0.000995 | -0.000037 | 0.014236 |
| nwh | -0.015835 | -0.009294 | 0.002056 | -0.011849. | 0.003251 |
| nw 1 | 0.013048 | 0.002963 | 0.001343 | -0.001475 | 0.010217 |
| nwm | 0. 023583 | 0.002630 | -0.001859 | -0 001562 | 0.024374 |
| nwy | 0.016552 | 0.004626 | -0.001020 | 0.001352 | 0.011594 |
| noc | 0011849 | 0.000348 | 0.003226 | 0.007732 | 0.000543 |
| nod | 0.000480 | -0.000042 | -0.001076 | 0.004132 | -0.002534 |
| ono | 0.010760 | 0.003916 | 0.002032 | 0.000281 | 0.004530 |
| OH. 1 | -0.002305 | -0.005148 | -0.001848 | 0.002648 | - 002044 |
| ore | 0.002697 | 0000102 | -0.005614 | -0.000174 | -. 008383 |
| pen | -0.002806 | -0.006170 | -0.001637 | 0.003342 | 0.001659 |
| rhi | 0.008858 | 0.002519 | 0.009480 | 0.002039 | -0.005181 |
| soc | 0.063208 | 0.046182 | 0.007060 | 0.005336 | 0.004631 |
| sod | 0.000831 | -0.000044 | -0.0005 12 | 0.002515 | -0.001129 |
| ten | -0.028315 | -0.031807 | 0.000013 | 0.005944 | -0 002465 |
| tex | 0.003708 | -0.011249 | -0.005440 | 0.005826 | 0.014570 |
| uth | 0.003966 | 0.000478 | -0.004322 | -0.005410 | 0.013221 |
| $v \mathrm{mt}$ | -0.011135 | -0.001267 | 0.001987 | -0.013766 | 0.001911 |
| vir | 0.000911 | -0.006515 | 0.001096 | 0.000362 | 0.005968 |
| was | 0.014188 | 0.008573 | -0.004066 | -0.001643 | 0.011325 |
| wev | -0.005796 | -0.006316 | 0.008472 | 0.002882 | -0.010835 |
| wis | 0.017349 | 0.008130 | 0.004049 | 0.001887 | 0.003284 |
| wyo | 0036087 | 0.014581 | 0.009379 | 0.000181 | 0.011946 |
| acl | 0.037495 | 0.001987 | 0.020905 | -0.030369 | 0.044973 |
| \$ \$ | 0.016525 | 0.006176 | 0.001243 | 0.00240 | 00 |

CENSUS

| 3856169 | - 13603 | -48142 | -4560 | 23010 | 16088 |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 398316 | 19453 | 3749 | 2191 | -990 | 6503 |
| 2699377 | 55005 | 25556 | $-12+42$ | 1254 | 40637 |
| 2258342 | -23732 | -33063 | -960 | 16670 | -6379 |
| 23417367 | 718296 | 360598 | -92628 | -48660 | 498987 |
| 2861207 | 9300 | -8960 | - 15098 | -20186 | 53545 |
| 3065974 | - 35008 | -43310 | - 1215 | - 14085 | 23603 |
| 588903 | -3658 | - 3450 | 5517 | -11857 | 6132 |
| 9654302 | 138867 | 27555 | -26259 | 24726 | 1128.15 |
| $540+451$ | -24425 | -71486 | - 11272 | 31013 | 27319 |
| 959215 | 10610 | 3416 | $-3436$ | 1020 | 9610 |
| 935683 | 11739 | 2722 | -86 | 2818 | 6284 |
| 11291203 | 237687 | 99611 | 20448 | 3962 | 113665 |
| 5428596 | -34670 | -46982 | 4771 | -7538 | 15080 |
| 2866173 | -19443 | - 18539 | -3626 | -6785 | 9506 |
| 2326235 | 12983 | 1826 | - 32 | -9151 | 20340 |
| 3623499 | -56185 | -53773 | 4004 | 34144 | -40561 |
| 4157253 | 97215 | 31963 | -2426 | 31034 | 36644 |
| 1110298 | 22278 | 10083 | 10155 | 2567 | -527 |
| 4171724 | 100902 | 45075 | 22135 | -22545 | 56236 |
| 5661335 | -66918 | -60181 | -21952 | -21870 | 37085 |
| 9158324 | 72041 | 2543 | 8705 | - 12047 | 72840 |
| 4013408 | 44474 | 21899 | -98 | 3590 | 19083 |
| 2497274 | 24104 | -15106 | 5369 | 11745 | 22097 |
| 4858 \$39 | 39004 | 7156 | 4244 | 20535 | 7069 |
| 778046 | 11207 | 4737 | 308 | 1776 | 4386 |
| 1546576 | 1176. | - 1140 | - 3668 | -4555 | 10539 |
| 793841 | 21074 | 9012 | 790 | -29 | 11301 |
| 911430 | -14433 | -8471 | 1874 | - 10799 | 2963 |
| 7298030 | 95225 | 21624 | 9799 | - 10762 | 74564 |
| 1292790 | 30488 | 3400 | -2403 | -2019 | 31510 |
| 17335623 | 286944 | 80188 | -17677 | 23436 | 200996 |
| 5811925 | 68864 | 2020 | 18749 | 44936 | 3158 |
| 642418 | 309 | -27 | -692 | 2655 | -1628 |
| 10678666 | -14898 | 41815 | 21702 | 3001 | 48379 |
| 2984855 | -6879 | - 15367 | -5515 | 7903 | 6100 |
| 2604411 | 7025 | 266 | -14622 | -453 | 21833 |
| 11736077 | -32927 | -72412 | -19210 | 39226 | 19469 |
| 934631 | 8279 | 2355 | 8860 | 1906 | -4842 |
| 3088114 | 195193 | 142614 | 21801 | 16477 | 14301 |
| 678440 | 564 | -30 | -347 | 1706 | -766 |
| 4545692 | -128711 | -144584 | 58 | 27021 | - 11206 |
| 14069400 | 52171 | -158260 | -76532 | 81969 | 204995 |
| 1451408 | 5756 | 693 | -6273 | -7853 | 19189 |
| 505424 | -5628 | -640 | 1004 | -6957 | 966 |
| 5291281 | 4819 | -34470 | 5797 | 1913 | 31580 |
| 4084245 | 57949 | 35014 | -16607 | -6712 | 46254 |
| 1935011 | -11216 | -12221 | 16394 | 5578 | -20966 |
| 4639675 | 80495 | 37719 | 18788 | 8753 | 15235 |
| 466633 | 16839 | 6804 | 4376 | 85 | 5574 |
| 630429 | 23638 | 1252 | 13179 | -19146 | 28352 |
|  | 70421 | 27838 | 2618 | 4258 | 28602 |

## 5. DISCUSSION

The decomposition (4.3) is valuable in that it shows that the empirical Bayes adjustment is relatively insensitive to changes in modeling assumptions. Notice that the components of (4.3) are not orthogonal, although they are nearly so. Modulo this lack of orthogonality, it can be concluded that adjusting by the raw adjustment factors leads to a gross change in census counts, but that a model-based, empirical Bayes adjustment offers a compromise to those who advocate no adjustment at all.

These results agree with Cressie's (1990) conclusions, based on a comparison of the homoskedastic model (1.8) with the heteroskedastic models (1.6) and (1.7). However, Cressie demonstrated that the mean squared prediction errors of the small areas are more sensitive to model assumptions. The same is true in this study of spatial prediction of adjustment factors. A comparison of methods for estimating the parameters $\boldsymbol{\gamma}$ demonstrated a similar pattern of sensitivity for predictors and mean squared prediction errors. Maximum likelihood estimators can be badly negatively biased for small sample sizes, leading to an oversmoothing of the raw data $\underset{\sim}{Y}$. Then, the empirical Bayes predictor is often $X \hat{Q}$; this over-reliance on the model (2.5) is an understandable source of discomfit to the Census Bureau.

Methods for estimating $\underset{\underset{\gamma}{\boldsymbol{\gamma}}}{ }$ that are free from the bias problem detailed above are clearly needed. Cressie (1990) explored a method-of-moments (m.o.m.) approach to estimating $\underset{\sim}{\boldsymbol{\gamma}}$. In the simple variance-components model (1.7), (2.5), and (2.6), he found the m.o.m. estimate of the variance-component parameter $\tau^{2}$ to be, in general, less biased towards zero than the m.l. estimate. However, there is no casy asymptotic distribution theory associated with m.o.m. estimation. It will be demonstrated elsewhere that the two requirements, of small bias and asymptotic distribution theory for the estimator of $\underset{\sim}{\boldsymbol{\gamma}}$, can be satisfied by restricted maximum likelihood (REML) estimation (Patterson and Thompson, 1971).

## ACKNOWLEDGEMENTS

My thanks go to Robert Parker, who helped me with the computations and preparation of figures. This research was partially supported by Joint Statistical Agreements JSA $89-23$ and JSA $90-41$, between the U.S. Bureau of the Census and Iowa State University. The findings and opinions expressed in this article are of the author alone, and do not necessarily reflect those of the Census Bureau. The research was also partially supported by the National Science Foundation under grant DMS-8902812.

## REFERENCES

Anselin, L., and Griffith, D.A. (1988). Do spatial effects really matter in regression analysis? Papers of the Regional Science Association, 65, 11-34.

Besag, J.E. (1974). Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society B, 36, 192-225.

Cressie, N. (1988). Estimating census undercount at national and subnational levels, in Proceedings of Bureau of the Census Fourh Annual Research Conference, U. S. Bureau of the Census, Washington, D. C., 127-150.

Cressie, N. (1989). Empirical Bayes estimation of undercount in the decennial census. Joumal of the American Statistical Association, 84, 393-401.

Cressie, N. (1990). Weighted smoothing of estimated undercount, in Proceedings of Bureau of the Census 1990 Annual Research Conference, Bureau of the Census, Washington, D. C., 301-325.

Cressie, N., and Chan, N.H. (1989). Spatial modeling of regional variables. Joumal of the American Statistical Association, 84, 393-401.

Diffendal, G. (1988). The 1986 Test of Adjustment Related Operations in Central Los Angeles County. Survey Methodology, 14, 71-86.

Doreian, P. (1980). Linear models with spatially distributed data. Spatial disturbances or spatial effects? Sociological Methods and Research, 9, 29-60.

Dow, M.M., Burton, M.L., and White, D.R. (1982). Network autocorrelation: A simulation study of a foundational problem in regression and survey research. Social Networks, 4, 169-200.

Dubin, R.A. (1988). Estimation of regression coefficients in the presence of spatially autocorrelated error terms. Review of Economics and Statistics, 70, 466-474.

Ericksen, E.P., and Kadane, J.B. (1985). Estimating the population in a census year: 1980 and beyond. Joumal of the American Statistical Association, 80, 98-109.

Ericksen, E.P., Kadane, J.B., and Tukey, J.W. (1989). Adjusting the 1980 Census of Population and Housing. Joumal of the American Statistical Association, 84, 927-944.

Freedman, D.A., and Navidi, W.C. (1986). Regression models for adjusting the 1980 Census. Statistical Science, 1, 3-17.

Hainer, P., Hines, C., Martin, E., and Shapiro, G. (1988). Research on improving coverage in household surveys, in Proceedings of Bureau of the Census Fourth Annual Research Conference, U. S. Bureau of the Census, Washington, D. C., 513-539.

Kitanidis, P.K., and Lane, R.W. (1985). Maximum likelihood parameter estimation of hydrologic spatial processes by the Gauss-Newton Method. Joumal of Hydrology, 79, 53-71.

Mardia, K.V., and Marshall, R.J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. Biometrika, 71, 135-146.

National Academy of Sciences. (1985). The Bicentennial Census: New Directions for Methodology in 1990, eds Citro, C.F. and Cohen, M.L. National Academy Press, Washington, D. C.

Patterson, H.D., and Thompson, R. (1971). Recovery of interblock information when block sizes are unequal. Biometrika, 58, 545-554.

Stephan, F. (1934). Sampling errors and interpretations of social data ordered in time and space, in Proceedings of the Amencan Statistical Joumal, New Series No. 185A, ed. F. A. Ross. Joumal of the American Statistical Association, 29, Supplement, 165-166.

Whitte, P. (1954). On stationary processes in the plane. Biometrika, 41, 434-449.
Wolter, K.M. (1986). Some coverage error models for census data. Joumal of the American Statistical Association, 81, 338-346.

SESSION 4

## Measuring Survey Error

# MEASUREMENT OF CONTENT DATA QUALITY IN THE 1990 CENSUS 

H.F. Woltman and K.F. Thomas'


#### Abstract

The research program for the measurement of nonsampling error associated with the content data collected for the 1990 Census involves studies designed to measure error due to respondents, enumerators, and missing data. In addition, a mathematical model, specific to the 1990 Census data collection and processing procedures, has been developed to allow estimates from these sources to be combined to produce a rough estimate of the "total" error. This paper will provide an overview of this research program.


KEY WORDS: Reinterview; Nonsampling Error; Total Error.

## 1. INTRODUCTION

Measurement of content error via specially designed studies has been and will continue to be an integral part of the research, evaluation, and experimentation phase of the decennial Census. This is an overview of the plans of the US Census Bureau for measuring content error in the 1990 Census. First, a brief discussion of the content development process which determines the subject areas and specific questions included on the census questionnaires is presented. Also discussed are the objectives of the content measurement error evaluation program and details of the methodology and sample design for each of the individual studies conducted as part of the evaluation. Finally, an overview of the error model developed to provide estimates of the total error, both sampling and nonsampling error components, associated with census statistics is presented.

## 2. OVERVIEW OF THE CONTENT DEVELOPMENT PROCESS

### 2.1 General

The Twenty-first Decennial Census was conducted on April 1, 1990. Long before Census Day, the Census Bureau had to make important decisions about the subjects and questions to be included in the census. By law, the Census Bureau had to submit the subjects planned for inclusion in the 1990 census questionnaires to Congress by April 1, 1987; specific questions had to be submitted to Congress by April 1, 1988.

The content development process determines which subjects and questions are included on the census questionnaires. There are two versions of the census questionnaires: a short form containing a limited number of basic population and housing questions, and a long form containing these basic questions as well as a number of additional questions. Special questionnaires, including the Individual, Military, and Shipboard Census Reports, are used for the enumeration of some persons. These forms have the same population questions as the short or long form questionnaires.

[^17]
### 2.2 Major Components of Content Planning

The process began with a review of the questions asked in the 1980 census. The next step was the identification of the projected 1990 data needs and an examination of current and emerging legislation that might affect the requirements for census information. At this point, the Census Bureau organized a complex structure of meetings, conferences, and working groups to assure that the various segments of the data-user community would be consulted in the content development process. This culminated in a list of questions that were the most likely candidates for inclusion in a series of tests. In most cases, general data needs were stated rather that specific content recommendations, and the Census Bureau developed questions to meet these needs.

A variety of meetings were held to obtain content recommendations, beginning with the Local Public Meetings in the spring of 1984 and followed by meetings of advisory committees and Federal agencies that got together periodically throughout the census cycle to discuss 1990 census activities.

Content recommendations were reviewed by Census Bureau analysts who applied the criteria discussed below to determine which items merited consideration and testing. The number of recommendations that were received from all sources was far too numerous to include on a census questionnaire, and many were not required by Federal legislation. The most promising questions were tested in the National Content Test (NCT), test censuses, and special purpose tests. Comprehensive evaluations were made to ascertain the accuracy and reliability of these items. A list of subjects planned for inclusion in the 1990 census was sent to Congress March 27, 1987.

Refinements in census content were made after additional testing in the summer and fall on 1987. The culmination of this process was the presentation of the proposed 1990 census questions and response categories to Congress on April 1, 1988.

### 2.3 Criteria for Determining 1990 Content

The Census Bureau was guided in the selection of questions by five basic criteria. First, only essential data were considered-those with a broad demonstrated societal need and those needed to meet Federal, state, and local statutory data requirements and to administer governmental programs.

These data must have been needed for relatively small areas (local governments and small statistical areas) or numerically small population groups. If data are required only at the mational or regional level, sample surveys are more appropriate vehicles.

Sccond, many of the questions asked in 1980 were repeated in 1990 because they provide basic information required for the country as well as a continuum of vital sociocconomic and housing data. The relative stability of census content over the last few decades stems, in part, from the relevance and usefulness of basic items and the need to measure how our population and housing stock have changed over time.

Third, there was no significant increase in the number of questions the Census Bureau asked in 1990, relative to the 1980 census. The Bureau had to strike the proper balance between the need for information and the length of the questionnaires. This was necessary because public cooperation-essential for a successful censuscould be undermined by a questionnaire that respondents find too burdensome.

Fourth, the forms did not contain questions that are intrusive, offensive, or widely controversial. For example, it is illegal to compel respondents to answer questions about religious beliefs or affiliations. Other controversial subjects that could influence or reduce response to the census also were avoided.

Fifth, the Census Bureau had to be able to formulate a clear, concise question on each subject that would yield accurate data. Wording and format were especially important because the census was conducted primarily by mail, using a self-administered questionnaire. If test results indicated that a question was likely to be misinterpreted, even by a segment of the population, it was unlikely to be included on the questionnaire.

Various tests were carried out to evaluate alternative question wording, order, and formatting. The National Content Test (NCT) was the major test vehicle. Other testing vehicles, however, including the test censuses and special purpose tests, were also used for designing and developing the 1990 census questionnaires (U.S. Dept. of Commerce, 1987).

### 2.4 Summary

The 1990 census was the bicentennial anniversary of census-taking in the USA. Many difficult choices about census content were made to complete 1990 census plans. This important task of selecting the census subject content was the result of careful review and testing of recommendations from a large variety of users of census data, including Federal and other governmental agencies, advisory committees, professional groups, and members of the general public.

## 3. OBJECTIVES OF THE 1990 RESEARCH, EVALUATION, AND EXPERIMENTAL PROGRAM

The Bureau of the Census has grouped most of the component studies of the 1990 Rescarch, Evaluation, and Experimental (REX) Program into three principal areas. We have titled these groupings "Content," "Coverage," and "Procedures and Processing."

The Content grouping comprises those REX studies that will ensure that we obtain information on the following:

1. We will assess the quality of data from the 1990 Census in order that we might provide information about the sources and magnitude of nonsampling errors introduced during data collection and processing.
2. We will assess the effect of sampling, nonsampling, coverage, and geographic errors on data use. We will identify major uses of the data and evaluate how errors in census data affect applications.
3. Using the data collected in the census and the evaluations as a base, we will determine more efficient and more accurate ways of collecting census data.
4. We will implement procedures that will allow for monitoring, documenting, and, if appropriate, correcting errors identified in data collection and processing.

The primary objectives that were identified for the 1990 REX Program, as for any research and evaluation effort, are:

1. to provide data for Census Bureau use in assessing and improving methods and operations for future censuses; and
2. to provide information to data users concerning the sources and effects of errors in census data.

In addition, the Bureau used the following specific criteria in selecting the components of the 1990 REX Program:

1. The proposal should be capable of integration within the staffing and fiscal resource limitations of the overall REX program;
2. The proposal should require an actual census environment for proper measurement. (Otherwise, it would be more appropriate to integrate the proposal into a test census or other special test.)
3. The proposal should not adversely effect the production and delivery of apportionment and redistricting counts by the legislated deadlines.
4. The proposal should not affect the quality of the census data in any way that would jeopardize their major uses.

With these criteria in mind, the content evaluation program encompasses four major evaluation projects. This includes the Content Reinterview Study (CRS), the Enumerator Variance Study (EVS), Imputation Evaluation and Research, and the Master Trace Study (MTS). In addition, a measurement error model has been developed that attempts to account for the major census operational processes and error sources.

## 4. INDIVIDUAL STUDIES

The content evaluation program encompasses four major studies designed to measure content error due to respondents, enumerators and imputation for missing data. The specific methodology, sample design, and sample size used for each study is discussed below.

### 4.1 Content Reinterview Survey (CRS)

Traditionally, the Content Reinterview Survey (CRS) is conducted following the Decennial Census and is designed to measure the response error associated with selected population and housing items. The objective is to measure simple response variance and response bias associated with the data items which are reinterviewed. The 1990 CRS will have this same goal of evaluating the quality of the data. The 1990 study will focus on new data items and will have innovations in field data collection techniques.

The Content Reinterview Survey will compare individual responses collected during the census to those collected in the reinterview. Measures of response error will be generated from the sample to estimate systematic and random nonsampling error associated with the census long form data.

The sample was restricted to long form households. Both households responding by mail and nonmail return households will be reinterviewed.

A single stage systematic sampling scheme was used. In order to obtain a final sample of 12,800 occupied units, an initial sample of 15,500 housing units was selected from the census address file. This number of households is comparable to the designated sample sizes of the 1970 CRS and the 1980 CRS.

All initial reinterviews are being conducted using Computer-Assisted Telephone Interviewing (CATI). The results of the CRS are used as the standard by which the validity of Census responses are measured; therefore, it is essential that the reinterview results are of the highest quality. Increased data quality is associated with the CATI system due to a reduction in interviewer bias because of the CATI interview (the computer controls skip patterns) and supervisory controls (the supervisor can monitor both the audio portion of the interview on the telephone and the video portion on the screen).

The Content Reinterview Survey has traditionally experienced high response rates due to extensive use of followup procedures. For 1990, introductory letters will be mailed to the sampled housing units prior to the start of the telephone interview. Households will be contacted by CATI until complete information for each person and the housing unit (or a specified number of failures to reach the sample unit) has occurred. Characteristics of each person 15 and over will be obtained by self-response. Cases which cannot be contacted by telephone will be sent to the field for personal visit followup.

It was necessary to obtain the household roster and telephone number from the questionnaire for the CRS sample. This is a requirement of the CATI system, and additionally, has advantages in terms of linking the census and reinterview data for identical persons.

Cross tabulations will be performed between the responses obtained in the census and in the reinterview for identical housing units and persons to compute measures of response variance and response bias.

For some items in the Content Reinterview Survey a set of detailed probing questions allows gathering data with a degree of accuracy not possible in the census. That is, the Content Reinterview Survey may be viewed as the "preferred" measurement technique. Comparison of the reinterview data with the census will provide estimates of response bias present in the census data. Items which will be evaluated using a response-bias (probing) type reinterview include:

| Population Items | Housing Items |
| :--- | :--- |
| Race | Tenure |
| Place of Birth | Monthly Rent |
| Citizenship | Meals with Rent |


| Population Items (cont.) |  |
| :--- | :--- |
| Education Level | Housing Items (cont.) |
| Ancestry | Numbing Facilities |
| Language Usage | Year structure Built Vans |
| Military Service |  |
| Employment Status |  |
| Disability |  |

Response variance is estimated for some items by asking the same question(s) asked in the census. Items which will be evaluated using a response-variance type reinterview include:

Spanish Origin<br>Year of Immigration<br>School Enrollment<br>Employer: Kind of Business<br>and Type of Company<br>Description of Building<br>Size of Lot<br>Agricultural Sales

### 4.2 Enumerator Variance Study (EVS)

Nonresponse followup (NRFU) enumerators are an important source of nonsampling error in census statistics for the 1990 Decennial Census. The purpose of the Enumerator Variance Study (EVS) is to estimate the contribution of NRFU enumerators to the total nonsampling error.

The enumerator variance study used a three stage design to collect data from NRFU enumerators in metropolitan areas:

1. Fifteen strata of district offices (DO's) were formed and 2 DO's were selected from each stratum.
2. Seventeen geographic address register areas (ARAs) (generally equivalent to a census tract) were selected from the DO's selected in the first stage.
3. Census blocks within sample ARAs were randomly allocated to enumerator assignments in order to interpenetrate enumerator assignments.

There are some limitations with this study:

- The greater part of each NRFU assignment is expected to be completed by a single enumerator. The NRFU assignment error variance should be smaller than the enumerator error variance for assignments completed by more than one enumerator.
- In order to minimize operational problems and cost, it was decided to interpenetrate census blocks rather than individual households. Enumerator assignment variance will not be estimable for geographic units smaller than two blocks. Within block assignment variance will be confounded with between block assignment variance.
- The results will be limited to mostly urban portions of the country.


### 4.3 Imputation Research

Imputation is the assignment of information for unreported items on a questionnaire. The effects of imputation on census data quality, and the magnitude of error it introduces into published census statistics, are not well understood.

Evaluation of the census imputation procedures will provide information on how much of census data bas been imputed and the ways imputation has affected the data. Research will help to choose among alternative imputation methods. Eventually, better imputation procedures will result from testing and adapting a variety of alternative methods.

### 43.1 Evaluating the Current Hot Deck Imputation Procedure

The sequential hot deck imputes missing data values by finding the "last similar" individual or household in the data file; thus it leans heavily on the sequential ordering of households in the file, which corresponds roughly to a geographical ordering in the physical universe. This sequential procedure has been used for two primary reasons: (1) It is computationally convenient, requiring only a single pass through the census data file; (2) data analysis has shown that for many census data items, strong correlations exist among neighbors, and these correlations drop off as the distance between households increases; hence, a nearby household will tend to be a better predictor than one that is far away in a data file (Thomas et al, 1984).

Evaluation of the hot deck will include several phases (Schafer, 1989, Schafer 1990).

## a. Descriptive Evaluation

Describe the set of rules by which the hot deck fills in missing items. These rules can be divided into two broad classes; consistency edit rules and imputation rules. Distinguishing between these two types of rules will be important, both for evaluating the current hot deck and for proposing improvements and alternatives.

## b. Estimating Imputation Error Rates

The current Total Error Model for the census (Woltman, Johnson, 1989) attempts to integrate the various sources of error in the census into a single model and derives an approximate mean-squared error for a census statistic under this model. The statistic, p , is a census proportion that estimates the proportion P of individuals or households in the population bearing some characteristics; i.e., p is the proportion of "positives" observed in the census, whereas $\mathbf{P}$ is the true proportion of positives in the population.

The Total Error Model specifies two parameters for the error due to imputation of missing characteristics: $\theta_{1}$, the probability of imputing a negative value given that the true value is positive, and $\phi_{1}$ the probability of imputing a positive value given that the true value is negative. Data from the Content Reinterview Study and, possibly, the 1990 Post-enumeration Survey (PES) can be used to provide point estimates for both 100 percent and sample data. Schafer (1989) also proposes a reparameterization of the error rates that separate the effect of the consistency edit and the effect of imputation.

## c. Data Analyses

Analyses proposed by Schafer (1990) are philosophically similar to the study by Thomas et al (1984) in that they look for spatial relationships in the observed census data, in the hope that these relationships will hold for the missing (and subsequently imputed) data as well. If some spatial relationship is found to accurately predict the observed data values, then it reasonable to suspect that it may be a useful device for imputation.

The basic idea is to apply a hot deck imputation rule to census data, and see how well this imputation rule predicts the data we actually observe. Suppose Y is a census item of interest (e.g., Hispanic origin) to be imputed. For each unit (household) in the data file, we can apply the imputation rule to find that the unit's donor; for example, the donor may be defined as the last household that matches on some other related item, such as number of persons. By comparing the donor's value of Y to the recipient's value of Y , we can assess the performance of this particular imputation rule within our dataset. Summary statistics and plots that may be useful for judging and comparing the performance of various imputation rules will be produced.

### 4.3.2 Developing Improved Imputation Procedure

The current hot deck algorithm incorporates a great deal of accumulated experience about the structure of the census data, both at the individual and household levels, and about the mechanism that causes data to be missing. Schafer (1989) presents a reformulation of the hot deck as a formal statistical model. With this model, we hope to retain the intelligence of the hot deck's imputation rules, while developing a more principled, less ad hoc imputation method.

Other methods of evaluating alternative sequential hot deck matching rules, for investigating the optimal tradeoff between matching on covariants versus geographical proximity are discussed in Schafer (1989). Finally alternatives to the sequential hot deck imputation procedure which relies heavily on the geographical information contained in the ordering of units within the data file will be investigated. Some of these include evaluating more cluster based approaches, e.g., defining the donor pool within blocks or block groups (Schafer 1989 and Blodgett 1990).

### 4.4 Master Trace Study (MTS)

The purpose of the Mater Trace Study (MTS) is to create and document a comprehensive database to support ongoing Research, Evaluation and Experimental (REX) studies and to provide data for future census evaluations. These data should be able to provide sample geographic and demographic information at various stages of census collection and processing. "Snap-shots" of a sample of census questionnaires (both short and long forms) at various stages of census processing and a post-censal reinterview will allow evaluation of the quality of individual responses to the answers, and of the effects of collection procedures, processing procedures and the edit and imputation process on responses to individual questions.

The goal of the Master Trace Study is to produce a database that contains an entry for all questionnaire items at each stage of census processing for each sample census record. These results will be used to assess the quality and accuracy of changes made to respondent or enumerator filled questionnaire entries during the census processing up to the point of final tabulations. The results of this study will also identify the sources of processing-induced error in census data in order that improved procedures and processes to minimize error in the future might be developed. This study will further provide data for use in assessing the extent and effects of content error on census data use.

A sample of 15,500 long form and 15,500 short form questionnaires were randomly sampled from across the nation for MTS. The sample of census long form questionnaires for the Master Trace is the same as the 1990 Content Reinterview Survey (CRS) sample. It is estimated that approximately $60-70$ forms per district be Master Trace.

### 4.5 Total Error Model

A measurement error model has been developed to provide estimates of the total error (both sampling and nonsampling error components) of census statistics (Woltman and Johnson, 1989). The term "total error" implies an integrated treatment of different sources of error in census or survey statistics. This model falls short, however, of treating all sources of error in a single model. Hence the term "total error model" is not used in a precise sense.

The applications of such a model are as follows:

1. The estimation of the total error of a census statistic including both sampling and nonsampling error components. Statistical inferences can then be based on the overall accuracy of statistics rather than just the sampling precision.
2. Gauging the relative impacts of different kinds of error on the total error. This can allow more informed judgement about what processes/operations need to be improved.
3. Applying this knowledge to the efficient allocation of resources in future censuses so as to minimize total error for fixed cost.

The model seeks to account for the following sources of error:

- Noncoverage of the total population
- Sampling error (in the case of statistics produced from the census sample)
- Response error
- Enumerator error
- Imputation crror

In addition, the model seeks to reflect these specific error sources as they are associated with the following census processes and operations. Namely:

- Whether or not a questionnaire is returned in the mail (i.e. mail return status),
- Editing for missing data of the questionnaires returned by mail,
- Followup of questionnaires that fail edit due to the specific edit tolerance rules applied,
- Followup of questionnaires not returned by mail, and
- The imputation procedure in the case where data are missing after completing all followup activities.

Specifically, then, the total error model seeks to account for the processes/operations giving rise to sampling and nonsampling errors in the proportion of housing units or persons having a particular attribute based on the observed and imputed values to a particular questionnaire item.

Estimates of various error parameters provided from the individual studies will be used in conjunction with the model to simulate total error estimates for various census statistics at various levels of tabulation geography.

## REFERENCES

Blodgett, R. (1990). Changing Point, internal Census Bureau memorandum.
Schafer, J.L. (1989). Proposal for Imputation Evaluation, internal Census Bur eau memorandum.
Schafer, J.L. (1990). Comments on Autocorrelation and Imputation, internal Census Bureau memorandum.
US Department of Commerce, Bureau of the Census (1987). The Content Development Process for the 1990 Census of Population and Housing.

Thomas, K.F., Harner, D.A., and Fay, R. (1984). Preliminary Evaluation Results Memorandum No.69, Intraclass Correlations Using a Sample of 1980 Census Data, internal Census Bureau memorandum.

Woltman, H.F., and Johnson, R. (1989). A Total Error Model for the 1990 Census, Proceedings of the Survey Research Section, American Statistical Association.

# COMPARING THREE BOOTSTRAPS FOR SURVEY DATA 

R.R. Sitter


#### Abstract

Various bootstrap methods for variance estimation and confidence intervals in complex survey data, where sampling is done without replacement, have been proposed in the literature. The oldest, and perhaps the most intuitively appealing, is the BWO method proposed by Gross(1980). Unfortunately, the BWO method is only applicable to very simple sampling situations. We first introduce extensions of the BWO method to more complex sampling designs. The performance of the BWO and two other bootstrap methods, the rescaling bootstrap (Rao \& Wu 1988) and the mirror-match bootstrap (Sitter 1990), are then compared through a simulation study. Together these three methods encompass the various bootstrap proposals.


KEY WORDS: Bootstrap; Jackknife; Two-stage sampling; Edgeworth expansion.

## 1. INTRODUCTION

For most complex survey designs, unbiased variance estimators of statistics that are expressible as linear functions of the observations can be derived. For nonlinear statistics and functionals this is generally not the case. Various methods for obtaining variance estimates have been proposed in the literature. Three commonly used methods are, the linearization(or Taylor) method, the jackknife method, and balanced repeated replications(BRR). These methods have some limitations; the linearization method requires theoretical calculation and programming of derivatives which can make it cumbersome to implement; the jackknife, when the primary sampling units are selected without replacement, has only been developed for stratified sampling; and the BRR method is only applicable to stratified sampling with restrictions on the number of units per stratum. In recognition of these limitations, various bootstrap procedures for variance estimation and confidence intervals in sample survey data have been proposed in the literature. This article attempts to compare the performance of three bootstrap methods, the rescaling bootstrap (Rao \& Wu 1988), the mirror-match bootstrap (Sitter 1990), and the without replacement bootstrap (BWO) (Gross 1980, Bickel \& Freedman 1984), in terms of variance estimation and confidence intervals in complex survey data where the sample is taken without replacement. Together these three methods encompass the various bootstrap proposals.

The three methods are outlined in section 2, with the algorithms for stratified sampling without replacement explicitly stated. The BWO, though it has intuitive appeal and seems the most natural of the three methods, is unfortunately only applicable to simple random sampling without replacement. To consider it as a competitor of the other two bootstraps, an extension is proposed in section 3 to stratified sampling, two-stage cluster sampling, and the Rao-Hartley-Cochran (1962) sampling designs. The results of a simulation study using various finite populations are given in section 4 . The methods are compared in terms of variance estimation and confidence intervals for the nonlinear statistics $r$, the ratio; $b$, the regression coefficient; and $c$, the correlation coefficient, and for $m$, the median. Stratified random sampling without replacement, with various strata sizes, numbers of strata, and within stratum sampling fractions is considered. The linearization, and jackknife methods

[^18]are included in the study for $r, b$, and $\boldsymbol{c}$, and a method based on Woodruff's (1952) confidence intervals for quantiles is included for the median.

## 2. THE THREE BOOTSTRAPS

To introduce necessary notation, stratified sampling without replacement is outlined at this point. In stratified random sampling, the finite population, consisting of $N$ units, is partitioned into $L$ nonoverlapping strata of $N_{1}, N_{2}, \ldots, N_{L}$ units, respectively; thus, $N_{t}+N_{2}+\ldots+N_{L}=N$. A simple random sample without replacement is taken independently from each stratum. The sample sizes within each stratum are denoted by $n_{1}, n_{2}, \ldots n_{L}$, and the total sample size is $n=n_{1}+n_{2}+\ldots+n_{L^{*}}$. A measurement, or possibly a vector of measurements, of some unit characteristic is represented as $y_{m}$, where the subscript $h$ refers to the stratum label and the subscript $i$ refers to the $i$ ith unit within the $h$ th stratum. The population parameter of interest $\theta=\theta(S)$, where $S=\left\{y_{\mu}: h=1,2, \ldots, L ; i=1,2, \ldots, N_{h}\right\}$, is usually estimated by $\hat{\theta}=\hat{\theta}(s)$, where $s=\left\{y_{m}: h=1,2, \ldots, L ; i=\right.$ $\left.1,2, \ldots, n_{k}\right\}$. The most common situation is when $\theta=\bar{Y}$, the population mean, and $\hat{\theta}=\bar{y}=\sum_{k=1}^{L} W_{k} \bar{y}_{k}$, its usual unbiased estimate, where $\bar{y}_{h}=\sum_{i=1}^{n_{k}} y_{k} / n_{k}$, and $W_{h}=N_{h} / N$. An unbiased estimate of Var $(\bar{y})$ is

$$
\begin{equation*}
\operatorname{var}(\bar{y})=\sum_{k=1}^{L} w_{k}^{2} \frac{1-f_{k}}{n_{k}} s_{k}^{2}, \tag{2.1}
\end{equation*}
$$

where $f_{h}=n_{h} / N_{h}$ and $s_{h}^{2}=\sum_{i=1}^{n_{h}}\left(y_{N}-\bar{y}_{h}\right)^{2} /\left(n_{h}-1\right)$.

### 2.1 The Rescaling Method

A method was proposed by Rao \& $W u(1988)$ for use when $\hat{\boldsymbol{\theta}}=\boldsymbol{g}(\bar{t})$, a function of means, where $\bar{i}=\left(\bar{i}_{1}, \ldots, \overline{b_{j}}\right)$. The general idea is to draw a resample vector with replacement from the original sample, rescale each resampled unit, and apply the original estimator to the rescaled vector. The rescaling factors are chosen so that the variance under resampling matches the usual variance estimator in the linear case. For example, suppose the original sampling is stratified random sampling without replacement, with $\bar{i}_{\varepsilon}=\sum_{k} W_{h} \bar{T}_{\varepsilon(k)}$, and $i_{s(k)}=t_{\varepsilon\left(y_{k}\right)}$, a function of $y_{k}$. Recall that the $y_{k}$ 's may represent a vector of measurements on the $i$ th unit within the $h$ th stratum, so many commonly used statistics can be written as $\hat{\theta}=g(\bar{t})$ (eg. the ratio, the regression coefficient, and the correlation coefficient). With the above formulation in mind the rescaling method is as follows:

1. Draw a simple random sample $\left\{y_{h i}^{*}\right\}_{i=1}^{n_{i}^{*}}\left(n_{k}^{*} \geq 1\right)$ with replacement from $\left\{y_{k}\right\}_{i-1}^{n_{k}}$. Let $C_{k}=\sqrt{m_{k}^{*}}\left(n_{h}-1\right)^{-1 / 2}\left(1-f_{k}\right)^{1 / 2}$, and calculate the following;

$$
\begin{aligned}
& \tilde{t}_{*(k)}^{*}\left.=\bar{t}_{\varepsilon(k)}+C_{k} \bar{t}_{k(k)}^{*}-\bar{t}_{\varepsilon(k)}\right) \\
& \tilde{t}_{k}^{*}=\sum_{k=1}^{L} W_{h} \tilde{H}_{\varepsilon(k)}^{*} \\
& \hat{\theta}^{*}=g\left(\tilde{t}^{*}\right)
\end{aligned}
$$

2. Repeat step 1 a large number of times, $B$, to obtain $\hat{\theta}_{1}^{*}, \hat{\theta}_{2}^{*}, \ldots, \hat{\theta}_{B}^{*}$,, and estimate $\operatorname{Var}(\hat{\theta})$ with $v_{\mathrm{t}}=E_{0}\left(\hat{\theta}^{*}-E_{0} \hat{\theta}^{*}\right)^{2}$, or its Monte Carlo approximation $v_{\mathrm{r}}=\frac{1}{B-1} \sum_{i=1}^{*}\left(\hat{\theta}_{i}^{*}-\hat{\theta}_{\hat{0}}^{*}\right)^{2}\left(E_{0} \hat{\theta}^{*}\right.$ and $\hat{\theta}_{(,)}^{*}$ can be replaced by $\hat{\boldsymbol{\theta}}$ ).

If $\boldsymbol{g}(\bar{t})=\bar{y}$ (univariate), then $v_{\mathbf{1}}\left(\bar{y}^{*}\right)=\operatorname{var}(\bar{y})$, the usual unbiased estimator of $\operatorname{Var}(\bar{y})$. To reduce computation one could rescale the entire sample before initiating the bootstrap step.

Rao \& Wu (1988) develop various rescaling algorithms for a wide class of sampling designs including, stratified sampling, two-stage cluster sampling, and the Rao-Hartley-Cochran (1962) method of pps-sampling. They all yield the usual variance estimates in the linear case. They show that for stratified sampling the method gives consistent variance estimates for $\hat{\theta}=\boldsymbol{g}(\bar{b})$, and in the linear case, with appropriate choice of $n_{k}^{*}$ in each stratum, the bootstrap histogram captures the second-order term of the Edgeworth expansion of $\bar{y}$ as $L-\infty$.

Some limitations of the method do exist, and are discussed in Sitter (1990).

### 2.2 The Mirror-Match Method

Sitter (1990) proposed a bootstrap method which we will call the mirror-match method. Generally, the method entails taking a resample without replacement from the sample to mirror the original sampling scheme, and then repeating this with replacement to match the usual variance estimates in the linear case. In the context of stratified random sampling, the mirror-match method is:

1. Choose $1 \leq n_{h}^{\prime}<n_{h}$ and resample $n_{h}^{\prime}$ without replacement from stratum $h$ to get $y_{h}^{\circ}=\left(y_{h}, y_{h 2}, \ldots, y_{h k_{h}^{\prime}}^{\circ}\right)$.
2. Repeat step $1, k_{h}=\frac{n_{h}\left(1-f_{h}^{*}\right)}{m_{h}^{\prime}\left(1-f_{h}\right)}$ times independently, replacing the resamples of size $n_{h}^{\prime}$ each time, to get $y_{k 1}^{*}, y_{k 2}^{*}, \ldots, y_{h n_{k}^{*}}^{*}$ where $f_{k}^{*}=n_{h}^{\prime} / n_{k}$, and $n_{k}^{*}=k_{k} n_{h}^{\prime}$ (if $k_{k}$ is non-integer a randomization between bracketing integers is used).
3. Repeat steps 1 and 2 independently for each stratum, and let $\hat{\theta}^{\circ}=\hat{\theta}\left(\underline{y}_{i}{ }^{*}, \underline{y}_{2}, \ldots, \underline{y}_{2}{ }^{*}\right)$.
4. Repeat steps $1-3$ a large number of times, $B$, to get $\hat{\theta}_{1}^{0}, \hat{\theta}_{2}^{\circ}, \ldots, \hat{\theta}_{\boldsymbol{B}}^{\circ}$, and estimate $\operatorname{Var}(\hat{\boldsymbol{\theta}})$ with $V_{m}=E_{.}\left(\hat{\theta}-E_{0} \hat{\theta}^{*}\right)^{2}$, or its Monte Carlo approximation $\nu_{m}=\sum_{i=1}^{\theta}\left(\hat{\theta}_{i}^{*}-\hat{\theta}_{(.)}^{*}\right)^{2} / B\left(E_{0} \hat{\theta}^{*}\right.$ and $\hat{\theta}_{(j)}^{*}$ can be replaced by $\hat{\theta})$. If $\hat{\theta}=\bar{y}$, then $V_{m}=\operatorname{var}(\bar{y})$ the usual variance estimate. If $f_{k} \geq 1 / n_{k}$, then choosing $n_{k}^{\prime}=f_{h} n_{k}$ implies the resampling fraction, $f_{k}^{\circ}$, in step 1 is the same as the original sampling fraction, $f_{k}$ (mirror). This choice has theoretical justification using Edgeworth expansions which is described in the next paragraph.

Using this basic idea, Sitter (1990) gives extensions to stratified sampling without replacement, two-stage cluster sampling, and the Rao-Hartley-Cochran (1962) method of pps-sampling. For stratified sampling he also shows that the method yields consistent variance estimates for non-linear statistics, and when $\hat{\theta}=\bar{y}$, for appropriate choice of $n_{k}^{\prime}\left(f_{k}^{*}=f_{k}\right)$, the bootstrap histogram captures the second-order term of the Egdeworth expansion as $L \rightarrow \infty$, as well as when $L$ is bounded and $n, N \rightarrow \infty$. The with replacement bootstrap (BWR), proposed by McCarthy \& Snowden (1985), is a special case of the mirror-match method; for the sampling designs where the BWR has been developed (i.e. for stratified let $n_{k}^{\prime}=1$ ).

Though the method does not share the problems indicated for the rescaling method it also has some limitations: 1) If the stratum sampling fractions are very small the choice of resample size to match the second-order term of the Edgeworth expansion of $\bar{y}$ is not available, and 2) even if the $n_{h}^{\prime \prime}$ s are chosen integers, a randomization between bracketing integers for $k_{k}$ is necessary.

## 23 The BWO Method

The following procedure was proposed by Bickel \& Freedman (1984) as an extension of the BWO method (Gross 1980) to stratified sampling. It was further discussed by McCarthy \& Snowden (1985). Suppose $N_{h}=k_{h} n_{k}+r_{h}$ for $0 \leq r_{k} \leq\left(n_{h}-1\right)$ for each stratum, with $k_{h}$, and $r_{h}$ integers. Construct two pseudo-populations for each stratum, population 1 by replicating $y_{k}, k_{h}$ times, and population 2 by replicating it $k_{k}+1$ times. For each stratum independently, sample $n_{k}$ units without-replacement from population 1 with probability

$$
p_{k}=\frac{\frac{\left(1-f_{h}\right)}{\left(n_{h}-1\right)}-a_{k l}}{a_{k l}-a_{h l}}
$$

where $\left.a_{k j}=\left[k_{h}+j-1\right] /\left[k_{k}+j\right) n_{h}-1\right]$ for $j=0,1$, and from population 2 with probability $1-p_{h}$. If this procedure is feasible then $\operatorname{Var},\left(\bar{y}^{*}\right)=\operatorname{var}(\bar{y})$. Unfortunately it is possible that $p_{h}<0$. An example of this is given by McCarthy \& Snowden (1985) for a single stratum. To illustrate the limitations of this extension, Table 1 gives values of $p$ for various $n$ and $N$ values of a single stratum in stratified sampling.

Table 1: Values of $p$ For The Bickel-Freedman BWO

|  |  | $N$ |  |  |  |  |  |
| ---: | :--- | :--- | :--- | :--- | :--- | :--- | :---: |
| $n$ | 25 | 30 | 40 | 50 | 75 | 100 |  |
| 2 | -253. | -404 | -739. | -1174. | -2628. | -4849. |  |
| 4 | -12.0 | -17.6 | -37.7 | -57.0 | -134.0 | -263.0 |  |
| 6 | -1.90 | -3.48 | -6.79 | -12.5 | -30.8 | -56.8 |  |
| 8 | -0.04 | -0.81 | -2.07 | -4.03 | -11.0 | -20.9 |  |
| 10 | -0.14 | 0.36 | -0.36 | -1.33 | -4.66 | -9.90 |  |
| 15 | 0.17 | 0.89 | 0.07 | 0.16 | -0.45 | -2.00 |  |
| 20 | 0.59 | 0.32 | 0.92 | 0.28 | -0.15 | -0.05 |  |

## 3. AN EXTENDED BWO METHOD

The BWO method for simple random sampling is intuitively appealing, avoids any rescaling, and is easily applicable. Unfortunately it does not extend to more complicated sampling designs. Due to this limited applicability, it would be difficult to justify consideration of the BWO as a possible competitor to the rescaling and mirror-match methods. With this in mind we propose, in this section, extensions of the BWO method to stratified sampling (which is generally applicable), two-stage cluster sampling, and the Rao-Hartley-Cochran method of unequal probability sampling. The extensions give the usual variance estimates in the linear case, and are applicable in general.

### 3.1 Stratified Sampling

Create the $h$ th stratum of the pseudo-population by replicating $\underline{y}_{h}=\left(y_{h}, \ldots, y_{h_{h}}\right), k_{h}$ times. Repeat this for each stratum and resample $m_{h}^{\prime}$ units from stratum $h$ without replacement. The method is similar, but the restrictions that $N_{h}=k_{h} n_{h}$, and that the resample size be $n_{h}$ are removed. Instead $n_{n}^{\prime}$ and $k_{h}$ are chosen to satisfy the following:

$$
\begin{equation*}
f_{h}^{*}=f_{h} \text {, and Var. }\left(\overline{( }_{h}^{*}\right)=\frac{\left(1-f_{h}\right)}{n_{h}} s_{h}^{2} \text {, } \tag{3.1}
\end{equation*}
$$

where $f_{k}^{*}=n_{A}^{\prime} / k_{k} n_{k}$ is the resampling fraction. Ignoring, for the moment, all integer restrictions, these two equations are satisfied by,

$$
\begin{equation*}
n_{h}^{\prime}=n_{k}-\left(1-f_{k}\right) \text {, and } k_{k}=\frac{N_{h}}{n_{k}}\left(1-\frac{1-f_{k}}{n_{h}}\right) \text {. } \tag{3.2}
\end{equation*}
$$

Clearly, the extension will be very close to the standard BWO in many situations.
Of course this only makes sense if $m_{k}^{\prime}$ and $k_{k}$ are both integers for all $h$, and clearly since $0 \leq f_{k} \leq 1, n_{k}^{\prime}=n_{k}-\left(1-f_{k}\right)$ is non-integer unless $f_{k}=0$, or 1 . To avoid this problem one randomizes between bracketing integer values in the following manner. Since the same technique will be used independently for each stratum, the method for a single stratum will be described with the $h$ subscript suppressed.

Let $k_{1}=\{k\}$ and $k_{2}=\lfloor k\rfloor$, where $k$ is given in (3.2) and let $n_{1}^{\prime}=n-1$ and $n_{2}^{\prime}=n$. Also let

$$
\begin{equation*}
p=\frac{\frac{1-f}{n(n-1)}-a_{2}}{a_{1}-a_{2}} \tag{3.3}
\end{equation*}
$$

where $a_{j}=\left[k_{j}\left(1-n_{j}^{\prime} / n k_{j}\right)\right] /\left[n_{j}^{\prime}\left(n k_{j}-1\right)\right]$ for $i=1,2$. It is straightforward to show that, for $n \geq 2,0 \leq p \leq 1$. So the method is generally applicable.

If $n \geq 2$, and $\left(k_{1}, n_{1}^{\prime}\right)$ are used with probability $p$, given in (3.3), and ( $k_{2}, n_{2}^{\prime}$ ) with probability ( $1-p$ ), then $\operatorname{Var} .\left(\bar{y}^{*}\right)=\operatorname{var}(\bar{y})$. To see this consider $k$ and $n^{\prime}$ as described above. Let $E_{20}$ and $V_{20}$ denote the conditional expectation and variance under the resampling given $k$ and $n^{\prime}$. Similarly, let $E_{1}$, and $\boldsymbol{V}_{1}$, denote the expectation and variance with respect to the randomization of ( $n^{\prime}, k$ ). Then

$$
\begin{aligned}
\text { Var. }\left(\bar{y}^{*}\right) & =E_{10} V_{2 \bullet}\left(\bar{y}^{*}\right)+V_{10} E_{2 \bullet}\left(\bar{y}^{*}\right)=E_{1}, V_{2 \bullet}\left(\bar{y}^{*}\right) \\
& =p V_{2 \bullet}\left(\bar{y}^{*} \mid k_{1}, n_{1}^{\prime}\right)+(1-p) V_{2 \cdot}\left(\bar{y}^{*} \mid k_{2}, n_{2}^{\prime}\right) \\
& =(n-1) s^{2}\left[p a_{1}+(1-p) a_{2}\right]=\frac{1-f}{n} s^{2} .
\end{aligned}
$$

### 3.2 Two-Stage Sampling

The same general idea can be used at each stage of two-stage sampling to extend the BWO method to that situation. Here $\hat{\bar{Y}}=\sum_{1}^{n} M_{i} \bar{Y}_{i} / n \bar{M}_{0}$, and the usual estimate of $\operatorname{Var}(\hat{\bar{Y}})$ is

$$
\begin{equation*}
\operatorname{var}(\hat{\bar{Y}})=\frac{1-f_{1}}{n} s_{1}^{2}+\sum_{i=1}^{n} \frac{f_{1}\left(1-f_{2 i}\right)}{n m_{i}} s_{2 i}^{2} \tag{3.4}
\end{equation*}
$$

where $f_{1}=n / N, f_{2 i}=m_{i} / M_{i}$,

$$
s_{1}^{2}=\frac{1}{n-1} \sum_{i=1}^{n}\left(\frac{M_{i} \bar{y}_{i}}{\bar{M}_{0}}-\hat{\bar{Y}}\right)^{2}, \text { and } s_{2 i}^{2}=\frac{1}{n\left(m_{i}-1\right)} \sum_{j=1}^{m_{1}}\left(\frac{M_{i}}{\bar{M}_{0}}\right)^{2}\left(y_{i j}-\bar{y}_{i}\right)^{2} .
$$

Create a pseudo-population by replicating each cluster in the sample $\boldsymbol{k}_{\boldsymbol{t}}$ times and each unit within the $\boldsymbol{i}^{\text {ab }}$ cluster $\boldsymbol{k}_{2 i}$ times. From this pseudo-population the resample vector is obtained by sampling $\boldsymbol{n}^{\prime}$ clusters without replacement and $m_{1}^{\prime}$ units without replacement from each of the resampled clusters which was a replicate of the ith cluster. Let $\hat{\bar{Y}}^{*}=\frac{i}{n^{\prime}} \sum_{1}^{n^{\prime}} \frac{M_{1}}{\bar{M}_{0}} \bar{y}_{i}{ }^{\prime}$, with $\bar{y}_{i}^{*}=\frac{i}{m_{i}^{\prime}} \sum_{1}^{m_{i}^{\prime}} y_{i j}{ }^{\bullet \prime}$. Then

$$
\begin{equation*}
\operatorname{Var}_{.0}\left(\hat{Y}^{*} \cdot{ }^{*}\right)=\frac{k_{1}(n-1)}{\left(k_{1} n-1\right)} \frac{\left(1-f_{1}^{*}\right)}{n^{\prime}} s_{1}^{2}+\sum_{i=1}^{n} \frac{k_{21}\left(m_{i}-1\right)}{\left(k_{2} m_{i}-1\right)} \frac{\left(1-f_{2 i}^{*}\right)}{n^{\prime} m_{i}^{\prime}} s_{21}^{2}, \tag{3.5}
\end{equation*}
$$

where $f_{1}^{*}=n^{\prime} / k_{1} n$, and $f_{2}^{*}=m_{i}^{\prime} \mid k_{21} m_{i}$, for any choices of $k_{1}, k_{21}, n^{\prime}$, and $m_{i}^{\prime}$. One can choose $n^{\prime}$ and $k_{1}$ to satisfy

$$
f_{1}^{*}=f_{1}, \text { and } \frac{k_{1}(n-1)}{n^{\prime}\left(k_{1} n-1\right)}=\frac{1}{n},
$$

and $m_{1}^{\prime}$ and $k_{21}$ to satisfy

$$
f_{2 i}^{*}=f_{2 i} \text {, and } \frac{k_{2 l}\left(m_{i}-1\right)}{n^{\prime} m_{l}^{\prime}\left(k_{2 i} m_{i}-1\right)}=\frac{f_{1}}{n m_{i}}
$$

for each $i$. If this is done, $\operatorname{Var}^{\circ}\left(\hat{\bar{Y}}^{*} \cdot{ }^{\circ}\right)=\operatorname{var}(\hat{\bar{Y}})$.
To avoid the problem of non-integer resample sizes in this extension of the BWO method, one can randomize between bracketing integer values in the following manner. Let $k_{11}=\left\lceil k_{1}\right\rceil$ and $k_{12}=\left|k_{1}\right|$, where $k_{1}=\left[1-\left(1-f_{2}\right) / n\right] / f_{1}$; let $n_{1}^{\prime}=n-1$ and $n_{2}^{\prime}=n$; let $k_{2 i 1}=\left[k_{2 i} \mid\right.$ and $k_{212}=\left|k_{21}\right|$, where

$$
k_{2 i}=\frac{m_{i}-1-f_{1} f_{2 i} n^{\prime} / n}{m_{i} f_{1} f_{2 i} n^{\prime} / n}
$$

and let $m_{i 1}^{\prime}=\left\lfloor m_{i}^{\prime}\right\rfloor$ and $m_{i 2}^{\prime}=\left[m_{i}\right\rceil$, where

$$
m_{i}^{\prime}=\frac{m_{i}-1-f_{i} f_{2 i} n^{\prime} / n}{f_{1} n^{\prime} / n}
$$

Here $n^{\prime}$ represents a random variable which takes values $n_{1}^{\prime}$ and $n_{2}^{\prime}$ with probabilities described in the following, and $m_{i}^{\prime}$ and $k_{2 i}$ depend on $n^{\prime}$. Define

$$
p_{1}=\frac{\frac{1-f_{1}}{n(n-1)-a_{2}}}{a_{1}-a_{2}}, \text { and } \quad p_{2 i}=\frac{\frac{f_{1}\left(1-f_{2 i}\right)}{n m_{i}\left(m_{i}-1\right)}-b_{i 2}}{b_{i 1}-b_{i 2}},
$$

where $a_{j}=\left[k_{1 j}\left(1-f_{1 j}\right)\right] /\left[n_{j}^{\prime}\left(n k_{1 j}-1\right)\right], b_{1 j}=\left[k_{2 i j}\left(1-f_{2 i j}\right)\right] /\left[n^{\prime} m_{i j}^{\prime}\left(k_{2 i j} m_{i}-1\right)\right] f_{1 j}=n_{j}^{\prime} / n k_{1 j}$ and $f_{2 i j}=$ $m_{i j}^{\prime} /\left(k_{2 i j} m_{i}-1\right)$, for $j=1,2$. Note that $p_{2 i}$ is conditional on the $n^{\prime}$ chosen at the first stage of resampling. Also note that $k_{i l j}$ and $m_{i j}, j=1,2$, are conditional on the $n^{\prime}$ chosen in the first stage. One can show that $0 \leq p_{1} \leq 1$ and, given $n^{\prime}, 0 \leq p_{2 i} \leq 1$ for each $i$ if $n \geq 2$, and $m_{i} \geq 2$. It is shown in Sitter (1989) that, if this algorithm is used, $\operatorname{Var}_{\text {.. }}\left(\hat{\bar{Y}}^{* *}\right)=\operatorname{var}(\hat{\bar{Y}})$.

### 3.3 The Rao-Hartley-Cochran Method (RHC) For PPS-Sampling

A simple method of sampling with unequal probabilities without replacement was proposed by Rao, Hartley, and Cochran(1962). The sampling method is as follows:

1. Randomly partition the population of $N$ units into $n$ groups $\left\{G_{s}\right\}_{s=1}^{n}$ of sizes $\left\{N_{\varepsilon}\right\}_{\delta=1}^{n}$, respectively.
2. Draw one unit from each group with probability $z_{j} / Z_{g}$ for the $g^{\text {th }}$ group, where $z_{j}=x_{j} \mid X, Z_{s}=\sum_{\mu G_{k}} z_{p} x_{j}=$ some measure of the size of the $j$ th unit, and $X=\sum_{j=1}^{N} x_{j}$.

An unbiased estimator of $\bar{Y}$, the population mean, is $\hat{\bar{Y}}=\sum_{\varepsilon=1} w_{g} y_{g} / n$, where $w_{g}=f / \pi_{g}, \pi_{g}=z_{g} / Z_{g}$ is the probability of selection of the gth sampled unit, $f=n / N$ is the probability of selection under simple random sampling without replacement, and $y_{g}$ and $z_{g}$ denote the values for the unit selected from the gth group. Note, by definition $\sum_{g-1}^{n} Z_{g}=1$. An unbiased estimate of $\operatorname{Var}(\hat{\bar{Y}})$ is,

$$
\begin{equation*}
\operatorname{var}(\hat{\bar{Y}})=\frac{\left(\sum_{\xi=1}^{\infty} N_{g}^{2}-N\right)}{\left(N^{2}-\sum_{g-1}^{\infty} N_{g}^{2}\right)} \sum_{\varepsilon=1}^{n} Z_{g}\left(\frac{y_{g}}{N_{z \xi}}-\hat{\hat{Y}}\right)^{2} . \tag{3.6}
\end{equation*}
$$

To extend the BWO method to this situation, let $\hat{Y}_{R H C}=N \hat{\bar{Y}}_{\mathrm{RHC}}=\sum_{g=1}^{n} \quad Z_{g} y_{g} / z_{g}$. A BWO procedure for this situation is:

1. Replicate $\left(z_{g}, y_{g}\right), k_{g}=Z_{g} / z_{g}$ times for $g=1, \ldots, n$ to create a pseudo-population.
2. Randomly partition the pseudo-population of $N^{*}=\sum_{g=1}^{n} k_{g}$ units into $n^{*}$ groups, $\left\langle\Gamma_{\varepsilon}^{*}\right\}_{\varepsilon=1}^{n *}$, of sizes $\left\langle\left. N_{\varepsilon}^{*}\right|_{\varepsilon=1} ^{m}\right.$,
3. Randomly select one ( $z_{i}, y_{i}$ ) pair from each of the $n^{*}$ groups with probability $z_{i}^{*} \mid z_{8}^{*}$, where $z_{z}^{*}=\sum_{k r_{i} z_{i}^{*}}$, and let $\hat{\theta}^{*}=\hat{\theta}\left(\underline{z}^{*}, \underline{y}\right)$.
4. Repeat steps $1-3$ a large number of times, $B$, to get $\hat{\theta}_{1}^{\circ}, \ldots, \hat{\theta}_{B}^{\circ}$, and estimate $\operatorname{Var}(\hat{\theta})$ with $v_{\mathrm{bwo}}=E_{0}\left(\hat{\theta}^{\circ}-E_{.} \hat{\theta}^{\circ}\right)^{2}$, or its Monte Carlo approximation $v_{\mathrm{bwo}}=\sum_{i=1}^{B}\left(\hat{\theta}_{i}^{0}-\hat{\theta}_{0}^{0}\right)^{2} /(B-1)\left(E_{0} \hat{\theta}^{\circ}\right.$ and $\hat{\theta}_{(,)}^{0}$ can be replaced by $\hat{\theta}$ ).

If $\theta=\boldsymbol{Y}$ and $\hat{\theta}=\hat{\boldsymbol{Y}}_{\text {PBIC }}$, we have

$$
\begin{equation*}
\operatorname{Var}_{\theta}\left(\theta^{\circ}\right)=\left(\frac{\sum_{\varepsilon=1}^{\infty} N_{\varepsilon}^{\bullet 2}-N^{*}}{\sum_{\delta=1}^{n} N_{\varepsilon}^{2}-N}\right)\left(\frac{N^{2}-\sum_{\varepsilon=1}^{n} N_{\varepsilon}^{2}}{N^{*}\left(N^{*}-1\right)}\right) \operatorname{var}(\theta) . \tag{3.7}
\end{equation*}
$$

This holds since $\hat{Y}_{\text {RHC }}^{*}=\sum_{1}^{n *} z_{g}^{*} y_{g}^{*} / z_{g}^{*}$, and thus

$$
\operatorname{Var},\left(\mathscr{Y}_{R R C}^{*}\right)=\frac{\sum_{g^{*}+1}^{* *} N_{g}^{\circ 2}-N^{0}}{N^{*}\left(N^{*}-1\right)}\left(\sum_{i=1}^{N^{\circ}} \frac{y_{i}^{0^{2}}}{z_{i}^{*}}-Y^{\circ^{2}}\right),
$$

where $\boldsymbol{Y}^{\boldsymbol{0}^{2}}$ is the pseudo-population total. Noting the replication of data points in the pseudo-population the result follows directly.

This suggests two alternatives. The simplest is to merely correct the variance estimate by an appropriate multiplicative factor. The second is to choose $n^{*}$ and $N_{8}^{*}$ for $g=1, \ldots, n^{\circ}$ so that the product of the first two factors in (3.7) is close to one. In many situations it should be possible to find bracketing values of $\boldsymbol{n}^{*}$ and $N_{g}^{*}$ and randomize between them, but a closed form solution seems unavailable. For a more detailed discussion of possible choices see Sitter (1989).

## 4. A SIMULATION STUDY

This section describes a simulation study that was done to compare the performance of the above methods for variance estimation and confidence interval estimation for stratified random sampling without-replacement. Eight finite populations are considered. A subset of the results for populations 3 and 7 , which excludes the extended BWO method and some sample sizes, also appears in Sitter (1990). For completeness, the linearization and jackknife methods are included for non-linear statistics, and a method based on Woodruffs (1952) confidence intervals for quantiles and the jackknife are included for the median. For a detailed description of these three methods see Kovar, Rao, and Wu (1988). Slight modifications are made to correct for the sampling fraction, since they are considering with replacement sampling. In the Kovar, Rao, and Wu study, very little difference was found between the six jackknife definitions which they considered for stratified sampling with replacement. In our case, the sampling within each stratum is without replacement, but in view of their results it seemed unlikely that there would be a substantial difference in the performance of these six jackknifes. With this in mind, only one jackknife, defined below, was used (their $v_{r 2}$ ).

The estimators considered in this study are, the ratio, the regression coefficient, the correlation coefficient, and the median as defined below. Assuming the notation for stratified sampling of section 2, let $y_{k i}$ denote the characteristic of interest of the $i$ th observation from the $h$ th stratum, and let $x_{h i}$ denote a related concomitant variable. Let $t_{h i 1}=x_{h i}, t_{h i 2}=y_{h i j} t_{k i 3}=x_{k i} y_{h i}, t_{k i 4}=x_{h i}^{2}$, and $t_{h i s}=y_{h i}^{2}$. The sample estimates, of the corresponding population parameters, used were: 1) the ratio, $r=\bar{t}_{2} / \bar{t}_{1} ; 2$ ) the regression coefficient, $\left.b=\left(\bar{t}_{3}-\bar{t}_{1} \bar{r}_{2}\right) /\left(\bar{t}_{4}-\bar{r}_{1}^{2}\right) ; 3\right)$ correlation coefficient, $c=\left[\bar{t}_{3}-\bar{t}_{1} \overline{t_{2}}\right] /\left[\left(\overline{t_{4}}-\bar{t}_{1}^{2}\right)\left(\bar{t}_{3}-\bar{t}_{2}^{2}\right)\right]^{1 / 2}$; and 4) the median, $m$, with $p=1 / 2$; where $\bar{t}_{j}=\sum_{k=1}^{\ell} W_{h} \bar{p}_{h j}$ with $\bar{t}_{k j}=\sum_{j=1}^{*_{k}} t_{h i j} / n_{h}$ for $j=1, \ldots, 5$.

### 4.1 Confidence Intervals

For the linearization and jackknife methods, confidence intervals are obtained using normal theory, and are given by $\left[\hat{\theta}-z_{a n} \sqrt{v}, \hat{\theta}+z_{e n} \sqrt{v}\right]$, where $v$ is the corresponding variance estimate of the particular method, and $z_{a n}$
is the upper $\alpha / 2 \%$ cutoff for the standard normal distribution. For the various bootstrap methods, one can obtain a histogram of the $\boldsymbol{B}$ bootstrap estimates $\hat{\boldsymbol{\theta}}_{b}^{*}$. Confidence intervals can be constructed using these values by the percentile method (Efron 1982). This method is used for the median. For non-linear functions of means the bootstrap-t method is used. Here the bootstrap histogram of $t_{b}^{*}=\left(\hat{\theta}_{b}^{*}-\hat{\theta}\right) / \sqrt{v_{j}}$, where $v_{j}^{*}$ is a jackknife variance estimator applied to the resampled data, is used to obtain $t_{v}^{*}$ and $t_{i}^{*}$ the upper and lower $\alpha / 2 \%$ cutoff points for $t^{\prime \prime}$. The bootstrap- $t(1-\alpha) \%$ confidence interval is then given by $\left[\hat{\theta}-t_{U}^{*} \sqrt{v_{j}}, \hat{\theta}-t_{\mathcal{L}}^{*} \sqrt{v_{j}}\right]$, where $v_{J}$ is a jackknife variance estimate applied to the whole sample. Abramovitch and Singh (1985) demonstrated the superiority of the one-sided bootstrap-f intervals over the normal theory intervals for the iid case using asymptotic expansions.

### 4.2 Generating The Finite Populations

The finite populations used are based on one hypothetical population given in Hansen \& Tepping (1985), which was further looked at in Kovar, Rao \& Wu (1988), as well as, one hypothetical population given in Kovar, Rao, and Wu (1988) which itself was based on a population given in Hansen, Madow, and Tepping (1983). The Hansen \& Tepping (1985) population was put forth as an approximation to real populations encountered in the National Assessment of Educational Progress study. This population consisted of $L=32$ strata, with $\left(x_{k}, y_{k}\right)$ bivariate normal variates distributed $N_{2}\left(\mu_{x h}, \mu_{y k}, \sigma_{x h}, \sigma_{y k}, \rho\right)$ for $h=1, \ldots, L$, where $\rho$ is constant across strata. The $W_{h}$ 's range from 0.013 to 0.042 , and the stratum means and variances are such that the coefficient of variation of $\boldsymbol{x}$ and $y$ are approximately $10 \%$ and $30 \%$ respectively. The parameter settings, as well as the stratum weights, are given in Sitter (1989). Kovar, Rao, and Wu (1988) extended this hypothetical population to gamma distributions with the same mean and covariance structure in the following manner. Let $x_{n}$ be distributed $\operatorname{Gamma}(\alpha, \beta)$ with shape parameter $\alpha=\mu_{x h}^{2} / \sigma_{x h}^{2}$ and scale parameter $\beta=\sigma_{x k}^{2} / \mu_{x h}$ so that $x_{h}$ has mean and variance $\mu_{x h}$ and $\sigma_{x h}^{2}$. Using a linear regression model, $y_{h}$ was obtained from $x_{h}$ by $y_{h}=\left(\sigma_{y h} \rho / \sigma_{x h}\right) x_{h}+\epsilon_{h}$, where $\epsilon_{h}$ is normally distributed with mean $\mu_{y h}-\sigma_{y h} \mu_{x h} \rho / \sigma_{x k}$ and variance $\sigma_{y h}^{2}\left(1-\rho^{2}\right)$. So $\left(x_{h}, y_{h}\right)$ has joint distribution with the prespecified moments.

The finite populations were generated from the hypothetical populations, and then simulations were performed on the obtained finite populations. The methods under consideration include a number of different bootstraps, which are computer intensive. Also the stratum sampling fractions could be important in comparing the different methods. This suggests using a number of different stratum sample sizes on each finite population. With these points in mind it is clear that cost constraints will limit the number of populations that can be considered. Some of the general results of Kovar, Rao, \& Wu (1988) were used to choose hypothetical population parameters in an attempt to create finite populations which would emphasize differences in the performance of the various methods.

To compare the performance of the various methods for variance estimation and confidence intervals for $r, b$, and $c$, four finite populations were generated. Two were generated using parameter settings similar to the Hansen and Tepping population settings, but with their $\sigma_{x h}^{2}$ multiplied by 20. Each of these two populations consist of $N=800$ units divided into $L=16$ strata, with the strata sizes, $N_{h}$, ranging from 22 to 69 . Population 1 was generated using a bivariate normal distribution with stratum means between 45 and 95 , stratum standard deviations between 12 and 44 and $\rho=0.8$ (Actual settings given in Sitter 1989). Population 2 was generated using the gamma distribution and linear regression model as described above with the same parameters as Population 1. Populations 3 and 4 were generated using a bivariate normal distribution and similar parameter settings, but aggregated into fewer strata. These two populations consist of $N=760$ and $N=516$ units respectively divided into $L=6$ strata, with the strata sizes, $N_{h}$, ranging from 80 to 200 and 36 to 144 respectively. Thus Populations 1 , and 2 have a larger number of strata with fewer units per stratum than Populations 3 and 4.

To compare performance of the methods for variance estimation and confidence interval for the estimated median, $m$, four finite populations, numbered 5 though 8 , were generated. All four were generated using normal distributions in a similar fashion to the above, but with only $y_{h}$ values. Population 5 is similar to Population 1, consisting of $N=800$ units divided into $L=16$ strata. The $\mu_{y k}$, and $N_{h}$ settings used for Population 5 are the same as those for $y_{k}$ of Population 1, but the standard deviation of $y_{k}$ used was decreased by a factor of $1 / \sqrt{2}$. Population 6 is similar to Population 4, consisting of $n=516$ units divided into $L=6$ strata, and the $\mu_{y_{k}}$, and $N_{h}$ settings, but the standard deviation of $y_{k}$ used was decreased by a factor of 10 . Populations 7 and 8 each consist of $N=800$ units divided into $L=32$ strata. The parameter settings used for Population 7 were considered by Kovar, Rao, and Wu (1988), in a comparison of the performance of the rescaling bootstrap method to various other methods, including linearization, the jackknife, and Woodruff's method. They considered with replacement sampling with an infinite population size. These parameter settings were obtained by first generating a bivariate population, as described in Hansen, Madow, and Tepping (1983), where ( $x, y$ ) are highly correlated, then dividing the $y$ units into 32 strata using the $x$ values, and finally calculating the resulting $\mu_{y h}, \sigma_{y k}$, and $W_{A}$ values. Noting the parameter settings of Population 7, and considering their method of generation, it is clear that a finite population generated from these settings will have little or no overlap of values across strata. So, from the $W_{k}$ values it is clear that the resulting median will likely be near the center of stratum 6. Population 8 was generated from slightly different $N_{k}$ values to create a population with median near a stratum boundary. The resulting stratum means and standard deviations, as well as the $R, B, C$, and median, where applicable, for the above 8 finite populations are given in Sitter (1989).

### 4.3 Measures of Performance

To simplify the study, equal within stratum sample sizes, $n_{h}=n_{0}$ for $h=1, \ldots, L$ are used; so the total sample size is $n=n_{0} L$. For all the bootstrap variance and confidence interval estimation methods under consideration, $\hat{\theta}_{0}^{\circ}$.) is replaced by $\hat{\theta}$ in the definitions of their Monte Carlo approximations, and all methods are compared to the true MSE of the estimator of interest in the following manner. First, the true MSE was estimated by selecting 3000 stratified random samples without replacement, and using MSE $=\sum_{b=1}^{3000}\left(\hat{\theta}_{b}-\theta\right)^{2} / 3000$. The relative bias and stability of the variance estimators, and the tail error rates and standardized length of the confidence intervals were then obtained as follows.

The simulation consisted of selecting $S$ new stratified random samples, and for each sample calculating the various variance and confidence limit estimates. The relative bias of a particular variance estimator $v$ was then measured by bias $=\frac{\overline{\mathbf{v}}-\text { MSE }}{\text { MSE }}$, where $\overline{\mathbf{v}}$ is the average of the $S$ variance estimates. The relative stability of a variance estimator was estimated by stab $=\frac{\sigma_{v}}{M S E}$, where $\sigma_{v}^{2}=\sum_{s}\left(v_{v}-M S E\right)^{2} / S$. These variance estimators were then compared in terms of their relative bias and relative stability.

For comparing the various confidence interval estimators, the error rates in the upper and lower tails were compared to the nominal error rates of $5 \%$ and $10 \%$ in each tail. The standardized interval lengths of the methods were also compared. The upper and lower tail error rates and the standardized interval lengths were obtained using

$$
\begin{aligned}
\mathrm{U} & =\frac{\text { of samples with } \theta<\theta_{U_{s}}}{S}, \\
\mathrm{~L} & =\frac{\# \text { of samples with } \theta<\theta_{L s}}{S}, \\
\text { length } & =\frac{\sum_{s}\left(\theta_{U_{s}}-\theta_{L s}\right) / S}{2 \tau_{u / 2} \sqrt{\mathrm{MSE}}},
\end{aligned}
$$

where $\left(\theta_{L^{\circ}}, \theta_{U_{s}}\right.$ ) is the estimated confidence interval from the sth sample, and $z_{\kappa / 2}$ is the upper $\alpha / 2 \%$ cutoff point of a standard normal distribution.

For the nonlinear statistics $r, b$, and $c$, the bootstrap variance estimates and confidence intervals were based on $B=200$ bootstrap replicates and $S=1000$ simulation replicates were used. With $S=1000$ the empirical coverage probabilities in each tail pertaining to nominal levels 0.05 and 0.1 will be within 0.01 and 0.02 respectively $95 \%$ of the time. For the median, the bootstrap variance estimates and confidence intervals were based on $B=500$ bootstrap replicates and $S=500$ replicates were used. With $S=500$ the empirical coverage probabilities in each tail pertaining to the nominal levels 0.05 and 0.1 will be within 0.01 and 0.02 respectively $85 \%$ of the time.

### 4.4 Non-Linear Statistics

Recall from section 4.2 that populations 1 through 4 were used to compare the methods on non-linear functions of means: $\boldsymbol{r}$, the ratio; $\boldsymbol{b}$, the regression coefficient; and $\boldsymbol{c}$, the correlation coefficient. For populations 1 and 2 , stratum sample sizes of 5,6 , and 7 were considered for each, while for populations 3 and 4 stratum sample sizes of 20 and 12 were used respectively.

Table 2 gives the average absolute difference between the observed and nominal error rates, both one-tailed and two-tailed confidence intervals, where the average is over the four populations and the various stratum sample sizes considered (ie. $\dot{L}_{5 k}=\sum_{i}\left|L_{i}-5 \%\right|$ and $L \mp U_{5 k}=\sum_{i}\left|L_{i}+U_{i}-10 \%\right|$ ). Table 2 also gives the average absolute relative bias in percentage and the average relative stability of the variance estimates. All of the methods perform well for these populations. Some minor comments follow:

1) Variance estimates: a) For the ratio, the linearization and jackknife methods perform equivalently, and slightly better than the bootstrap methods in terms of both relative bias and stability; b) There is little to choose between the bootstrap methods, and the jackknife method, for $b$ and $c$; and $c$ ) the linearization method yields higher relative bias, but slightly lower relative stability for $b$ and $c$.
2) Confidence intervals: a) For the ratio, the methods perform equally well. As the non-linearity increases $(r-b-c)$, the bootstrap methods track the one-tailed error rates better than do the linearization and jackknife methods.

Kovar, Rao, \& Wu (1988) observed a similar trend in their study; and b) As non-linearity increases, choosing the resample sizes in the mirror-match method to match the second-order term of the Edgeworth expansion of the mean seems to become more beneficial. Though the same is not true for the rescaling method, this choice of resample size performs much better than was shown in the Kovar, Rao, \& Wu (1988) study, where choosing the resample size in the rescaling method to match the Edgeworth expansion of the mean yielded very poor performance, which is contrary to what one would expect.

Table 2: Average Absolute Error in the Upper and Lower Tails, Relative Bias, and Relative Stabilities Over Considered Populations (ratio, regr., and corr.)

|  | Avg. Abs. Dev. from Nominal error rate |  |  |  |  |  | \|\%bias ${ }^{\text {a }}$ | stab ${ }^{\text {b }}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | 5\% |  |  | 10\% |  |  |  |  |
| Method | $\underline{L}$ | U' | $L+U$ | $\underline{L}$ | $\tilde{U}$ | $L+U$ |  |  |
|  | Ratio |  |  |  |  |  |  |  |
| lin | 59 | 58 | . 84 | 39 | 1.05 | 1.19 | 2.86 | . 22 |
| jack | 59 | 56 | . 83 | . 42 | 1.05 | 1.20 | 2.88 | . 22 |
| ext bwo | . 48 | . 64 | . 84 | . 36 | . 89 | 1.15 | 3.24 | . 25 |
| $\mathrm{m}=\mathrm{m}^{\text {e }} 1$ | . 40 | . 71 | . 86 | . 44 | . 95 | 1.09 | 3.38 | . 25 |
| $\mathrm{m}-\mathrm{m} 2$ | . 45 | 52 | . 65 | . 34 | . 90 | . 89 | 3.21 | . 24 |
| rst ${ }^{1}$ | . 48 | . 84 | 1.14 | . 34 | 1.10 | 1.29 | 3.40 | . 24 |
| [8 2 | . 50 | 55 | . 65 | . 31 | 1.14 | 1.08 | 3.43 | . 24 |
|  | Regression |  |  |  |  |  |  |  |
| lin | . 84 | 1.33 | 1.96 | 1.24 | 1.14 | 1.68 | 5.05 | . 30 |
| jack | . 70 | . 91 | 1.04 | 1.10 | . 74 | 1.19 | 3.61 | . 34 |
| ext bwo | . 89 | . 60 | 59 | 1.24 | . 67 | 1.21 | 3.41 | . 33 |
| m-ml | . 81 | . 41 | . 70 | 1.24 | . 80 | 1.56 | 3.31 | . 32 |
| $\mathrm{m}-\mathrm{m} 2$ | . 84 | . 41 | . 82 | 1.14 | . 70 | 1.16 | 3.34 | . 32 |
| 51 | 1.06 | . 32 | . 99 | 1.15 | . 84 | 1.49 | 3.11 | . 32 |
| rs 2 | . 84 | . 30 | . 71 | 1.31 | . 79 | 1.25 | 3.25 | . 32 |
|  | Correlation |  |  |  |  |  |  |  |
| lin | 2.44 | 1.11 | 1.58 | 3.10 | 1.51 | 2.01 | 4.06 | . 38 |
| jack | 1.78 | 1.33 | . 60 | 2.23 | 1.79 | 1.04 | 3.41 | . 42 |
| ext bwo | 1.15 | 1.18 | . 55 | 1.46 | 1.21 | . 85 | 3.40 | . 42 |
| m-m1 | . 96 | 1.07 | . 44 | 1.54 | 1.50 | . 64 | 2.90 | . 42 |
| m-m 2 | . 83 | 1.30 | . 60 | 1.40 | 1.37 | . 85 | 2.48 | . 40 |
| rs1 | 1.01 | 1.03 | . 76 | 1.75 | 1.43 | 1.32 | 2.34 | . 41 |
| rs 2 | . 95 | 1.05 | . 80 | 1.50 | 1.10 | 1.00 | 258 | . 42 |

- average ( $\mid \%$ rel. bias |) over populations and strata sample size considered.
- average (rel. stability).
- m-m 1: $n_{h}^{\prime}=f_{h} n_{h}, m-m 2: n_{h}^{\prime}=1$

4 Rescaled 1: resample size chosen to match 3rd moments, Rescale 2: resample size $=n_{n}-I$

### 4.5 The Median

The four populations considered for the median, described in section 4.2 , group naturally into populations 5 and 6 , and populations 7 and 8 . Populations 7 and 8 represent the situation where stratification is done using a highly correlated concomitant variable. For populations 5 and 6 stratum sample sizes of 5 and 6 , and 12 were considered, respectively. For populations 7 and 8 , stratum sample sizes of 4 through 7 were considered for both.

Table 3 summarizes the results for population 6 (population 5 gave qualitatively similar results). It gives the observed one-tailed and two-tailed error rates, and the standardized lengths for nominal levels of $5 \%$ and $10 \%$, as well as, the relative bias in percentage and the relative stability for the variance estimates. In concurrance with theoretical results, the jackknife performs poorly. The Woodruff based variance estimate's performance depends on the choice of $\alpha$, but for this population, it outperforms the bootstrap methods for a fairly wide range. In terms of confidence intervals, no clear winner is evidenced. The bootstrap methods overstate the coverage probabilities and have shorter standardized length than the Woodruff intervals, which understate the coverage probabilities. It should be noted that the bootstrap methods do not track the one-tailed error rates well.

Table 3: Population 6: Variances and C.I.'s For The Median

| Method | Nominal error rate |  |  |  |  |  | Std. length |  | \% bias | stab |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $5 \%$ |  |  | 10\% |  |  |  |  |  |  |
|  | L | U | $\mathrm{L}+\mathrm{U}$ | L | U | L + U | 5\% | 10\% |  |  |
| wood | 3.8 | 4.8 | 8.6 | 8.2 | 7.4 | 15.6 | 0.96 | 1.02 |  |  |
| $\alpha=.01$ |  |  |  |  |  |  |  |  | 22.7 | 0.64 |
| $\alpha=.025$ |  |  |  |  |  |  |  |  | 7.4 | 0.59 |
| $\alpha=.05$ |  |  |  |  |  |  |  |  | 6.6 | 0.63 |
| $\alpha=.1$ |  |  |  |  |  |  |  |  | 2.7 | 0.68 |
| $\alpha=.2$ |  |  |  |  |  |  |  |  | 17.8 | 0.88 |
| jack | 13.8 | 12.4 | 26.2 | 18.2 | 15.2 | 33.4 | 1.39 | 1.39 | 234.8 | 6.02 |
| ext bwo | 11.4 | 3.4 | 14.8 | 16.0 | 6.6 | 22.6 | 0.92 | 0.92 | 17.2 | 0.85 |
| $m-m^{4} 1$ | 11.0 | 2.8 | 13.8 | 15.6 | 6.6 | 22.2 | 0.96 | 0.92 | 22.2 | 0.85 |
| $m-m 2$ | 7.2 | 4.8 | 12.0 | 14.4 | 8.2 | 22.6 | 0.96 | 0.93 | 30.0 | 0.95 |
| [s'1 | 10.2 | 4.0 | 14.2 | 15.0 | 6.6 | 21.6 | 0.94 | 0.94 | 25.0 | 0.87 |
| 152 | 10.6 | 3.8 | 14.4 | 15.6 | 7.2 | 22.8 | 0.96 | 0.99 | 38.7 | 1.00 |

Table 4: Variances For The Median

| $n_{0}=$ | Population 7 |  |  |  | Population 8 |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | 5 |  | 7 |  | 5 |  | 7 |  |
| Method | bias | stab | bias | stab | bias | stab | bias | stab |
| wood |  |  |  |  |  |  |  |  |
| $\alpha=.01$ | 220.5 | 2.46 | 42.7 | 0.65 | 612.4 | 6.75 | 900.7 | 9.69 |
| $\alpha=.025$ | -9.2 | 0.44 | -28.5 | 0.46 | 570.7 | 6.27 | 550.7 | 7.92 |
| $\alpha=.05$ | 18.9 | 0.59 | -6.4 | 0.47 | 409.9 | 5.58 | 751.8 | 10.58 |
| $\alpha=.1$ | 68.8 | 1.05 | 32.9 | 0.74 | 623.9 | 8.23 | 1523.2 | 21.18 |
| $\alpha=.2$ | -27.5 | 0.70 | -41.7 | 0.67 | 1095.6 | 14.10 | 1523.2 | 21.18 |
| jack | 83.3 | 3.08 | 77.6 | 2.79 | 12.2 | 2.13 | 20.5 | 1.62 |
| ext bwo | 9.1 | 0.86 | 16.0 | 0.87 | 25 | 1.32 | 19.1 | 1.17 |
| $m-m^{2}$ | 14.2 | 0.84 | 8.5 | 0.81 | 18.6 | 1.39 | 5.9 | 1.01 |
| $\mathrm{m}=\mathrm{m} 2$ | 23.2 | 0.95 | 18.2 | 0.79 | 3.0 | 1.26 | -35.7 | 0.76 |
| rsi | 38.1 | 1.19 | 960.6 | 13.15 | 26.3 | 0.83 | large | large |
| ${ }_{5} 5$ | 8.2 | 0.87 | 1.1 | 0.68 | 144.2 | 2.49 | large | large |

- bias given in \%
- mirror-math 1: $n_{h}^{\prime}=f_{h} n_{h}$ mirror-match $2: n_{h}^{\prime}=1$
* rescale 1: resample size chosen to match 3 rd moments, rescale 2 : resample size $=n_{0}-1$

Table 5: Population 7: C.I's For The Median

| Methods | Nominal error rate in each tail |  |  |  |  |  | Std. length |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | 5\% |  |  | 10\% |  |  | $1-\alpha$ |  |
|  | L | U | $\mathrm{L}+\mathrm{U}$ | L | U | $L+U$ | 0.8 | 0.9 |
| $n_{0}=5$ |  |  |  |  |  |  |  |  |
| wood | 6.0 | 2.2 | 8.2 | 5.8 | 17.6 | 23.4 | 1.26 | 0.77 |
| jack | 14.8 | 6.2 | 19.0 | 14.8 | 13.4 | 28.2 | 1.09 | 1.09 |
| ext bwo | 5.8 | 7.0 | 12.8 | 5.8 | 17.6 | 23.4 | 0.76 | 0.77 |
| $m=m^{* 1}$ | 5.8 | 2.2 | 8.0 | 5.8 | 15.4 | 21.2 | 0.83 | 1.04 |
| $m-m^{2}$ | 5.8 | 2.2 | 8.0 | 5.8 | 13.6 | 19.4 | 0.83 | 0.85 |
| $\mathrm{rs}^{\text {b }} 1$ | 7.2 | 17.6 | 24.8 | 7.2 | 17.6 | 24.8 | 0.81 | 1.00 |
| rs 2 | 9.0 | 6.2 | 15.2 | 9.2 | 6.2 | 15.2 | 0.69 | 0.89 |
| $\pi_{0}=7$ |  |  |  |  |  |  |  |  |
| wood | 0.2 | 2.6 | 2.8 | 6.4 | 19.0 | 25.4 | 1.11 | 0.69 |
| jack | 18.6 | 8.0 | 26.6 | 21.6 | 8.0 | 29.6 | 1.06 | 1.06 |
| ext bwo | 9.0 | 2.0 | 11.0 | 9.0 | 2.0 | 11.0 | 0.77 | 0.98 |
| $m-m 1$ | 9.0 | 2.0 | 11.0 | 9.0 | 12.8 | 21.8 | 0.76 | 0.96 |
| $m-m 2$ | 1.8 | 2.2 | 4.0 | 9.0 | 15.8 | 23.8 | 1.02 | 0.73 |
| $\text { rs } 1$ | 2.4 | 63.0 | 65.4 | 5.2 | 91.0 | 97.2 | 1.02 | 0.20 |
| rs 2 | 2.6 | 10.2 | 12.8 | 11.2 | 11.8 | 23.0 | 0.83 | 0.74 |

Table 4 summarizes the variance estimate results for populations 7 and 8. It illustrates the greater robustness of the extended BWO of section 2.3 and the mirror-match methods to different stratifications, compared to the Woodruff based and the rescaling methods; the jackknife again performs poorly. It should be noted that though Kovar, Rao, \& Wu (1988) include the median in their simulation study, the rescaling method, as introduced by Rao \& Wu (1988), is only applicable to functions of means.

Table 5 summarizes the confidence interval results for population 7 with stratum sample sizes 5 and 7 . For this population, the mirror-match method performed the best.

## 5. CONCLUSIONS

The mirror-match and BWO methods perform better on average over the situations considered here. It is not surprising that the rescaling method does well for stratified random sampling with the non-linear functions of means considered. In such a simple situation a rescaling of i.i.d. resampling adequately approximates without replacement sampling. One would expect that as the estimators and the sampling design become more complex, methods which better mimic the original sampling will perform better. From this viewpoint the mirror-match method, and perhaps even more so, the extended BWO method are promising. The results for the median support this; especially where stratification is done by a highly correlated concomitant variable. Clearly, both theoretical and empirical study of these methods for more complex sampling plans is needed, and is being investigated.

## REFERENCES

Abramovitch, L., and Singh, K. (1985). Edgeworth corrected pivotal statistics and the bootstrap. Annals of Statistics, 13, 116-132.

Bickel, P.J., and Freedman, D.A. (1984). Asymptotic normality and the bootstrap in stratified sampling. Annals of Statistics, 12, 470-482.

Efron, B. (1982). The Jackknife, the Bootstrap, and Other Resampling Plans, Philadelphia, Society for Industrial and Applied Mathematics.

Gross, S. (1980). Median estimation in sample surveys. Proceedings of the Section on Survey Research Methods. American Statistical Association, 181-184.

Hansen, M.H., Madow, W.G., and Tepping, BJ. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. Joumal of the American Statistical Association, 78, 776-807.

Hansen, M.H., and Tepping, B.J. (1985). Estimation of variance in NAEP. Unpublished Manuscript.
Kovar, J.G., Rao, J.N.K., and Wu, C.FJ. (1988). Bootstrap and other methods to measure errors in survey estimates. Canadian Joumal of Statistics, 16, Supplement, 25-45.

McCarthy, PJ., and Snowden, C.B. (1985). The bootstrap and finite population sampling. Vital and Health Statistics (Ser. 2, No. 95), Public Health Service Publication 85-1369, Washington, DC: U.S. Government Printing Office.

Rao, J.N.K., Hartley, H.O., and Cochran, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. J. Roy. Statistical Social Series, B, 24, 482-491.

Rao, J.N.K., and Wu, C.F.J. (1988). Resampling inference with complex survey data. Joumal of the American Statistical Association, 83, 231-241.

Sitter, R. (1989). Resampling procedures for complex survey data. Ph.D. Thesis, Department of Statistics and Actuarial Science, University of Waterloo.

Sitter, R. (1990). A resampling procedures for complex survey data. Technical Report 149, Department of Mathematics and Statistics, Carleton University.

Woodruff, R.S. (1952). Confidence intervals for medians and other position measures. Joumal of the American Statistical Association, 47, 635-646.

# SIMPLE CORRECTION FOR THE BIAS OF TRADITIONAL VARIANCE ESTIMATORS 

Y. Leblond ${ }^{\prime}$


#### Abstract

SUMMARY It is widely known that traditional variance estimators of non-linear estimators of the characteristics of a population are generally biased when relatively small samples are used. For example, the estimators obtained with the Taylor linearization method and the jackknife estimator respectively tend to underestimate and overestimate the variance of commonly used estimators. We are proposing a simple method whereby the bias of traditional variance estimators can be reduced. This method consists of multiplying the traditional variance estimator by an adjustment factor that depends only upon the size of the sample n, and additional information. This correction factor tends toward 1 when n increases, and thus the modified variance estimator obtained is asymptotically equivalent to the traditional estimator. This factor is relatively simple to justify and easy to calculate. It is obtained by using a modelassisted method. The application of this method to the variance estimator of the ratio estimator is discussed, as are the results of a simulation study that shows the robustness of the method.


KEY WORDS: Ratio estimator; Bias; Variance; Taylor; Jackknife; Superpopulation model.

## 1. INTRODUCTION

In the recent literature on the estimation of sampling variance, several authors have tackled the problem of obtaining a variance estimator whose mean square error (MSE) is minimal in a given class of estimators. The work of $\mathrm{Wu}(1982,1985)$, Wu and Deng (1983), and Deng and Wu (1987) on the estimation of variance of ratio and regression estimators in a simple random or simple stratified design is particularly noteworthy. These authors compared the different variance estimators, considering only the variability introduced by the random selection of the sample. They also compared the bias of different estimators using a superpopulation model. Royall and Cumberland (1978, 1981, 1981a), and Valliant (1987) also considered the problem of variance estimation when studying the properties of different estimators using certain superpopulation models.

All of this work showed that it is difficult to determine the "best" variance estimator without making certain modelling hypotheses concerning the population under consideration. However, these hypotheses are rarely verified in practice, which means that we have to find new, more robust methods of variance estimation than those currently in use.

The problem considered here involves correcting a given variance estimator in order to reduce its bias under the sampling plan. Our purpose here is not to propose a variance estimator with a lower bias than its peers, or that has the lowest bias in a given class of variance estimators, but rather to define a method of reducing the bias of the estimator used, whatever it may be.

To do this, we are attempting to define a modified variance estimator $\hat{V}_{w}$ of a traditional variance estimator $\hat{V}$ with the following properties:

[^19]1. the bias of $\hat{V}_{\boldsymbol{w}}$ has a lower absolute value than that of $\hat{\boldsymbol{V}}$,

2 - $\quad \hat{V}_{M}$ is asymptotically equal to $\hat{\boldsymbol{V}}$,
3 - $\quad \hat{V}_{w}$ is simple, and easy to calculate and justify.
In section 2 we will describe the method used to obtain the modified variance estimators of three traditional variance estimators of the ratio estimator. In Section 3 we will present some results of simulations demonstrating the soundness of our bias reduction method.

## 2. RATIO ESTIMATOR

Let us take a simple random sampling plan (SI) of $n$ units out of $N$. Let us assume that $f=n / N$ and define the ratio estimator of the mean of the population $\bar{Y}$ by:

$$
\bar{y}_{Q}=\bar{x} \frac{\bar{x}}{\bar{y}}
$$

where $\bar{X}=\sum_{i=1}^{N} \frac{x_{1}}{N}, \quad \bar{y}=\sum_{i=1}^{n} \frac{y_{i}}{n}$ and $\bar{x}=\sum_{i=1}^{n} \frac{x_{i}}{n}$.

The estimation of the mean square error (MSE) is considered to be $\operatorname{MSE}\left(\bar{y}_{Q}\right)=E\left(\bar{y}_{Q}-\bar{Y}\right)^{2}$, where $E\left(\left(^{\cdot}\right)\right.$ is the expectation in relation to plan SI. Let us consider three currently used variance estimators of $\bar{y}_{Q}$ : the estimator obtained with Taylor's linearization $\left(\hat{V}_{T}\right)$, the Taylor's estimator with Royall's correction ( $\hat{V}_{T 2}$ ), and the jackknife estimator ( $\hat{V}_{\rho}$ ), which are defined respectively by:

$$
\begin{aligned}
& \hat{V}_{T}=\frac{1-f}{n} \sum_{i=1}^{n} \frac{\left(y_{i}-\hat{R} x_{i}\right)^{2}}{n-1}, \\
& \hat{V}_{T 2}=\left(\frac{\bar{X}}{\bar{x}}\right)^{2} \hat{V}_{T} \\
& \hat{V}_{J}=(1-f) \frac{n-1}{n} \sum_{j=1}^{n}\left(\hat{R}_{j}-\hat{R}_{(,)}\right)^{2}
\end{aligned}
$$

where $\hat{R}=\bar{y} / \bar{x}, \hat{R}_{(j)}=\sum_{l=j}^{n} y_{i} / \sum_{i \neq j}^{n} x_{i}$ and $\hat{R}_{(\cdot)}=\sum_{j=1}^{n} \hat{R}_{(j)} / n$.
Let $\hat{V}$ be one of these three variance estimators. Leblond (1989) has shown that, under certain conditions, the bias of $\hat{V}$ is

$$
\begin{aligned}
\operatorname{Biar}(\hat{V}) & =E\left\{\hat{V}-\operatorname{MSE}\left(\bar{y}_{Q}\right)\right\} \\
& =B_{2}+O\left(n^{-3}\right)
\end{aligned}
$$

where $B_{2}=O\left(n^{-2}\right)$ is a complex expression dependent upon $y_{k}$, and where $O(\cdot)$ indicates that the argument is limited. Furthermore, Leblond (1989) has shown that for $\hat{V}_{T}, B_{2}$ is always negative, and for $\hat{V}_{y}, B_{2}$ is always positive.

We propose a method for correcting $\hat{V}$ in order to reduce the magnitude of the bias to a term $O\left(n^{-3}\right)$. This method involves obtaining an estimator $\hat{B}_{2}$ of $B_{2}$ and defining the modified variance estimator $\hat{V}_{M}$ as follows:

$$
\begin{equation*}
\hat{V}_{M}=\hat{V}-\hat{B}_{2} . \tag{1}
\end{equation*}
$$

However, due to its complex form, $B_{2}$ is difficult to estimate. In order to simplify the estimation of $B_{2}$, we will consider a model assisted approach. The model used is one generally considered to underlie the ratio estimator:

$$
E_{\xi}\left(y_{i}\right)=\beta x_{i}, y_{\xi}\left(y_{i}\right)=\sigma^{2} x_{i} \text { el } \operatorname{cov}_{\xi}\left(y_{i}, y_{j}\right)=0, i \neq j .
$$

With this model-based approach, we will consider the predicted bias of $\hat{\boldsymbol{V}}$, or

$$
E_{\ell}[\operatorname{Bias}(\hat{V})]=E_{\xi}\left(B_{2}\right)+O\left(n^{-3}\right) .
$$

The advantage of using the predicted bias notion lies in the simple form of $\boldsymbol{E}_{\boldsymbol{\xi}}\left(\boldsymbol{B}_{2}\right)$. In fact, $\boldsymbol{E}_{\xi}\left(\boldsymbol{B}_{2}\right)=\boldsymbol{K}_{\mathbf{g}} \sigma^{2}$ where $K_{x}=0\left(n^{-2}\right)$ depends only upon the values of $x$ (see Table 1).

Table 1: Asymptotic Bias (to the term $O\left(n^{-3}\right)$ ) of the variance estimators.

| Bias | $\hat{V}_{T}$ | $\hat{V}_{\pi}$ | $\hat{V}_{J}$ | Estimation |
| :---: | :---: | :---: | :---: | :---: |
| $\mathbf{B}_{2}$ | $<0$ | $?$ | $>0$ | complex |
| $\mathbf{E}_{f}\left(\mathbf{B}_{2}\right)$ | $-2 \mathrm{M}_{x}$ | $-\mathrm{M}_{\mathrm{x}}(1+\mathrm{f})$ | $\mathrm{M}_{\mathrm{x}}(1-\mathrm{f})$ | simple |

where $M_{x}=\frac{(1-f)}{n(n-1)} \frac{S^{2} x}{\bar{X}} \sigma^{2}$.
Thus, we obtain $\hat{V}_{\boldsymbol{m}}$ by assuming that, in (1), $\hat{B}_{2}$ is equal to the estimator of $E_{\xi}\left(B_{2}\right)$. In order to obtain this estimator, it is sufficient to estimate $\hat{\sigma}^{2}$. However, it is necessary to estimate $\hat{\sigma}^{2}$ from $\hat{\sigma}^{2}$ such that

$$
\begin{equation*}
\hat{\sigma}^{2}=C_{x} \hat{y} \tag{2}
\end{equation*}
$$

and

$$
\begin{equation*}
E_{q}\left(\hat{\sigma}^{2}\right)=\sigma^{2}+O_{p}\left(n^{-1}\right) \tag{3}
\end{equation*}
$$

where $C_{x}$ is a term that depends only upon the values of $\boldsymbol{x}$, and $O_{p}$ indicates that the argument is limited as probability in relation to plan SI. Condition (2) makes it possible to obtain a simple modified variance estimator, $\hat{V}_{w}$, which is equal to the multiplication factor $x$, the traditional estimator. In order to determine $C_{x}$ for each of the three estimators under consideration, we can demonstrate the following result:

$$
E_{\xi}\left[\left(\frac{\bar{x}}{\bar{x}}\right)^{s} \hat{V}_{T}\right]=\frac{\sigma^{2}}{C_{x(g)}}+O_{p}\left(n^{-2}\right)
$$

where

$$
C_{x(g)}=\frac{(n-1)}{\bar{x}(1-f)}\left(\frac{\bar{x}}{\bar{x}}\right)^{x} .
$$

For $\hat{V}=\left(\frac{\bar{x}}{\bar{x}}\right)^{8} \hat{V}_{T}$, we then obtain $C_{x}=C_{x(\xi)}$, which makes it possible to obtain $\hat{\sigma}^{2}$ as required by (2) and verifying (3).
$C_{x}$ is thus obtained for $\hat{V}_{T}$ and $\hat{V}_{r_{2}}$ by assuming that $g=0$ and $g=2$ in $C_{x(\mathcal{S})}$ respectively. The value of $C_{x}$ for $\hat{V}_{J}$ is the same as that for $\hat{V}_{\pi}$, because

$$
\hat{V}_{72}=\hat{V}_{J}+O_{p}\left(m^{-2}\right) .
$$

If we assume that $\hat{B}_{2}=C_{x} K_{x} \hat{V}$ in (1), we obtain

$$
\hat{V}_{m}=\hat{V}-K_{s} C_{x} \hat{V}=A_{x} \hat{V},
$$

where $A_{x}=1-K_{x} C_{x}$ is known as the correction factor. $A_{x}$ converges in probability toward 1 . Thus, the modified estimator converges asymptotically toward its original version. The explicit form of the modified variance estimators of the three estimators discussed here is thus obtained by substituting the appropriate values of $K_{x}$ and $C_{x}$ :

$$
\begin{aligned}
& \hat{V}_{M T}=\left(1+\frac{2}{n}\left(C V_{x}\right)^{2} \frac{\bar{x}}{\bar{x}}\right) \hat{V}_{T} \\
& \hat{V}_{M T 2}=\left(1+\frac{1+f}{n}\left(C V_{x}\right)^{2} \frac{\bar{x}}{\bar{X}}\right) \hat{V}_{T 2} \\
& \hat{V}_{M J}=\left(1-\frac{1-f}{n}\left(C V_{x}\right)^{2} \frac{\bar{x}}{\bar{X}}\right) \hat{V}_{J}
\end{aligned}
$$

where $C V_{x}=S_{x} / \bar{X}$ is the coefficient of variation of $x$ in the population. Please note that the correction factor for $\hat{V}_{M T}$ is higher than 1, and that for $\hat{V}_{\boldsymbol{M}}$, this factor is lower than 1, which should, in accordance with the results shown in Table 1, allow a reduction in the right direction in the bias.

Furthermore, in cases where $x_{i}=1$ for any $i$, the three modified variance estimators equal

$$
\hat{v}(\bar{y})=\frac{1-f}{n} \sum_{i=1}^{n} \frac{\left(y_{1}-\bar{y}\right)^{2}}{n-1} .
$$

Thus, in such a situation, the modified variance estimator is equal to the generally used unbiased estimator. This is a sign that the correction made is well founded.

Finally, we should point out that bias $B_{2}$ of order $n^{-2}$ of $\hat{V}$ is not completely eliminated by the correction made. In fact, this correction is carried out by subtracting the unbiased asymptotic estimator of $E_{\ell}\left(B_{2}\right)$ from $\hat{V}$. Thus, for any given population, $E_{\xi}\left(B_{2}\right)$ and $B_{2}$ are not necessarily equal (unless the population completely conforms with the model $\xi$ used). Nevertheless, the results of the simulations discussed in the next section show that, in most of the cases studied, the correction factor allows a substantial reduction in the bias of the traditional variance estimators.

## 3. RESULTS OF THE SIMULATIONS

The purpose of the simulation study was to compare the modified and traditional variance estimators discussed in Section 2. The relative bias and mean square error (MSE) in some superpopulation models were used as the criteria for comparison. In order to do this, we considered the populations generated from the following general superpopulation model:

$$
\begin{equation*}
E_{\xi}\left(y_{j}\right)=\alpha+\beta x_{i}+y x_{i}^{2}, V_{\xi}\left(y_{i}\right)=\sigma^{2} x_{i}^{k} \text { and } \operatorname{cov}_{\xi}\left(y_{i}, y_{j}\right)=0, i \neq j \tag{4}
\end{equation*}
$$

Each determination of the parameter $\theta=\left(\alpha, \beta, y, \sigma^{2}, \mathcal{g}\right)$ led to a superpopulation model. We generated 30 finite populations of size $N=100$ for each of these superpopulations, using a gamma density function. The use of a gamma density function is appropriate because it has an asymmetrical distribution, like most of the populations studied in economic surveys. We generated 2,000 samples in accordance with simple random selection of fixed size without replacement for each of these 30 populations; where, for each sample, the ratio estimator and the various variance estimators were calculated. We considered the sample sizes $n=5,10,15,20$.

Let $\bar{y}_{\text {pre }}$ and $\hat{V}_{r m}$ be, respectively, the ratio estimator and a variance estimator calculated on the basis of an $r^{e}$ sample of the $s^{*}$ population of the $t^{*}$ superpopulation; and let $\bar{Y}_{s}$ be the mean of the $s^{*}$ population of ther ${ }^{*}$ superpopulation.

The MSE of the ratio estimator, the bias, and the MSE of $\hat{V}$ for the $s^{*}$ population of the $t^{*}$ superpopulation are defined respectively by:

$$
\begin{gathered}
\operatorname{MSE}_{r=}\left(\bar{y}_{Q}\right)=\sum_{r=1}^{2000} \frac{\left(\bar{y}_{Y_{r x s}}-\bar{Y}_{z}\right)^{2}}{2000}, \\
\operatorname{Bias}_{x}(\hat{\eta})=\sum_{r=1}^{2000}\left[\hat{V}_{r m}-\operatorname{MSE}_{r=}\left(\bar{y}_{Q}\right)\right] / 2000 \\
\operatorname{Var}_{s t}(\hat{V})=\sum_{r=1}^{2000}\left[\hat{\nu}_{r s r}-\left(\sum_{r=1}^{2000} \hat{\nu}_{r s r} / 2000\right)\right]^{2} / 1999 .
\end{gathered}
$$

Subsequently, the relative bias and the MSE of $\hat{V}$ for each superpopulation t are determined by

$$
\begin{gathered}
\operatorname{Brel}_{\mathrm{g}}=\sum_{s=1}^{30} \frac{\operatorname{Bias}_{\mathrm{k}}(\hat{V}) / \operatorname{MSE}_{\mathrm{k}}\left(\overline{\mathrm{Y}}_{\mathrm{Q}}\right)}{30}, \\
\operatorname{MSE}_{\mathrm{t}}(\hat{V})=\sum_{s=1}^{30} \frac{\left(\operatorname{Bias}_{\mathrm{m}}(\hat{\mathrm{~V}})\right)^{2}+\operatorname{Var}_{\mathrm{k}}(\hat{V})}{30} .
\end{gathered}
$$

These last quantities represent an approximation of the expected values of the relative bias and the MSE of the variance estimators under consideration, using the superpopulation model $t$. They serve as a basis for comparison between the various estimators.

The populations generated from the superpopulation model defined in (4) with $\alpha=0, \gamma=0$ and $g=1\left(\beta>0, \sigma^{2}>0\right)$ correspond to what we call favourable populations for modified variance estimators, due to the fact that the latter are obtained on the basis of the same model. Thus, for such populations, we expect that the correction made to the variance estimators should eliminate the bias $B_{2}$ of order $n^{-2}$. On the other hand, the populations arising from (4), where at least one of the values of $\alpha, \gamma, g$ in $\underline{\theta}$ is different from those used to create the favourable populations (see Table 2) are said to be unfavourable. These unfavourable populations serve as a basis to check the robustness of the modified variance estimators in relation to the reduction of bias. The superpopulation models used to generate the unfavourable populations, six in number, are the same as those used by Valliant (1990).

Table 2: Parameters used to generate the unfavourable populations

| Model | $\alpha$ | $\beta$ | Y | $\sigma^{2}$ | g |
| :---: | ---: | ---: | :--- | :---: | :---: |
| 1 | 100 | 1.5 | 0 | 0.25 | 1.5 |
| 2 | 100 | 1.5 | 0 | 0.25 | 2.0 |
| 3 | 100 | 1.8 | -0.0008 | 0.25 | 1.5 |
| 4 | 100 | 1.8 | -0.0008 | 0.25 | 2.0 |
| 5 | 100 | -0.3 | 0.0009 | 0.25 | 1.5 |
| 6 | 100 | -0.3 | 0.0009 | 0.25 | 2.0 |

The study of the performance of the different variance estimators involved comparing the relative bias of $\hat{\boldsymbol{V}}$ and $\hat{V}_{\boldsymbol{w}}$ for each type of estimator (Taylor, Royall, and jackknife) for both the favourable and unfavourable populations. We did not compare the types of estimators: the work of Wu and Deng (1983) has already shown that it is difficult to find the "best" variance estimator, as its performance depends upon the model used. Also, our simulation studies attempted to answer the question "Does the modification of a variance estimator allow a major reduction in bias for a population that does not conform to the model underlying this modification?" The results obtained allow us to answer this question in the affirmative.

The results obtained for the favourable populations conformed to the theory discussed in section 2 (see Figures $1,2,3$ ). The three modified variance estimators have a bias that is of a smaller order than their traditional versions. This is verified by the fact that the bias of $\hat{V}$ has a higher absolute value than that of $\dot{\boldsymbol{V}}_{\boldsymbol{w}}$, and that the bias of $\hat{V}_{M}$ tends more rapidly toward 0 than that of $\hat{V}$, when $n$ increases. Furthermore, the three modified variance estimators have a similar relative bias, around 0 , due to the elimination of the term of order $O\left(n^{-2}\right)$. The difference between the relative bias of the traditional estimators conforms with the results shown in Table 1, concerning $E_{\S}\left(B_{2}\right)$ : the absolute value of the bias of $\hat{V}_{T}$ is about twice that of $\hat{V}_{T 2}$ and $\hat{V}_{f}$.

For unfavourable populations, the correction allows a significant reduction of the bias only for $\hat{\boldsymbol{V}}_{T}$ and $\hat{\boldsymbol{V}}_{s}$ (see Figures 1 and 3). The estimator $\hat{V}_{\pi}$ does not share this robustness in a marked fashion: for many unfavourable populations the correction has little impact (see Figure 2), and, for some superpopulation models other than those described in Table 2, it even tends to increase the absolute value of the bias. This may be
explained by the fact that the form itself of this estimator is linked to the favourable superpopulation model from the beginning. In fact, $\hat{\nabla}_{12}$ is approximately unbiased under this model. The unfavourable case in Figures 1 to 3 refers to results obtained for populations generated in accordance with model 4 , and reflects the general trend of results obtained with other models.

Figure 1: Relative bias (\%) of the estimator $\hat{V}_{T}$ and its modification.


Figure 2: Relative bias (\%) of the estimator $\hat{V}_{\pi 6}$ and its modification.


Figure 3: Relative bias (\%) of the estimator $\hat{V}_{J}$ and its modification.


Using simulation, we also compared the results on the MSE of the variance estimators for the six superpopulations considered (see Table 3). The modified variance estimators $\hat{V}_{M T}$ and $\hat{V}_{M T 2}$ have, as a general rule, an MSE that is higher than that of their traditional versions. For these two estimators, the increase in their MSE is the price that has to be paid to obtain a smaller relative bias. This "price", however, seems to be higher for $\hat{V}_{j}$. The jackknife estimator is different: $\hat{V}_{\boldsymbol{M}}$ has a smaller MSE than that of $\hat{V}_{j}$. These statements are valid for most of the populations generated, even though only the results obtained for model 5 are shown in Table 3.

Table 3: Increase (decrease) in the MSE due to the modification of traditional variance estimators under model 5

$$
\left(\Delta \operatorname{MSE}\left(\hat{V}_{\mu}\right)=100\left(\frac{\operatorname{MSE}\left(\hat{V}_{\mu}\right)-\operatorname{MSE}(\hat{\eta}}{\operatorname{MSE}(\hat{V}}\right)\right) .
$$

| n | $\Delta \operatorname{MSE}\left(\hat{V}_{\boldsymbol{m}_{\pi}}\right)$ | $\Delta \operatorname{MSE}\left(\hat{V}_{\boldsymbol{w}^{\prime 2}}\right)$ | $\Delta \operatorname{MSE}\left(\hat{V}_{\boldsymbol{M} \boldsymbol{J}}\right)$ |
| ---: | :---: | :---: | :---: |
| 5 | 28.2 | 7.3 | -5.0 |
| 10 | 15.7 | 6.3 | -9.4 |
| 15 | 8.9 | 3.5 | -6.1 |
| 20 | 6.8 | 3.0 | -4.0 |

In summary, in light of the results obtained, we recommend the use of $\hat{\nu}_{m T}$ as the most valid and robust variance estimator (that is to say, the one with a relatively small bias and a small MSE, as independently possible 8 from the distribution of the population under consideration). On the other hand, if we wish to use a jacknnife estimator, we should use $\hat{V}_{w}$, as this has both a bias and an MSE that are lower than those of $\hat{V}_{J}$.

## 4. CONCLUSIONS

We have described a method that can be used to reduce the bias of variance estimators traditionally used to estimate the variance of the ratio estimator. This method involves multiplying these variance estimators by a correction factor, thus allowing a reduction of the bias. The simulation results show that this correction may be effective and robust when applied to the Taylor and jackknife variance estimators. This correction has the effect of lowering the MSE of the jackknife estimator.

It would be interesting to generalize the method described here to the estimation of the variance of other types of regression estimators: combined, separate ratio, generalized regression, etc.

## REFERENCES

Deng, L.Y., and Wu, C.F.J. (1987). Estimation of variance of the regression estimator, Journal of the American Statistical Association, 82, 568-576.

Leblond, Y. (1989). Contributions à la théorie d'estimation pour sous-populations, Doctoral Thesis, Université de Montréal.

Royall, R.M., and Cumberland, W.G. (1978). Variance estimation in finite population sampling, Joumal of the American Statistical Association, 73, 351-358.

Royall, R.M., and Cumberland, W.G. (1981a). An empirical study of the ratio estimator and estimators of its variance, Joumal of the American Statistical Association, 76, 66-67.

Royall, R.M., and Cumberland, W.G. (1981b). The finite population linear regression estimator and estimators of its variance: an empirical study, Joumal of the American Statistical Association, 76, 98-102.

Valliant, R. (1987). Conditional properties of some estimators in stratified sampling, Joumal of the American Statistical Association, 82, 509-519.

Valliant, R. (1990). Comparison of variance estimators in stratified random and systematic sampling, Joumal of Official Statistics, 6, 115-131.

Wu, C.F.J. (1982). Estimation of variance of ratio estimator, Biometrika, 69, 183-189.
Wu, C.F.J. (1985). Variance estimation of the combined ratio and combined regression estimators, Joumal of the Royal Statistical Society, Ser. B, 47, 147-157.

Wu, C.F.J., and Deng, L.Y. (1983). Estimation of variance of the ratio estimator: an empirical study. In Scientific Inference, Data Analysis and Robustness, Ed. G.E.P. Box et al., 245-277. New York: Academic Press.

## SESSION 5

## Improving Data Collection

# ENHANCING THE QUALITY OF TIME CRITICAL ESTIMATES THROUGH THE USE OF MIXED MODE CATI/CASI COLLECTION 

G.S. Werking and R.L. Clayton'


#### Abstract

A chronic problem in the preparation of time critical estimates is the significant limitations inherent in mail collection methods. To address this issue, the Bureau of Labor Statistics has conducted an extensive 7 year research effort into the use of computer assisted telephone intervicwing (CATI) and the computer assisted self interviewing (CATI) methods of touchtone and voice recognition self-response. This paper will summarize some of the significant results of this research covering both performance and cost data. The paper will conclude with a discussion of a large scale implementation program of these techniques for a monthly sample of 350,000 establishments.


KEY WORDS: Employment Statistics; Revisions; Touchtone Collection; Voice Collection; Cost Analysis.

## 1. INTRODUCTION

### 1.1 Employment Statistics in the United States

On the first Friday of each month, the Bureau of Labor Statistics (BLS) releases data on the United States' employment situation for the previous month. On release day, the Commissioner of Labor Statistics appears before the Joint Economic Committee of Congress and provides a detailed analysis of the current month's data and trends; at the same time, the data are made available to the news media and the financial and business communities. This closely watched set of statistics is the earliest indicator available on the previous month's economic activity and is used as a major gauge of the health of the U.S. economy. The data in the release cover employment, hours and earnings by detailed industry which are derived from the Bureau's 350,000 unit monthly establishment survey-the Current Employment Statistics (CES) survey-along with labor force and unemployment data which are derived from the Bureau's 60,000 unit household survey--the Current Population Survey (CPS).

The establishment survey data have many important economic uses. Due to the CES survey's size and timeliness in conjunction with the importance of the basic payroll statistics which it collects, the CES monthly estimates are not only used as principal economic indicators themselves but also are included in the development of many of the Nation's other major economic indicators including: Personal Income for the Gross National Product, the Index of Leading Economic Indicators, the Index of Coincident Indicators, the Industrial Production Index, Real Earnings measures and Productivity measures. While the timeliness and accuracy of the CES statistics are essential in analyzing the current economic conditions in the United States, the CES survey bas relied on a mail data collection process since its inception in the early 1900s. This collection process results initially in the publication of "preliminary" estimates for higher level aggregates using only partial sample returns followed by "final" estimates two months later which use the full sample. The process of producing preliminary and final estimates for a given month periodically yields a substantial revision to the monthly estimates. These revisions not only affect the basic CES statistics but also the other statistics which use the CES estimates as input. It was

[^20]the issue of revisions in the CES survey's monthly estimates which initiated the Bureau's research program into automated telephone data collection approaches.

This paper provides an overview of the Bureau's 7 year research program into automated telephone collection techniques and summarizes some of the most significant results. The following sections will: describe the CES survey process; discuss the research program evaluating Computer Assisted Telephone Interviewing (CATI), Touchtone Data Entry (TDE) and Voice Recognition (VR) data collection methods; detail some of the significant research results covering both performance and cost data; and conclude with a discussion of the large scale implementation program of these methods in the CES survey.

### 1.2 Current Employment Statistics Survey

The CES survey, at 350,000 units, is the largest monthly sample survey in the United States. It is conducted by the Bureau as a Federal-State cooperative program. Under the Federal-State survey environment, the Bureau specifies the survey's sample design and operational procedures while the State conducts all data collection and edit reconciliation activities. The Bureau produces and publishes extensive monthly industry detail at the 2,3 and 4 digit industry levels for the Nation as a whole with the States producing monthly State and area ( 270 Metropolitan Statistical Areas) estimates.

The CES estimates are widely regarded as highly accurate economic statistics. Once each year, universe employment counts for the previous year are made available from the Unemployment Insurance tax records; these counts are used to annually benchmark (realign) the CES sample estimates to the full universe counts. The annual benchmark process provides for a more accurate current monthly estimate along with providing a comprehensive estimate of overall survey error. The average difference in the CES final sample estimate versus the universe count over the past 5 years is under $.2 \%$, with 4 years during the 1980 s where the difference was approximately zero. While the CES final monthly estimates are regarded as highly accurate relative to the universe counts; the preliminary monthly estimates, which are based on approximately $50 \%$ of the mail sample returns, have been periodically subject to large revisions when compared to the final estimate available 2 months later. Over the years, improvements in reducing the size of the monthly revisions have been made; however, periodic large revisions have been viewed as a byproduct of conducting a large decentralized mail data collection process.

The decade of the 1980s brought about a number of changes for the CES program which would significantly alter the urgency and options for resolving the monthly revision issue. The 1980s created a far more quality conscious uscr constituency and while the CES products had not necessarily deteriorated, the CES users' expectations on quality and "fitness for use" had greatly increased. Much of this new way of thinking is directly attributable to the efforts of Deming, Juran and others on the subject of quality management. The 1980s also saw a much greater focus on the uses and the importance of the CES payroll statistics in assessing the current bealth of the U.S. cconomy, but with the rise in the visibility of the CES statistics camc a corresponding user frustration with monthly revisions. The 1980 s also ushered in dramatic new technological breakthroughs, most notably-the age of microcomputers. This new technology offered survey agencies many new opportunities for improving data collection control and quality including: CATI, Touchtone Data Entry, Voice Recognition, Computer Assisted Personal Intervicwing (CAPI) and FAX. Several of these methods would ultimately offer options to significantly improve timeliness and quality at an equivalent or reduced ongoing cost.

The 1980s saw the Bureau shift from experimental research in the CES survey to full production testing of some of the most advanced state-of-the-art automated collection techniques available, with major implementation of these techniques scheduled for 1991.

## 2. CES RESEARCH PROGRAM

### 2.1 Research Goals

In the early 1980s, the Bureau began an extensive 7 year research effort into the causes of late response and alternative collection methods which could significantly increase response rates for the preliminary estimates. The focus of the survey research centred around obtaining answers to three basic questions:

Are data available at the establishment in time to respond by the publication deadline for the preliminary estimates?

Are there data collection methods which can ensure an $80-90 \%$ response rate under these tight time constraints?

Can the cost of these data collection methods be controlled at about the same level as the current mail collection costs?

At the conclusion of the research program, a mixed mode CATI/TDE collection approach emerged which satisfied the response rate and cost constraints for the survey. The following sections provide a brief description of these personal computer ( PC ) based data collection methods, the research tests, the response rate results and the cost analysis. Further details on these tests are documented in the research papers listed in the references.

### 2.2 Data Collection Methods

The CES survey has a very limited data collection time period available to meet the preliminary estimates publication deadline. The CES survey's reference date is the payroll period containing the 12 th of the month; thus, there are only $21 / 2$ weeks available to collect, keypunch, edit, tabulate, validate and publish the data. In order to meet these tight time constraints, a collection method must be able to obtain the required data as soon as they become available within the establishment. The principal data collection methods studied are described below.

Mail - The CES questionnaire is a single page which provides space for the employer to record 12 months of data. The employer receives the questionnaire in the mail on or about the 12th of the month (i.e., the survey reference date) and subsequently fills in the row of data items corresponding to the current month. There are five basic data items collected: all employees, women worker employment and production (or nonsupervisory) worker employment, hours and earnings. Once completed, the employer mails the form back to the State agency where it is keypunched and set aside for the next month's mailing. As indicated earlier, this process currently yields a $50 \%$ response rate in the $21 / 2$ weeks available for the preliminary estimates.

Computer Assisted Telephone Interviewing - Under CATI collection, the employer is mailed the CES questionnaire at the beginning of the year and retains it for recording the monthly data. Each month as the payroll data become available, the employer fills in the data items for that month and waits for the prearranged CATI call from the State agency. When the State agency calls, the data are collected under CATI, edited and a time for the next month's collection call is arranged.

Touchtone Date Entry - Under TDE reporting, the employer does the same activities as under CATI except instead of waiting for the State agency's CATI call, the employer now calls an 800 telephone number connected to the touchtone PC located at the State agency. The employer then touchtone enters the data items following the prompts in the automated interview. As each data item is entered by the employer, it is read back for respondent verification.

Voice Recognition - VR data reporting is identical to touchtone collection except the employer no longer needs to have a touchtone phone. The employer now reads the data as they appear on the form and the voice PC translates and reads back the data to the employer for verification. The VR system is speaker independent and accepts continuous speech; it recognizes the digits 0 through 9 and "yes" and "no".

## 23 Research Tests

The Bureau began developing a PC-based CATI system in 1983 for use in a two State test scheduled to begin in 1984 (Figure 1; note all figures appear at the end of the paper). The CATI system developed by the University of California at Berkeley was selected for the test and was subsequently used throughout the research effort. A small random sample of 200 units was selected in cach State and collection procedures and systems were refined over the next 7 years. The research tests were highly successful in the response rates they achieved and the tests were expanded to 9 States in 1986 and then to a total of 14 States in 1988. The composition of the test sample was also changed in 1986. Instead of selecting random samples of the full CES sample, the subsequent research tests focused only on random samples of habitually late CES respondents (i.e., those units which had a response rate of under $20 \%$ for the preliminary estimates publication deadline). Thus, the success of the new collection methods of CATI and TDE was measured in terms of their ability to move samples of reporting units with a $0-20 \%$ preliminary estimates response rate to a stable ongoing $80-90 \%$ response rate. By the end of the CATI research phase in 1990, the Bureau was collecting over 5,000 units monthly under CATI and had conducted well over a quarter of a million CATI interviews.

While CATI was proving to be highly successful in improving response rates, it also became clear by 1985 that ongoing CATI collection would be more expensive than the existing mail collection. At this time, a separate path of research was begun on how to reduce the cost of CATI, while still maintaining the high monthly response rates which it was achieving. While improvements were made in reducing the length of time required for a CATI interview, it was a new alternative PC-based telephone reporting method which would offer dramatic reductions in the collection costs of CATI.

By 1985, many U.S. banks were operating a version of touchtone entry verification for check cashing at drive-in windows. The Bureau identified a PC-based touchtone reporting system suitable for survey research testing and by 1986 was conducting a small two State test of this technique for collecting data. TDE was not viewed as a direct replacement for mail nor as a competitive method to be tested against CATI. CATI's role was to take habitually late responders and turn them into timely responders through personal contact and an educational process, while TDE's role was to take these timely CATI responders and maintain their response rates at the same high level, but at a greatly reduced unit cost. Over the 5 years of data collection, TDE has also proven to be a very successful and reliable method of telephone data collection. The research phase for TDE is now also being concluded with over 5,000 units reporting monthly under TDE across 14 States; in total the Bureau has collected over 100,000 schedules using this new automated reporting method.

As a natural follow-up to TDE, the Bureau is currently conducting several small research tests of a new Voice Recognition reporting system. Preliminary results for VR reporting have replicated the same high monthly response rates achieved under TDE, but with the important advantage that respondents find VR reporting more natural and generally prefer it over TDE. At this time, the cost of the VR hardware is approximately 15 times that of TDE; however, within several years as the initial costs of VR drop, this collection method will be a viable replacement for TDE.

### 2.4 Research Results

Over the past 7 years, the Bureau has been able to establish that the payroll data is available in most firms prior to the publication deadline for the preliminary estimates and that CATI collection has the ability to take traditionally late mail responders (i.e., $0-20 \%$ response rate for preliminary estimates) and within 6 months turn them into timely responders with response rates of $82-84 \%$ (Figure 1). These response rates have been remarkably stable over the years as the CATI sample has been expanded from 400 units to 5,000 units and the number of participating States increased from 2 States to 14 States. The research results indicate that the data do exist at most establishments in time to meet the publication deadline and that CATI collection can raise mail response rates by $60-80 \%$ for these late respondents and this rate can be maintained in the targeted range of $80-90 \%$ over long periods of time. The principal limiting factor in a respondent's ability to make the publication deadline was found to be the length of the firm's pay period (Figure 3). Employer pay periods are generally weekly, biweekly, semi-monthly or monthly. Weekly and semi-monthly payrolls can almost always be collected in time for publication with biweekly pay periods available most of the time; however, most monthly payroll
systems close out well after the publication deadline. Monthly payrolls have been one of the largest factors in limiting the CATI response rates to the $82-84 \%$ range.

Several other important results have come out of the CATI research. Under CATI, approximately $60 \%$ of the respondents will have their data available on the prearranged date for the first call with the remaining $40 \%$ using the first call as a prompt call. This rate has varied little across States or over the years of testing. A small test is scheduled to be conducted to see if an advance postcard notice to the respondent shortly before the prearranged CATI contact date will significantly limit the number of callbacks required.

The average time for a CATI interview depends on the number of items to be collected, the time efficiency of the interview instrument, and the experience of the data collector. The average time for a CATI call (Figure 1) was reduced by one-third as the CATI instrument was streamlined and interviewers became better trained. Another very important concern in the testing was the effect of CATI on sample attrition. There was some concern that employers would not want to be constantly bothered by telephone contacts and would drop out of the program. However, the sample attrition rate for CATI was about one-third of that for mail with almost no loss of large reporters under CATI. In summary, CATI appears to have maximized the achievable response rate for the preliminary estimates while also enjoying broad support from the respondents.

Due to the increased cost associated with CATI collection, the Bureau initiated research into touchtone collection. During the 4 years of testing, TDE has demonstrated the ability to take timely CATI reporters having $82-84 \%$ response rates and maintain these high rates under completely automated TDE reporting (Figure 1). The importance of this result lies in the cost savings under TDE collection versus CATI collection. One of the major concerns for TDE collection was that, unlike CATI where respondent contacts are scheduled throughout the day, TDE respondents might tend to call during the same time period thus generating busy signals and require an excessive number of touchtone PCs to handle peak load reporting. Fortunately, this was not the case, and while the touchtone PCs are on-line 24 hours a day, most calls are relatively uniformly distributed between 8 am and 5 pm (Figure 4). TDE respondents tend to require the same proportion of prompt calls as CATI respondents-approximately $40 \%$. Methods are currently also being tested to reduce the TDE prompting workload. One major advantage for the respondent is that TDE collection requires only one-half of the time of a CATI interview with the average TDE interview lasting only 1 minute and 45 seconds. Additionally, touchtone phones are widely available at most establishments; current estimates indicate that over $80 \%$ of employers could report under touchtone data collection. While TDE reporting offers many advantages to the survey agency, its strongest feature is respondent acceptance; respondents prefer touchtone reporting over mail and CATI due to its speed and convenience.

A general summary observation concerning the development of a CATI research program is that it's not particularly important which CATI hardware or software system is used during the research phase as long as it is reasonably flexible for change. The final results from testing may suggest very different CATI requirements for production implementation than that originally required for the research program. The most important and time consuming activity is the development and refinement of the methods and procedures for respondent contact. Once effective methods and procedures are developed, the requirements for a "right" system become more obvious.

### 2.5 Cost Analysis

With the performance testing and respondent acceptance for CATI and TDE proving to be highly successful, the final phase of research shifted to analyzing the transitional costs of CATI and the ongoing costs of TDE collection,

The major "labor" and "non-labor" cost categories were studied for mail, CATI, and TDE collection (Figure 2). The study looked not only at estimates of current cost, but also at projected costs over the next 10 years using the current rate of increase for the major cost items. Since CATI was to play only a 6 month transitional role (i.e., moving late responding mail units to on-time responding CATI units) prior to conversion to ongoing TDE, the major focus of the cost analysis was on the cost tradeoffs between ongoing mail versus ongoing TDE data collection.

For the labor categories, the monthly mailout, mailback, check-in and forms control operations of mail were replaced by a single annual mailout-only operation under TDE, thus eliminating a large monthly clerical operation in the States. The batch keypunching, keypunch validation, and forms control operations under mail were completely eliminated under TDE, where the respondent touchtone enters the individual firm's data and validates each entry. Another major quality and cost-efficiency advantage of TDE was that procedures for telephone nonresponse follow-up became far more feasible under TDE than mail. Under TDE, an accurate up-to-date list can be generated of respondents who have not yet called in their data, this list can then be used to conduct brief telephone prompting calls. Under mail, telephone follow-up activities of apparent nonrespondents were awkward since the State staff did not know whether the respondent's form was not yet completed, currently in the mail, in the State check-in process or at keypunch; in addition, respondents who bad recently sent their form tended to resent the additional reminder for an activity that they perceived as completed. Due to the voluntary nature of the program and the uncertainty of a respondent's response status, telephone prompting under mail was only used for critical (large employers) units.

There were no significant cost savings made for edit reconciliation as the number of edit failures under TDE remained at about the same level as under mail. This was also true for postcard reminders where the number of postcards used under mail collection for late respondents was approximately the same as the number used under TDE, where respondents received an "advance" postcard notice to touchtone their data by the due date.

In the non-labor categories, the cost of postage under mail (currently 50 cents per unit) is replaced by the cost of a telephone call and the amortized cost of the TDE machine (currently 46 cents per unit). Postage is a continually increasing cost with an annual price increase of approximately $5 \%$ (Figure 2). The rising cost of postage is driven by annual labor cost increases ( $+5.7 \%$ ) and by fuel costs (also generally increasing) with labor accounting for over $80 \%$ of total postage costs. In contrast, under TDE the cost of telephone calls has been decreasing in recent years ( $-1.7 \%$ ), along with the cost of microcomputers ( $-19.5 \%$ ).

Excluding the additional new requirement of full telephone nonresponse prompting for TDE, there are demonstrable cost savings in shifting from mail to TDE. Perhaps more importantly, under a 10 year projection of future costs of these two collection methods (Figure 5), these savings grow substantially. Attempts will be made to redirect future cost savings from TDE to help offset the full nonresponse prompting activities

There are several major conclusions which have emerged from the performance and cost analysis review. The traditional view that mail is the least expensive collection option available to statistical agencies is no longer true. The major technological breakthroughs of the 1980s in automated telephone collection can not only reduce the collection cost below that of mail but can also improve timeliness and control over the collection process. Additionally, over the next $5-10$ years, the cost of mail will become even less cost competitive with these high-technology/low-labor collection approaches due to the increasing labor and postage costs associated with mail. In the Bureau's follow-up interviews with respondents who were converted to TDE, results have shown that respondents have very little trouble adapting to this new method of reporting. Virtually all respondents showed a strong preference for TDE reporting over mail or CATI and found it to be a natural extension of other similar business applications. TDE can be viewed as a reliable replacement method for survey data collection; over the past 4 years of collection there have been no major equipment failure problems or disruptions of the collection process. Minor equipment problems bave been easily resolved using a back-up PC when required. In addition, future back-up protection for the State TDE collection process will involve the use of a call forwarding option to reroute calls to a central site should major problems occur at the State.

## 3. IMPLEMENTATION

### 3.1 Major Issues

By the end of 1989 , the Bureau had completed a very successful research program and bad sustained the high performance levels over 7 years. However, there is a significant difference between the completion of successful research and full scale implementation of new methods. While over 10,000 units were being actively collected under these new lechniques, these units represented under $\mathbf{3 \%}$ of the CES sample. Proposed collection changes
for a monthly sample of 350,000 units which has been collected for well over half a century under a decentralized State mail collection environment requires not only a very strong demonstrable user need but also broad-based support at national, regional and State levels.

As it turned out, the user need had begun to change in the early 1980s. During the 1980s, the U.S. economy experienced the longest sustained peace-time growth period in its history, with over 19 million jobs created, and unemployment rates at their lowest levels since the carly 1970s. By the mid 1980s, economic policy was firmly focused on establishing non-inflationary economic growth. The monthly CES employment growth and wage data were being closely monitored for signs of wage-induced inflationary pressures resulting from strong job growth during a period of low unemployment. With this greatly increased use and visibility of the monthly data, came a corresponding user frustration with the periodic large revisions to the preliminary estimates. While monthly revisions to the preliminary estimates had always been a part of the CES survey process and even though the size of these revisions had been reduced over the years, large revisions of over 100,000 in the preliminary monthly employment estimates of over-the-month change were now being viewed as unacceptable. These user demands for greater accuracy in the preliminary estimates would lead the Bureau to develop proposals for the implementation of automated CATI and TDE collection methods into the U.S. government's largest monthly survey.

While user demand is critical, major changes of this magnitude could not be undertaken without full support at the State level where data collection actually occurs. One guideline which remained constant throughout the research program was that the collection system was ultimately the States' data collection system and therefore must be designed to integrate well into their survey environment and create as minimal an organizational impact as possible. To that end, the CATI and TDE systems remained open for change throughout the research program. As many State suggestions and requirements as possible were taken into account with each new release of the systems. The success of much of the development work can be credited to the resiliency and endurance of the 14 research States as they made constant recommendations for improvements in systems and procedures. In the end, the CATI interview instrument had moved from an awkward simulated household survey type interview approach to a fast and efficient "screens" and "windows" approach well suited for capturing and editing longitudinal economic data. Thus, at the conclusion of the research phase, the systems and procedures were well tested and refined across a wide range of States. This approach to testing brought with it a strong sense of confidence in the methods and the systems at the State level. This would prove to be essential for the Bureau's proposed quick production implementation timetable of these state-of-the-art collection methods.

### 3.2 Approach and Impact

The main focus of the implementation proposal was the control of revisions in the preliminary estimates. Over the past 5 years, approximately $40 \%$ of all revisions were over 50,000 with $13 \%$ of the revisions exceeding 100,000 (i.c., large revisions) (Figure 6). The goal of the implementation study was to identify a minimum set of late responders which, if obtained by the publication deadline, would control the size of revisions to what was considered to be in an acceptable level (under 50,000 revision in over-the-month change). While one obvious approach was to convert all 175,000 late respondents to the new collection methods, this approach was considered to be lengthy and costly. While there was a need to responsibly control the size of revisions (i.e., not necessarily completely eliminate all revisions), there was also a corresponding need to resolve this problem in as timely a fashion as possible (i.e., convert the smallest number of units necessary to conirol revisions to under the 50,000 level).

Establishment surveys, unlike household surveys, generally have differential weighting for individual units with very large units being "certainty" units in the sample design. In the CES sample design, units with $100+$ employees make up only $20 \%$ of the sample (i.e., 75,000 units), but account for over $83 \%$ of the unweighted sample employment. These units tend to have a much lower response rate for the preliminary estimates so that if the late respondents' employment trend differs from the early respondents, these units can create a substantial revision in the sample estimates. Revision impact studies were conducted to assess the affect of large employers $100+$ on the preliminary estimates. Late respondents in size class $100+$ were included in the original sample used for the preliminary estimates and the estimates were recalculated. These estimates were then compared to the original preliminary estimates to determine the impact of $100+$ employers on revisions. The results indicated that from one-half to two-thirds of the revision was attributable to these units. These studies were
repeated over several months with similar results. Applying these projected reduction rates in revision size to revisions over the past 5 years resulted in over $97 \%$ of all revisions being below the 50,000 level as compared to the current level of only $60 \%$. This greatly reduced targeted sample size for conversion to CATI/TDE provided for an accelerated implementation time schedule consistent with controlling conversion costs to the minimum level necessary to protect against large revisions in the preliminary estimates. The Bureau will begin the implementation of CATI/TDE collection methods in 25 States in 1991 with planned expansion to all States in 1992. With the implementation of the new collection methods, the Bureau will be able to help resolve one of the most difficult and visible quality issues affecting the CES user community.

## 4. SUMMARY

The decade of the 1980s has brought about many changes for survey agencies. Some of the changes can be viewed in terms of our accomplishments made over the decade while others are more subtle and need to be viewed in terms of the changes in the survey environment in which we operate.

The 1980 s created a far more quality conscious user constituency which is quick to identify and point out our product limitations. While our products may not have deteriorated, our users' expectations on quality and "fitness for use" have greatly increased. This is an issue which we as statistical agencies must be able to respond to in order to maintain our credibility with the user community. The 1980s also ushered in dramatic new technological breakthroughs most notably-the age of microcomputers. The new technology has offered survey agencies many new opportunities for improving data collection control and quality including: CATI, CAPI, TDE, VR and FAX. Some of these options offer improved quality and control at lower ongoing costs. The decade of the 1990s may well offer even greater opportunities for using technology to improve our data collection timeliness and quality at lower costs.

As we look at the status of our statistical programs, we often find very rigid environments. The data collection approaches for our surveys often date back to their inceptions. Our data collection cost assumptions and cost studies are usually well outdated and often simplistic in approach. Since data collection generally represents the largest part of a survey's cost, it is usually well entrenched in the agency's organizational structure and can be quite difficult to restructure in order to accommodate large scale change. It is within this survey environment that we will face the major challenges and opportunities of the 1990 s.

The challenge for statistical agencies in the 1990s will be threefold:

- to be responsive to the changing quality needs of our users;
to attempt to have our research stay up with the rapid change of technology and automated data collection approaches; and perhaps more importantly
to find ways to implement successful research into our ongoing programs.
These challenges will determine the cost and quality competitiveness of our programs and our agencies in the future.


## REFERENCES

Clayton, R., and Harrell, L. (1989). Developing a Cost Model of Alternative Data Collection Methods: MAIL, CATl and TDE, Proceedings of the Section on Sunvey Research Methods, American Statistical Association.

Clayton, R., and Winter, D. (1990). Speech Data Entry: Results of the First Test of Voice Recognition for Data Collection, Proceedings of the Section on Survey Research Methods, American Statistical Association.

Groves, R.M.J. editor et al. (1988). Telephone Survey Methodology, New York: John Wiley and Sons.

Phipps, P.A., and Tupek, A.R. (1990). Assessing Measurement Errors in a Touchtone Recognition Survey," Presented at the Measurement Error Conference, Tucson, Arizona.

Ponikowski, C., and Meily, S. (1988). Use of Touchtone Recognition Technology in Establishment Survey Data Collection, Presented at the First Annual Field Technologies Conference, St. Petersburg, Florida.

Statistical Policy Working Paper 15 (1988). Quality in Establishment Surveys; Office of Management and Budget.

Statistical Policy Working Paper 19 (1990). Computer Assisted Survey Information Collection, Office of Management and Budget.

Werking, G.S., Tupek, A.R., Ponikowski, C., and Rosen, R. (1986), A CATI Feasibility Study for a Monthly Establishment Survey, Proceedings of the Section on Survey Research Methods, American Statistical Association, 639-643.

Werking, G.S., and Tupek, A.R. (1987). Modernizing the Current Employment Statistics Program, Proceedings of the Business and Economic Statistics Section, American Statistical Association, 122-130.

Werking, G., Tupek, A., and Clayton, R. (1988). CATI and Touchtone Self-response Applications for Establishment Surveys, Joumal of Official Statistics, pp. 349-362.

Figure 1. Research Summary

| M 41 |  | 1984 | 1985 | 1986 | 1987 | 1988 | 1989 | 1990 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Resp. Rates | 47\% | 47\% | 48\% | 49\% | 49\% | 51\% | 52\% |
|  | Unhs | 400 | 400 | 2000 | 3000 | 5000 | 5000 | 5000 |
| CATI | Resp. Rates | 83\% | 8\% | 82\% | 6\% | 83\% | 84\% | 82\% |
|  | \% Call Back | 44\% | 42\% | 20\% | 51\% | 42\% | 61\% | 41\% |
|  | Ar. Minutes | 5.6 | 5.6 | 5.0 | 4.8 | 4.4 | 3.5 | 3.8 |
| $\begin{aligned} & \text { TDE } \\ & \text { \& } \\ & \text { Volce } \end{aligned}$ | Unhs |  |  |  | 400 | 800 | 2000 | 5000 |
|  | Resp. Rates |  |  |  | 78\% | 100\% | 6\% | 62\% |
|  | * Call back |  |  |  | 45\% | 45\% | 43\% | 40\% |
|  | Av. Minutes |  |  |  | 1.6 | 1.8 | 1.7 | 1.7 |

Figure 2. Data Collection Costs
(arrows show direction of recent price chenge)

| Cost Calegory | Mail | CATI | Swi-response (TDE 8 VR) |
| :---: | :---: | :---: | :---: |
| LABOR |  |  |  |
| mail out | $\cdots$ |  |  |
| man leturn | $\underline{\square}$ |  |  |
| dala enliy | $\cdots$ |  |  |
| edil and edil reconctialion | $\cdots$ | $\leq$ | $\rightarrow$ |
| norresponse lonlowip |  |  | $\rightarrow$ |
| NON-LABOR |  |  |  |
| postago | $\cdots$ |  | $\cdots$ |
| leiephones |  | $\longrightarrow$ | $\square$ |
| microcomputers |  | $\cdots$ | $\square$ |

Recen Annual Price Chanoe Factors

| Labor | $+5.7 \%$ | ECI. Slate and Local Government |
| :--- | ---: | :--- |
| Postage | $+5.0 \%$ | U.S. Poslal Service |
| Telephone | $-1.7 \%$ | CPI-U, Inlrastale foll calls |
| Microcompulers | $-19.5 \%$ | PPI Experimenlal Price Indexes (16 bif compuler) |



# FOLLOW-UP STRATEGY FOR ECONOMIC SURVEYS 

J.-M. Berthelot and M. Latouche ${ }^{1}$


#### Abstract

At the time of the development of the generalized data collection and data capture function, particular attention was paid to optimizing the follow-up strategy. The objective is to minimize the amount of resources used without significantly affecting data quality. The resources allocated to recontact are concentrated on the questionable units that may have significant impact on the estimates. This paper describes the overall strategy for follow-ups in economic surveys, as well as the results of an empirical study using the developed strategy.


KEY WORDS: Data Collection and Capture; Recontact; Score Function.

## 1. INTRODUCTION

One of the major thrusts of Statistics Canada over the past few years has been to try to improve the efficiency of its operations. One wants to reduce respondent burden, increase the efficiency of the production process and improve the quality of the end product. To respond effectively to these objectives, it became necessary to reconsider the design of some production processes. This led to the creation of the Generalized Survey Function Development Project (GSFD). The aim of this project is to develop generalized survey tools such that the resulting methods, operations and software would meet the requirements of most surveys (Colledge, 1987).

The Generalized Data Collection and Data Capture Function (DC2) is one of the survey functions developed within the GSFD. This function consists of the activities necessary for the acquisition, validation, and conversion into machine readable format of survey data (GSFD, 1989). The aim of the generalized data collection and capture function is to provide a solid methodological basis on which most surveys will be able to rely. The overall objective is to maximize the effectiveness of the generalized data collection and capture function in terms of costs, time, resources, and particularly data quality (Bilocq, 1988).

The data collection and capture process has always been a demanding operation in terms of human resources and time. In particular, one of the tasks that can be very expensive is the follow-up. The follow-up consists of contacting the respondent, after the initial contact, in order to obtain missing information, to verify, and if necessary, to correct questionable information. Follow-up is used for total or partial non-responses, error correction and for confirmation of outlier data.

Consequently, during the development of the generalized data collection and capture function, particular attention is given to optimizing the follow-up strategy. The objective is to minimize the amount of resources used without significantly affecting data quality. The resources allocated for recontacts are concentrated on questionable units that may have a significant impact on the estimates. This paper describes the overall strategy for follow-ups in economic surveys, the development of a score function for the classification of questionable units in accordance with their potential impact on the estimates, and the application of the strategy to data from the Annual Retail Trade Survey.

[^21]
## 2. THE PRESENT SITUATION

Until now, the follow-up strategy has been to recontact every respondent who either did not return the questionnaire, returned an incomplete questionnaire, or returned a questionnaire with questionable data. The objective of this strategy is to ensure that all errors are identified and corrected, regardless of their impact at the estimation level.

In order to ensure the identification of all questionable units, data undergoes extensive editing. Many rules are used to edit the data. Reject lists produced by the application of these rules must be manually verified, resulting in a considerable work load. The use of this approach often results in over-editing of the data. Reject lists include only a small percentage of units that are truly inaccurate. In the long run, this leads to a loss of credibility in the editing rules, which could affect the effectiveness of the editing process.

Different editing rules may be applied at various stages of survey processing. Consequently, it is sometimes necessary to recontact a respondent several times, in order to verify all the potential errors found in a questionnaire. This can result in a large respondent burden.

Due to budget and time constraints, it is practically impossible to recontact all the units with non-responses or questionable data. A decision is usually made to recontact only certain respondents. This decision is made at a certain point in the data collection and capture process, depending upon time and budget constraints. The limits imposed on recontact efforts are not often based on a pre-established sampling plan. Thus the impact of these limits on data quality cannot be evaluated and included in the estimation of the survey error.

In summary, the present strategy is very expensive, takes a great deal of time, and results in a considerable respondent burden. The process is not very efficient due to the large number of editing rules used. The error ensuing from the limits imposed by time and budget constraints cannot be measured and included with the estimates. In addition, there is no evidence that it is necessary to identify and correct every error to ensure an acceptable level of data quality.

## 3. THE PROPOSED STRATEGY

In an environment of budgetary restrictions, it is practically impossible to apply the present strategy. Consequently, a new strategy that accounts for operational constraints must be developed. Particular attention must be given to certain factors when developing this follow-up strategy for economic surveys. There must be an attempt to reduce the resources used for follow-up, to improve the editing process, to automate the process, to limit recontacts, and to measure and evaluate the error generated by the follow-up process.

The strategy proposed is relatively simple and is based on one major principle, the development of an integrated approach for all the phases of data processing. This principle, however, has repercussions on each of the processing phases. The rest of this section deals with these repercussions.

### 3.1 Data Capture

Data capture consists of converting the information found on a questionnaire into machine readable format. During this phase, it is important to ensure that the data processing system is designed so that it can transfer the questionnaire information to an electronic medium, without modifications, regardless of the information given on the questionnaire. In this way, all manual processing related to the constraints imposed by the data processing system are avoided. It is no longer necessary, for example, to edit the questionnaires manually to ensure that the fields defined as compulsory by the data processing system are present on all the questionnaires. By general agreement, the data found on a questionnaire become constraints that must be respected. It is, therefore, necessary to ensure that the data processing system is flexible enough to capture all the information found on a questionnaire, without any obstructions due to missing information. The recontacts generated by data capture errors in the old approach are thus eliminated.

### 3.2 Editing

The objectives of editing are to permit the identification of errors, questionable data and inconsistencies in the data. In order for the edit process to be effective, it is necessary to structure the editing of data while taking account of these objectives. In the proposed strategy, the editing has been broken down into three distinct phases: the editing at data capture, editing for questionable data, and consistency edits. The following paragraphs give a brief glimpse of the content of these three phases.

The goal of editing at data capture is to maximize agreement between the information captured and the information found on a questionnaire. It is performed by using syntax rules, range edits, and basic consistency rules. It is done in an interactive way during data capture. When a reject is identified by the editing process, the operator must simply ensure that the information captured corresponds to the information found on the questionnaire. If such is the case, the operator must simply state that the message bas been received and continue the data capture of the rest of the questionnaire. If the opposite is true, the operator must correct the information captured so that it corresponds to that found on the questionnaire. In this way, a questionnaire can be captured all at once without other manual interventions.

The objective of editing for questionable data is to identify quantitative data that are outlying or erroneous. Only non-responses and units identified by this procedure will be subject to follow-up. In this way, the potential number of follow-ups is limited. The identification of outlying or erroneous data is made by means of statistical methods that call upon the distribution of data as observed during one or several earlier cycles of a survey (Hidiroglou and Berthelot, 1986).

The objective of editing for consistency is to ensure that the quantitative variables of a questionnaire are consistent with one another. Inconsistencies are identified using a set of linear rules imposed on a group of variables. If an inconsistency is found, it must be corrected by the automatic imputation of the variables responsible for this inconsistency. In this way, all the follow-ups related to the presence of inconsistencies in a questionnaire are eliminated. Editing for consistency must be developed jointly with editing for questionable data, as their roles are complementary.

## 33 Recontact

Recontact is the action of going back to the respondent to obtain missing information or to confirm or correct information considered questionable by the editing rules. The use of selective recontact is one of the main characteristics of the strategy. Recontact efforts are concentrated at two levels.

Firstly, it is necessary to follow-up cases of total non-response. A non-response may correspond to an out-ofscope unit as much as to a poor respondent that is part of the survey population. This follow-up is indispensable to ensure that an appropriate imputation will be made during a later processing phase.

Secondly, it is critical to concentrate the rest of the resources on units that have a significant impact on the estimates. In order to do so, a score function may be used. Its' objective is to evaluate the impact on the estimates of questionnaires that include questionable data. A questionnaire sample plan that uses the score function value as a variable of interest enables the optimal allocation of resources available for recontact.

### 3.4 Imputation

Since the now strategy encourages selective recontact, inconsistent or missing data will be found in the file after recontact is completed. In order to obtain a complete and consistent file, it is necessary to impute values for these records. The imputation of these data must be done automatically. The Generalized Edit and Imputation System (GEIS) developed at Statistics Canada is an ideal candidate for this task. It uses a set of linear rules to identify inconsistencies, localize the fields to impute, and impute the values, while respecting the consistency rules. In addition, this system guarantees consistency at the micro-data level (Whitridge et al. 1988).

### 3.5 The Measurement of Error

Since the strategy promotes the restriction of recontacts, it is necessary to be able to measure the error generated by the fact that only a portion of the potential errors are followed-up. To do this, a sample plan aimed at recontacting potential errors must be established before beginning data collection. The sample plan is designed using the results of one or more earlier cycles of a survey. The sample plan must involve the whole sample, that is, all units in the sample must have a non-zero probability of being selected for recontact.

For each respondent, an error for any variable can be evaluated using the difference between the value after recontact and the value before recontact (including imputed values). For each variable, the evaluation of error at the level of the estimate is obtained through the recontact sample plan. The principle put forward is similar to the one used in quality control. A sub-sample is used in accordance with a pre-established plan to evaluate error in the total sample.

The proposed strategy promotes the use of a selective recontact procedure. A systematic recontact effort is made for total non-respondents in order to ensure that the status of the respondent (active, inactive or out of scope) is obtained, in order to guarantee, an appropriate imputation, if necessary, during a subsequent phase. Therefore, recontact resources are concentrated on questionable units that may have a significant impact on estimates. The internal consistency of all respondents is guaranteed by the use of the GEIS. The use of selective recontact combined with the GEIS should reduce operational costs and processing time without affecting data quality in any major way.

## 4. SCORE FUNCTION

One of the essential elements of the proposed strategy is the use of a 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 of the questionnaire's variables 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. Four criteria guided the research for a score function: the size of the responding unit, the size and number of potential errors, and practical considerations.

Depending upon the variable of interest, the size of a responding unit is a good indicator of the effect that this respondent may have on the estimates, and should be taken into account when designing a score function. An indication of size can generally be obtained, either from the data reported in the current cycle, the data of an earlier cycle, or administrative data.

A measure of the error that questionable data could generate is an important factor for discriminating between respondents. If it were possible to recontact the real errors in decreasing order of importance, a point would eventually be reached below which it is no longer necessary to recontact respondents, since the residual error would be negligible. The size of the error can be evaluated using differences or trends between raw values and historical data. In the context of a sample survey, the selection weight of the units must also be used during the evaluation of error.

Following the same line of thought, the number of questionable data items in a questionnaire is another important factor. Of the units having the same relative impact, it should be a priority to recontact those with the greatest number of questionable data items in order to confirm or correct as much information as possible.

Practical considerations must be taken into account in order to meet the needs of various parties concerned. Certain variables may be considered essential by subject matter experts and for this reason require particular attention. In addition, methodologists may find that it is difficult to impute certain variables. In both cases, the variables in question should have a greater weight in the calculation of the global score of a respondent. The response rate at the cell level must also be considered. A low response rate in a cell can increase the need for recontact. The function's distribution should be similar from one publication cell to another. In this way, the
function's parameters could be determined at a higher level of aggregation than the publication cell, and thus increase the stability of these parameters.

A score function must be used within certain operational limits. The function must be easy to put into place and flexible enough for use with various surveys. Its use must not disrupt data processing. A score function must use only respondent information that is the subject of the study, in conjunction with the input parameters.

The development of a score function must take these criteria into account. The objective is to develop a general formula that strikes a balance between these various factors.

## 5. APPLICATION OF THE STRATEGY

In order to evaluate the relevance of the proposed strategy, a simulation was applied to data from the 1987 Annual Retail Trade Survey. The main objectives were to verify the applicability of certain principles put forward in the strategy, and to evaluate the use of a score function with the survey data.

### 5.1 Data used

The Annual Retail Trade Survey (ARTS) is the responsibility of the Industry Division of Statistics Canada. It is a census of Canadian retailers with Total Net Sales and Receipts of at least one million dollars a year. The population is divided into chains and independents. The chains usually correspond to large enterprises with more than four business locations in the same trade group; the other enterprises are considered to be independent. For the simulation, only the independents were used.

For the purpose of the study, 12 quantitative economic variables were used. A subset of the population including two trade groups for all the provinces, and two provinces for all trade groups was used. The subset totaled 2,053 questionnaires. The trade groups were chosen from the food products sector (trade group 20) and from the motor vehicle equipment/manufacturing industry (trade group 120). The two provinces retained were Prince Edward Island and Alberta.

The strategy was applied to the 1987 raw data provided by respondents. The 1987 final data, as produced by Industry Division, were used as control data to evaluate the effectiveness of the simulation results. The 1985 and 1986 final data were used for the calculation of the parameters for editing and the score function.

### 5.2 Methodology

The methodology used for the simulation respected the proposed follow-up strategy as much as possible. It had to be adapted, however, in order to account for operational constraints.

### 5.2.1 Data capture and editing

To ensure that the raw data corresponded well to the data found on a questionnaire, the 2,053 questionnaires were recaptured. No modification was made to the data during the recapture. The data was keyed twice to ensure the quality of data capture.

An editing plan for questionable and potentially erroneous data was established in several phases. Basic validation rules were developed using the questionnaire structure and the knowledge of Industry Division specialists. This enabled us to identify partial or total non-response and the out-of-scope units.

Subsequently, the variables were grouped in such a way as to form homogeneous variable sub-sets using the method described in Bilocq and Berthelot (1990). The link between the variables in each of the sub-sets was modeled. These two phases were performed using 1985 and 1986 data from the population as a whole.

Questionable or erroneous data were identified by applying the basic rules and the models of the links between the variables to the 1987 raw data. The results of this editing plan (success or failure) were kept for each of the variable and each questionnaire.

### 5.2.2 Recontact

For the simulation, only those units that included non-responses, errors, or questionable data were considered for recontact. Recontact was simulated by replacing the raw data from a questionnaire by its final data. This assumed that a single recontact was enough to correct all the problems found on a questionnaire. This hypothesis was necessary to perform the simulation. Indeed, it was out of the question to recontact respondents more than three years after the end of the survey's reference period.

Recontacts were divided into two mutually exclusive categories: automatic and selective. An automatic recontact was used for total non-response, units potentially out of scope, and questionable units that were part of a small publication cell. Small cells were defined as any province by trade group cross-classification that contained less than ten reporting units.

On the other hand, a selective recontact was designed using a score function. The score function was evaluated for each questionnaire that failed the edit for questionable data. The score obtained gives an indication of the impact of a given questionnaire on the estimates. The higher the score, the greater the potential impact. For the purpose of the simulation, only units with a score above a predetermined threshold were recontacted. This corresponded to making a recontact with a probability of 1 for units above the threshold and a recontact with a probability of 0 for the others. This differs significantly from the strategy that proposes a sampling plan with probabilities greater than 0 for all units.

During the development of a score function for the ARTS, three different formulas were explored. Each of these emphasizes a specific aspect or a particular operational constraint. After various analyses, one formula was retained. It is a compromise between simplicity, precision, and the production of a similar function value distribution for each publication cell. The formulas considered, and the analyses related to them, are documented in Latouche and Berthelot (1990).

The general formula retained emphasizes the absolute difference between the raw current value and the final value of the preceding cycle. It is defined at the level of each publication cell.

For a given cell $p$ one has:
$y_{k, L, i}$ the raw value for respondent $k$ within cell p for variable $i$ at time $t$;
$y_{k, i, t-1}$ the corresponding final value at time $t-1$;
$w_{k, i}$ the selection weight at time $t$,
$\hat{Y}_{., ., t-1}$ the total of cell p for variable $i$ at time $t-l$;
For a respondent $k$, the function is then defined by:

$$
\operatorname{SCORE}_{k}=\sum_{i=1}^{I} \frac{W_{k, i, t} \times\left|Y_{k, i, t}-Y_{k, i, t-1}\right| \times Z_{k, i, t} \times v_{\ldots, i, t}}{\hat{Y}_{\ldots, i, t-1}}
$$

Where

$$
z_{k, i, t}=\left\{\begin{array}{ll}
0 & \text { if } y_{k, i, t} \\
1 \text { if accepted by the edit } \\
y_{k, i, t} & \text { is questionable }
\end{array}\right\}, \text { and }
$$

$\nu_{u, t}$ the weight indicating the importance of variable $i$ at time $t$.
For the simulation, a uniform weight was used for each of the variables, that is, no variable was considered more important than another. Since the ARTS was a census, a selection weight equal to one was used.

To evaluate the behaviour and the effect of the score function, different rates of recontact were used for the selective recontacts. The recontact rates considered were zero, seventeen, thirty-four, fifty, and one hundred percent. Zero percent corresponded to making only the automatic recontacts. One hundred percent corresponded to recontacting all units that included questionable data, while fifty percent represented a midpoint. Seventeen and thirty-four percent were used as a compromise between zero and fifty percent.

These different recontact rates were used to identify the point at which additional recontact would not perceptibly improve data quality. From this point, the additional cost of a recontact is no longer justified and the use of resources for recontact is no longer optimal.

### 5.2.3 Consistency and Imputation

Statistics Canada's Generalized Edit and Imputation System was used to ensure that all records were complete and consistent. A set of linear rules was imposed on the data after the recontacts were simulated. This set is used to define a consistent record, done by determining a multi-dimensional space within which a record is considered valid. The use of this type of rule enables the use of linear programming to automatically identify and correct the variables that cause inconsistency in a record (Schiopu-Kratina and Kovar, 1989).

The GEIS was used with each of the five recontact rates. Five files containing the final simulation results were created following the application of the GEIS. These files were analyzed in order to evaluate the relevance of the proposed strategy.

### 5.3 Results

At the data editing stage, 1,163 of the 2,053 questionnaires failed at least one of the rules. Of these 1,163 failures, 584 were considered automatic recontacts. They were distributed into 245 total non-responses, 322 units potentially out of scope, and 17 questionable units within small cells (trade group x province).

There remained a total of 579 questionnaires containing questionable data that were eligible for selective recontact. The five selective recontact rates applied to the 579 units determined the total number of recontacts made in each case. These percentages correspond to $0(0 \%), 97(17 \%), 194(34 \%), 285(50 \%)$ and $579(100 \%)$ selective recontacts. These recontact rates are approximately the same for each of the publication cells.

To evaluate the error generated by the strategy, we assumed that the final value corresponded to the real value that is, that the final data did not include any error and represented reality. The effectiveness of the strategy was evaluated for each of the variables by comparing, at various levels, the estimates obtained using the strategy against those obtained using the 1987 final data. The measurement used was named "pseudo-bias." For a given level of aggregation "a" and for a variable i , the pseudo-bias is defined as follows:

$$
\text { pseudo-bias }_{a, i}=\frac{\hat{Y}_{a, i, 87}-Y^{\prime} a, i, 87}{Y_{a, i, 87}^{\prime}} \times 100 .
$$

Where $\hat{Y}_{\text {ai, }, 87}$ corresponds to the total in accordance with the simulation and $\boldsymbol{Y}_{\text {a,i,87 }}^{\prime}$ corresponds to the total in accordance with the final data.

The results below are encouraging for the part studied at the time of the simulation. Figure I below shows the pseudo-bias for the variable Total Net Sales and Receipts for the five selective rates of recontact at three levels of aggregation. For each of the selective rates of recontact the pseudo-bias is relatively small for this variable. Generally, the pseudo-bias decreases as the selective recontact rate increases. The major point is that the pseudobias reduction is greater when the selective recontact rate increases from $17 \%$ to $34 \%$ than when it climbs from $34 \%$ to $50 \%$. With a rate of $34 \%$, pseudo-bias is only $-0.18 \%$ and it improves little with a rate of $50 \%$, moving only to $-0.14 \%$. Those 91 extra recontacts between these two rates improve the estimate, but not substantially. At first glance, it appears that a selective recontact rate of $34 \%$ is enough. A similar pattern was observed for the other variables studied.

FIGURE 1


To ensure that the most appropriate selective recontact rate is chosen for the ARTS, comparisons were made at the publication cell level. Figure II below shows the pseudo-bias distribution of the publication cells for the five recontact rates. The pseudo-bias mean by cell, with a confidence interval of plus or minus one standard error for the five selective recontact rates, is shown. The pseudo-bias dispersion decreases substantially when the recontact rate goes from $0 \%$ to $17 \%$ and then to $34 \%$. The mean and the standard error of the pseudo-bias are relatively constant as soon as a selective recontact rate of $34 \%$ is used.


An interesting fact to note from Figures I and II is that the pseudo-bias does not equal zero when $100 \%$ of the selective recontacts are made. Various reasons may explain this fact. First, errors may have been identified during the analysis, subsequent to the processing of the data by Industry Division economists. Second, the editing
scheme used in the simulation is different from that used in the survey. Thirdly, the imputation methodology used in the simulation is different from that used in the survey. These three factors are sufficient to explain the difference between the simulation and the survey when $100 \%$ of the selective recontacts are made.

The results shown in Figures I and II seem to indicate that the additional expenses required to recontact more than $34 \%$ of the questionable data are not justified by the resulting small gain in quality. By taking into account the principal objective of the strategy, which is to optimize the use of resources without affecting data quality too much, a selective recontact rate of $34 \%$ is judged adequate.

These results were obtained from the analysis of the Total Net Sales and Receipts variable only. In order to ensure that this conclusion is generally valid, the pseudo-bias for a selective recontact rate of $34 \%$ was studied for the other variables.

Table I shows the pseudo-bias for the variables reported by over $95 \%$ of respondents for the 2,053 questionnaires used for the simulation.

Table I: Pseudo-bias, frequently reported variables

| VARLABLE | SURVEY TOTAL <br> IN $\$ 1,000$ | PSEUDO-BLAS <br> $\%$ |
| :--- | :---: | :---: |
| SALES | $8,515,601$ | 0.28 |
| TOTAL NET SALES | $9,015,559$ | -0.18 |
| OPENING INVENTORY | $1,405,293$ | -0.34 |
| CLOSING INVENTORY | $1,480,133$ | 0.34 |
| PURCHASES | $6,847,548$ | 0.33 |
| SALARIES | 937,354 | 0.78 |

The pseudo-bias for the frequently reported variables as a whole is relatively small. The largest value corresponds to the variable Salaries for which the simulation produces an over-estimation of $0.78 \%$. A recontact rate of $34 \%$ seems to be an acceptable compromise for these variables.

Table II shows the pseudo-bias for the other variables. These are reported infrequently, not because of partial non-response but rather because the characteristics they measure do not apply to all the respondents.

Table II: Pseudo-bias, sporadically reported variables

| VARLABLE | SURVEY TOTAL <br> IN $\$ 1,000$ | PSEUDO-BLAS <br> $\%$ | PRESENCE <br> $\&$ |
| :--- | :---: | :---: | :---: |
| GROSS COMMISSIONS | 14,232 | 8.33 | 5 |
| REPAIRS | 398,542 | -3.20 | 46 |
| RENTALS | 41,344 | 0.09 | 10 |
| FOOD SERVICES | 9,325 | -35.18 | 2 |
| OTIIER OPERATING REVENUE | 36,513 | -37.55 | 15 |
| NON OPERATING REVENUE | 16,349 | -1.46 | 22 |

The pseudo-bias is acceptable for three of the six sporadically reported variables. However, it is excessive for Gross Commissions, Food services and Other Operating Revenue. For the Gross Commissions and Food services variables, this fact is explained by their very low frequency. Such infrequent events make editing and imputation particularly difficult. To remedy this situation, a greater weight should be given to these two variables in the score function, and a more stringent editing plan should be applied. The Other Operating Revenue variable is a special case, as it is only slightly related to the main variables, and may vary greatly from one respondent to another. Consequently, it is very difficult to establish an adequate editing plan for this variable. Even if the problem is different, the solution is roughly the same: increase the weight of this variable in the function and develop a tight editing plan that accounts for the characteristics specific to this variable.

Figure III shows a histogram of the pseudo-bias distribution for publication cells for the Total Net Sales variable, for a selective recontact rate of $34 \%$.

FIGURE 3


This histogram shows that the results of the simulation are never more than $4 \%$ away from the survey results. For most of the publication cells, the pseudo-bias is less than $1.5 \%$, and the distribution is centred at zero. The results of the simulation for the Total Net Sales variable are very encouraging. The pseudo-bias distribution at the level of the publication cells is centred at zero, and does not have a large variance. A similar pattern was observed for the other frequently reported variables.

## 6. CONCLUSION

The results obtained from the simulation of the proposed follow-up strategy for economic surveys are promising. It is clear that recontacting those units that have an important impact on the estimates is necessary to obtain an acceptable level of quality. The simulation also demonstrated that it is not necessary to recontact all questionable units. Recontacting a limited number of questionable units is enough to obtain estimates of acceptable quality.

Selective recontact of units reporting questionable data combined with an automatic edit and imputation system, makes it possible to obtain estimates of an acceptable level of quality. The use of this approach in production could save data processing resources. The resources saved could be reallocated to other quality assurance and data analysis tasks.

Although the results of the simulation of the proposed strategy are interesting, an important weakness was identified: the results are not conclusive for the sporadically reported variables. However, this problem may be solved by using a more stringent editing plan, and increasing the relative importance of these variables in the calculation of the score function.

The simulation described here represents a first try at making the proposed follow-up strategy operational. Although the results are promising, there is still a lot of research to be done. For example, the possibility of doing a selective follow-up of total non-responses and out-of-scope units must be considered, a survey plan to enable the evaluation of the error generated by the strategy must be developed, the score function formula must be modified to account for non-responses, and the use of the strategy for a sampling survey must be simulated. The use of the approach proposed by Särndal (1990) to evaluate the precision of the estimates when imputed values are present deserves to be studied. The possibility of extending the strategy developed for economic surveys to social surveys must also be considered.

## REFERENCES

Bilocq, F. (1988). La fonction générale de collecte et de saisie, presented at colloque annuel sur les statistiques de l'ACFAS, Moncton, N.B., May 1988.

Bilocq, F., and Berthelot, J.-M. (1990). Analysis on grouping variables and on detection of questionable units, working paper BSMD-90-005/F, Statistics Canada, March, 1990.

Colledge, M. (1987). The Business Survey Redesign Project - Implementation of a New Strategy at Statistics Canada, presented at the US Bureau of the Census Third Annual Research Conference, Washington, March 1987.

GSFD, Generalized Survey Function Development Team (1989). Methodological and Operational Concepts in the Collection and Capture Module, technical report, Statistics Canada, 1989.

Hidiroglou, M.A., and Berthelot, J.-M. (1986). Statistical Editing and Imputation for Periodic Business Surveys. Survey Methodology, June 1986, 12, 1, 73-83.

Latouche, M., and Berthelot, J.-M. (1990). Use of a score function for error correction in business surveys at Statistics Canada, presented at the International Conference on Measurement Errors in Surveys, Tucson, Arizona, November 1990.

Särndal, C.E. (1990). Estimation of precision in the generalized estimation system when imputation is used, technical report, Statistics Canada, March 1990.

Schiopu-Kratina, I., and Kovar, J.G. (1989). Use of Chemikova's algorithm in the generalized edit and imputation system, working paper BSMD-89-001E, Statistics Canada, January 1989.

Kovar, J., MacMillan, J., and Whitridge, P., (1988). Overview and strategy for the generalized edit and imputation system, working paper BSMD 88-007E, Statistics Canada, April 1988.

Proceeding of Statistics Canada Symposium 90
Mcasurement and Improvement of Data Quality
October 1990

# THE QUALITY OF DATA IN CROP YIELD SURVEYS 

R. Fecso ${ }^{\prime}$


#### Abstract

Crop yield surveys conducted by many countries involve sevcral stages of subsampling which are determined by the enumerator and often repeat visits to the ultimate sampling unit. These procedures and the often difficult measurcment tasks associated with the ficld work provide ample opportunities for data quality problems. Studies of typical problems uncovered in surveys done in the U.S. will be reviewed. The management strategy for improving the data will be reviewed along with results of recent quality measurement studies.


KEY WORDS: Total quality management; Crop yield measurement; Bias.

## 1. INTRODUCTION

Agriculture is the largest economic activity in many countries and crop yield is often the most variable year-toyear component of agricultural production. Thus, crop yield and associated surveys have been of interest for some time. For cxample, Francis Bacon looked at relationships between weather and agricultural fluctuations in the 17 th century. Through the middle of this century research on crop yield measurement flourished. The methods developed over 30 years ago to measure crop yicld, although often modified, remain similar to those used today in many countries.

This paper focuses on one method of crop yicld measurement, called objective measurement or crop cutting, as practiced by the National Agricultural Statistics Scrvice (NASS) of the U.S. Department of Agriculture. The main purpose of the paper is to review and begin documentation of methodological studies which are not widely known. This documentation is important to us because we did not have much research activity in these surveys for the past several years. As we are now building a new rescarch group to study yield measurement, and most of the staff is new, a bascline documentation of existing work will help them focus on important areas to study. Wc hope that this work will also encourage dialogue with other countries concerning their experiences with yield measurement.

### 1.1 Quality Focus

Statistical agencies are joining industry and servicc organizations in their interest in quality philosophies. NASS began a limited total quality management (TQM) effort using objective yicld survey data as a demonstration project about five ycars ago. The demonstration was well received by management, but another data series was picked to be the first formal application because it was a bigger management concern. Hopefully, those TQM developments will be revived along with the yield research program. Thus, some of those TQM developments will be incorporated in this discussion. Basically, the approach focuses on how to provide adequate information to determine where to look to improve data quality from a complex survey design with a complex estimator knowing that resources for research are limited.

[^22]
## 2. CROP SURVEYS

This section will provide a brief background on crop yield surveys. Yield data is important since is affects commodity trading prices on exchanges, allocation of transportation services (rail cars, barges, etc.), agricultural credit, and a variety of commercial and government programs. There are two basic methods of collecting yield data: reported data and objectively collected (field enumerated) data.

Yield data reported by growers and other individuals knowledgeable about their local agricultural conditions dates back to 1862 in the U.S. These surveys are usually done by mail and ask respondents to judge observed condition of the crop or expected yield. Further detail on these surveys can be found in Fecso (1990). This discussion will concentrate on the objective yield (OY) surveys.

### 2.1 Objective Yield Surveys

Collection of data from growing sites forms the basis of OY surveys. NASS conducts these surveys for many major crops (corn, soybeans, cotton, wheat, rice, and potatoes) and specialty crops (fruit and nut tree and grapes). A broad generalization of the survey design, which varies by crop, will be given here. For further detail, see Francisco, Fuller and Fecso (1987).

The basic survey is an annual area frame survey. A stratified sample of some 15,000 area segments of about one square mile in agricultural strata is used to locate fields of the crops of interest. Acreage is estimated from this survey and the fields are subsampled for use in OY measurement. The subsampling scheme results is a selfweighting sample. Usually, there are under 2,000 fields selected per crop, restricted to the largest producing states (usually 10 states which account for over 80 percent of the crop's production).

In each field selected, two plots, called units, are selected using a random rows and/or paces method. Unit size depends upon the crop. Tightly spaced plants like wheat have units of about one square yard while crops like corn use 15 foot long rows. Each enumerator is assigned about 15 samples which may be divided among several OY crops in their area. The small assignment which is necessary to assure timeliness of survey results makes measurement of enumerator effects difficult.

Measurements taken at the units in the field vary by crop and month of the survey. For example, the end of season yield for corn is determined by the following use of collected data:
Estimate of gross yield per acre of corn for the field

$=$| $\frac{(43,560 \text { square feet in an acre) }}{\text { (Length of all rows in the sample }[4 \times 15=60)] \text { (Width across } 8 \text { rows } / 8)}$ |
| :--- |
| $\mathrm{X} \quad$ |
| $\mathrm{X} \quad \frac{\text { Weight of Ears husked in the field }}{\text { [Number of ears husked in the field }}$ |
| $\mathrm{X} \quad \frac{\text { Weight of Grain Shelled }}{\text { Weight of Ears Shelled }}$ |
| $\mathrm{X} \quad \frac{\text { (Grain weight tested for moisture content) (1-[Moisture } \%+100])}{\text { (Ear Weight in Bag - Bag Weight) (Shelling fraction) }}$ |
| $\mathrm{X} \quad$ (Adjustment of weight to bushels) |

The detailed end of season measurements are made to facilitate forecasting. In earlier months, the plants have not developed the characteristics measured at maturity. The plant characteristics which are present at the time, the most highly correlated with final measurements, and reasonably easy to collect are measured on the units and used to forecast the components of the product above. For example, in early months, the count of stalks
is used to predict future ear numbers. More detail on the models can be found in Reiser, Fecso and Chua (1989) and Reiser, Fecso and Taylor (1987).

The above samples provide data to measure the gross or biological yield. Economic yield, the statistic of interest, requires an adjustment for loss due to harvesting. In a half sample of fields, two additional units are located after harvest. The grain left in the unit is gleaned and expanded to a per acre basis.

## 3. METHODOLOGICAL STUDIES

There is a long history of methodological study of objective yield measurement, e.g., Mahalanobis (1946) and Zarkovich (1966). The intent of this section is to outline some of the findings at NASS which are not likely to be known outside of NASS. The presentation will be organized by the components of the yield estimating equation. For each component, a brief description of problems or methods developed will be given. This is not a complete profile of all errors, although this is a long range goal, but a view of the complexity of controlling data quality in these surveys.

### 3.1 Unit Placement

The two units are located in a field using a random method designed to give all plants nearly an equal chance of selection. NASS units are determined by a randomly chosen row and then a random number of paces into the field along the row. (When rows are not visible a random number of paces along the edge of the field is substituted.) The random number of rows and paces is determined by the acreage of the field. An algorithm determines the random numbers when the fields are selected. Selection is limited to the quarter of the field which is most accessible. This restriction has not shown biasing effects when studied. Plants along the edge and in tractor turn rows are a concern. The algorithm was changed several times in the 50 's and 60 's to ensure a reasonably uniform distribution of units near the edges and in turn rows.

In tree crops, another level of randomization is necessary. Enumerators select random numbers to determine a path up the limbs to an end branch for counts. Strict procedures for marking the tree and random number selection are needed to avoid preselection of easier to reach limbs.

Pacing along a row provides the enumerator with the opportunity to create bias. Seeing the condition of the field ahead can result in shorter or longer steps which reach a visually appealing place. A buffer space is used to help reduce the possible bias. When the last pace is completed, the enumerator measures or places a pole down at the tip of their toe. The pole is usually live feet long. (For some crops, a tape measure is used to mark the five feet buffer.) The end of this buffer marks the start of the unit to be sampled. A controlled experiment in wheat indicated that counts using a buffer are lower then counts not using a buffer. Differences, on average, have been observed between the counts in unit one and unit two, another indication that subtle biases in selection can occur using pacing.

### 3.2 Acreage Measurement

The yield per acre in a field is determined by adjusting unit yield to a per acre basis. Unit size varies by crop. Units in fields where rows are evident have lengths of fixed size, but row widths vary. In fields where rows are not obvious, a fixed-size rectangular unit is created.

The area expansion to a per acre bias is subject to errors from both the length and width of the unit. Remeasurement of plots made using a tape to measure fifteen feet past the five foot buffer found several plots 20 feet long. Plots using fixed metal frames of about three feet in length may have end point inclusion or exclusion of plants which is effectively a change in plot length.

Row widths are variable. To control the variability, enumerators make two measurements, across one and four row spaces, for units used to determine gross yield. The measurements are across one and five row spaces in additional units for harvesting loss
measurements. The ratio of the four to the one row measurement had an interesting distribution. There was a large spike at four with a normal, small variance, error around four. There was also a spike at five. (The five to one post-harvest data behaved conversely.) It was speculated that instructions to measure between five rows (four spaces) may have been confused or pre- and post-harvest instructions were mixed. A year later this phenomena was mentioned in training classes. With no other changes in training, the extra spike disappeared.

## 33 Counts in the Unit

Different agronomic characteristics are measured or observed within the unit depending upon the crop and month. Fatigue and other task difficulty contribute to error in these data. Counts tend to diminish on average as work proceeds within fields, indicating fatigue problems associated with repeated counting of large numbers of items. Other tasks are difficult. Identification of maturity stage can be difficult at the borderline of a maturity category. An error in maturity stage determination contributes to measurement error in the forecasts. Counting nodes and lateral branches on soybean plants is difficult. Measuring length of corn ears or kernel rows, finding all the cotton bolls, and determining kernel formation are other difficult measurements.

### 3.4 Weight Determination

Weight per fruiting body is needed. To obtain this statistic, a subsample of fruiting bodies from the units in the field is sent to a state office for further measurements. The subsampling can cause errors. For example, the average weight of corn ears sent to the state office has been found to be larger than the average weight of all ears in the unit. Counts of wheat heads in the field sometimes vary considerably from those counted in the office. Finally, lab equipment must be controlled to obtain correct weights, moisture content, and grain to othermatter ratio.

### 3.5 Harvest Loss

Many of the problems mentioned above may occur during the harvest loss measurement. Other errors include losses from birds between harvest and the sample visit and finding all the grain on the ground, especially when a tire has run over the area.

## 4. ONGOING QUALITY EFFORTS

The assortment of typical data problems just presented leads to the question of what to do about the problems. NASS uses several activities to help control and reduce data error, including training, quality control methods and procedures, and methodological studies and improvements.

### 4.1 Training

NASS uses a train the trainer approach which follows its organizational structure. A national training session is conducted to train the statisticians from the state offices who will be in charge of the survey for their state, Often, the training is several months prior to the start of the survey. These statisticians then hold a training session in their state. They repeat the material to train the enumerators and supervisory enumerators.

There are several areas of concern with the training program. First, the consistency of enumerator training between states can be a problem in a train the trainer system. Recent budget cuts have added to this problem. The initial yield training session has been combined with training for another survey. Also some of the more expensive training supplies such as color inserts for training manuals have been eliminated.

On the bright side, there is some growing interest in using more formal evaluation methods in our training program. Concepts quizzes are used more often and evaluations of the results are stimulating discussion about optimal training procedures. For example, it appears that little is gained from allowing more than three hours of pre-training home study by enumerators. Also, quiz score improvement is most related to over two years experience. Finally, a general survey of enumerator's opinions about job needs and concerns has been done. Interesting insights are expected from the analysis which is now beginning.

### 4.2 Quality Control Measurements and Procedures

Quality control methods for survey operations which are typical of those used in many survey organizations are also used in NASS OY surveys. Since the first operational OY surveys of corn and cotton in 1961, NASS bas produced many of the usual quality measures. These measures include refusal rates, inaccessible rates, samples per enumerator, cost per sample, sampling error, and edit failure rates.

The main real-time quality control procedures are supervisory enumerator rechecks of work and edits of completed forms. Supervisory enumerators return to a sample done in the first day's work of each enumerator and one randomly selected sample of later work to recheck counts. This process is mostly a continuation of training as the enumerator often accompanies the supervisor. Too little data is collected for statistical adjustment of the full survey data. Also, some counts change naturally as the plant matures between time of the visits.

Data is key entered and verified as the survey progresses. Edits of the data are done in batch mode periodically throughout the survey. Edits currently are simple range and consistency checks. An area with potential to improve data quality is the development of smarter edits. For example, looking for outlier and leverage points in regressions of variables such as weight and number of fruiting bodies. Range checks can vary based on categorical information, for example, narrow row soybean yield has a higher mean than wide row plantings. Data capture with portable data recorders was studied seven year ago and is again a research project. Data transmission problems which halted past work are mostly resolved, but cost may still be a limitation given the small sample size per enumerator. Computerization of weight measurement in the state office has been successfully initiated, thus reducing transcription errors.

### 4.3 Methodological Studies and Improvements

NASS uses research studies designed to test specified hypothesis about errors or to test new procedures as well as general validation studies to guide the process of survey change for improvement. Examples of the results of special studies have been presented. Validation studies have a long history at NASS. The design of these studies, prior to 1987, typically consisted of one or two states with 16 to 32 fields selected. The units in the field would be replicated up to 32 times per field to measure procedural error sources. Although results have varied by state, year and crop, a meta analysis of these studies by Warren(1985) indicates procedural bias of 6 to 9 percent for corn. Although not formally documented, similar results are found in soybeans. These biases remain despite thirty years of study.

The recurring bias level led us to evaluate our validation and research methods. With yield being the product of many measurements, all of which had several potential sources of error, we questioned the ability of control procedure to achieve the state of "do it right the first time." Fortunately, we had an alternative. A double sampling approach to adjust for bias was developed. (Fecso,1986) Briefly, the new design collects actual yield for a subsample of fields by weighing the harvested grain. This is a rare example of a survey in which a true value can be obtained. Bias is estimated as the difference between the survey procedures and the harvested weight for the paired observations.

This new validation approach has been used for soybeans with the following results:

| YEAR | Estimated Bias (bushels) | Standard Error of Estimate |
| :--- | :---: | :---: |
|  | 2.2 | .9 |
| 1987 | 1.9 | .8 |
| 1989 | 3.2 | .9 |

The design met its main criterion, to detect biases of over $5 \%$ (about 1.6 bushels). The estimated biases for these three years have been in the 6 to 9 percent range. Large shifts in the bias will also be detectable, an important feature when survey conditions are changing over time. Preliminary data for 1990 indicates that a shift may have occurred.

Although this double sampling estimator puts the survey estimate on a fully statistical basis (rather than assuming some bias level or inferring from two states to all states), the variance of the estimate is larger than we would
like. Further work on allocation of resources can help as some aspects of the survey appear to be oversampled compared to others.

We may want to pursue further development of process control methods for survey operations as we restaff for yield research. Process control techniques used in industry can be modified for survey use. Simple time series of measures like the four-row to one-row measurement or the unit one to unit two difference can provide clues to changes in the essential survey conditions. More complex statistical techniques such as LISREL (Reiser, Fecso and Chua, 1989) could also be useful.

## 5. CONCLUSION

The intent of this paper is to provide a feel for the number and complexity of tasks associated with the visually and mechanically measured values needed to measure yield. Interested readers are encouraged to contact the author for more detail and especially to open dialogue about their experiences with crop yield measurements.

Total quality management principles can be applied to our renewed yield research efforts. We need to determine our most important quality problems so we may better balance the need for increased data and accuracy with the pressure of budget and personnel limitations and the desire for more timely data. The shifting emphasis on these elements over time creates changes to the survey which often create bias in the estimate, sometimes unexpectedly. Thus, we look to techniques such as the validation methodology and process measurements to point out accuracy changes and to help us detect important error sources. If successful we will have an efficient way to direct our limited resources to develop needed methodological improvements in accuracy, ease of conduct or cost reduction. Besides these "quality control" ideas, there is anticipation that exploring the panel, spatial and multivariate nature of the data will bring further efficiencies.

## REFERENCES

Fecso, R. (1986). Sample Survey Quality: Issues and Examples from an Agricultural Survey, Proceedings of the Section on Survey Research Methods, American Statistical Association.

Fecso, R. (1990). A Review of Errors of Direct Observation in Crop Yield Surveys, in publication review.
Francisco, C. A., Fuller, W.A., and R. Fecso (1987). Statistical Properties of Crop Production Estimators, Survey Methodology, 13, 1, 45-62.

Mahalanobis, P. C. (1946). Recent Experiments in Statistical Sampling in the Indian Statistical Institute. Joumal of Royal Statistical Society, 109(4), 325-370.

Reiser, M., Fecso, R., and Chua, M. (1989). Some Panel Aspects of the Objective Yield Survey, Proceedings of the Section on Survey Research Methods, American Statistical Association.

Reiser, M., Fecso, R., and Taylor, K. (1987). A Nested Error Model for the Objective Yield Survey, Proceedings of the Section on Survey Research Methods, American Statistical Association.

Warren, F.B. (1985). Corn Yield Validation Studies, 1953-83, Washington, D.C., USDA, SRS, Staff Report No. YRB-85-07.

Zarkovich, S.S. (1966). Quality of Statistical Data, Rome, Food and Agricultural Organization of the United Nations.

## SESSION 6

Measurment and Improvement of Quality of Administrative Records When Used as a Replacement to Traditional Survey Taking

# CANADIAN TAXFILER FAMILY DATA: ASSESSMENT OF QUALITATIVE CRITERIA AND TRENDS 

J.M. Leyes ${ }^{1}$


#### Abstract

Statistics Canada has created family data from individual tax records. These data have been evaluated in this paper relative to survey/census data for five selected criteria (i.e. concepts and definitions, population coverage, population coverage bias, income coverage bias and range of variable). The 1988 tadiler family data were compared to the 1989 population estimates for a cross-sectional comparison. The tax data time series for the period 1982-1987 were also compared to the Survey of Consumer Finances, an annual supplement to the Labour Force Survey. In addition, there is a discussion of future directions that are expected to contribute to improvements in the coverage of the tax family data.


KEY WORDS: Biases; Census of Population; Macro-comparisons.

## 1. INTRODUCTION

In 1979, Statistics Canada began a formal review of the potential of using administrative records for social statistical applications for small area data (Statistics Canada, 1979). Based on this review, it was concluded that the highest coverage of the population and the greatest potential for social administrative data would arise through the use of the personal income tax records, or the T 1 records. In this paper the discussion is limited $t 0$ taxfiler data comparisons with (a) the population estimates and (b) the Survey of Consumer Finances.

From the beginning of this data development program, the work with the personal income tax file was subject to a number of expected a priori shortcomings. The shortcomings included the following:
i. Population Coverage Bias. The income tax system is based on individuals only. Since only $60 \%$ of Canadians were filing tax returns in the mid to late 1970's, coverage was deemed inadequate for social statistical applications.
ii. Age Coverage Bias. The age profile of taxfilers also differed from the age profile of the population. This was judged to be an unacceptable bias.
iii. Income Distribution Bias. The elderly and the young frequently have low incomes and do not file tax returns. Data from the tax file would, therefore, be inadequate for public policy purposes directed at these larget groups.
iv. Income Coverage Bias. Since not all income received by Canadians is taxable, the income coverage of the T1 was considered incomplete.
v. Dimensionality of Variables. Since any single administrative record has a specific and narrow application in program administration, the range of data variables was judged inadequate as a source of social data.

[^23]vi. Concepts and Definitions. The concepts and definitions used in household surveys and censuses of population can only be approximated through the use of an annual tax file.

Each of the above represented a limitation or shortcoming of data derived from administrative records in general, the T 1 in particular. Since the time that the above assumptions were identified, data for taxfiler families have been developed. Beginning in section 3 of this paper, the first three assumptions are assessed with empirical findings from the taxfiler family data for the period 1982-88.

## 2. THE DEVELOPMENT OF TAXFILER FAMILY DATA

The taxfiler family concept has been designed to emulate the census family concept. A census family:
" [rjefers to a husband and a wife (with or without children who have never married, regardless of age), or a lone parent of any marital status, with one or more children who have never married, regardless of 8age, living in the same dwelling. For census purposes, persons living in a common-law type of arrangement are considered as now married, regardless of their legal marital status; they accordingly appear as a husband-wife family in census family tables." (Statistics Canada, 1982, p. 29)

This concept is suitable for household collection methods since respondents are asked to report on the relationships between all residents of a dwelling. With administrative records, secondary information such as reported marital status, value of exemptions/tax credits, ages of taxfilers, addresses, child care expenses, and so forth, are used to form families.

It has not, therefore, been possible to emulate the exact census family concept. The major sources of difficulty arise with older children (whether they have been married or not when they reside with their parents) and with common law couples. In general, the census family concept works reasonably well for families with dependent children, and some success has been achieved in estimating single parent and common law families.

In 1984, Statistics Canada began estimating families from the individual taxfiler (T1) data. The creation of families from the T1 is based on a six-step process.
i. Taxfilers, reporting the Social Insurance Numbers (SIN) of their spouses, are matched to form husband-wife families.
ii. Other husband-wife families are formed from taxfilers who declare themselves married but do not report spousal SINs.
iii. Filing children ${ }^{2}$ who reside with their parents are matched to their parents.
iv. There is an intermediate step to unduplicate records, to identify one-filer husband-wife family units, to assign a unique postal code to family members, and to assign a family composition type to each family unit.
v. Common law spouses are matched from the pool of individuals classed as single parent families and nonfamily persons.
vi. In Step 6, non-filing family members are imputed.

With this brief introduction and description of the taxfiler family data, it is now possible to consider some data findings.

[^24]
## 3. SOME EMPIRICAL FINDINGS

### 3.1 Population Coverage Comparison: 1982-88 Taxfiler (T1) Family File (T1FF) to 1983-89 Population Estimates ${ }^{3,4}$

In this section, two tables have been prepared. Table 1 contains comparisons at the Canada level for both population and family composition types. Table 2 contains some population comparisons for the provinces, territories and Canada.

### 3.1.1 Population Coverage Bias: 1988 Taxfiler Family File (T1FF) to Statistics Canada's 1989 Population Estimates

The T1 family file (T1FF) data that are reported in Table 1 have been placed into six classifications for comparison.

- Section one (lines 1-2) includes a count of the number of taxfilers and the year-to-year percent change.
- Section two (lines 3-7) compares the estimated taxfiler family (T1FF) population and the official Statistics Canada population estimates. (In creating the T1FF data, a record is created for each family member and for each non-family person, including the imputed population. Thus, line three of Table 1 is the T1FF population count.)
- Section three (lines 8-12) compares the number of husband-wife T1FF families to the official Statistics Canada estimate of husband-wife families. (Husband-wife families, in both cases, are defined to include common law families.)
- Section four (lines 13-17) compares the number of single parent T1FF families to the official Statistics Canada estimate of single parent families.
- Section five (lines 18 -21) provides an estimate of common law familics. The only time that there is another estimate of common law families is during census years. Thus, there is only a comparison for 1985.
- Section six includes the count of T1FF non-family persons for each year from 1982-1988. (Only the SCF includes an annual estimate of non-family persons.)

There are several highlights in Table 1:

- The T1FF population (taxfilers plus imputed) increased at a slightly lesser rate than the number of taxfilers alone (i.e. Table 1, lines 2 and 4);
- The T1FF population varied betwcen 93.7 and 95.9 percent of the official estimates over the $1982-88$ period (Table 1, line 7);

The T1FF coverage of husband-wife families had almost reached 98 percent in 1988 (Table 1, line 12);

[^25]Single parent overcoverage decreased over the $1982-88$ pcriod (line 17). At the same time single parent coverage was decreasing, common law coverage was increasing as a result of improved linkage procedures ${ }^{6}$; and

The T1FF estimate of non-family persons was lower than the SCF estimate for each year (lines 22 and 24) for the first five years and higher for the two most recent years. The year-to-year changes were similar for each source (lines 23 and 25).

Table 1.
Summary of Comparisons Between T1FF and Population Estimates Data: 1982-1988

| $\begin{aligned} & \text { ILINE } \\ & \text { NO. } \end{aligned}$ | TAX YEARS |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | 1982 | 1983 | 1984 | 1985 | 1986 | 1987 | 1988 |
| 'MUMBER Of TAXFILERS (Including all provinces and territories.) |  |  |  |  |  |  |  |
| 1. Taxfilers ( , 000) | 15,166 | 15,243 | 15,467 | 15,526 | 15,971 | 16,687 | 17,251 |
| 2. \% change | , 166 | 0.5 | 1.5 | 0.4 | 2.9 | 4.5 | 3.4 |
| TOTAL POPULATIOM (including all provinces and territories) |  |  |  |  |  |  |  |
| 3. Tax Pop. ( , 000) | 23,628 | 23,725 | 23,736 | 23,839 | 24,016 | 24.838 | 25,155 |
| 4. \% change |  | 0.4 | 0.0 | 0.4 | 0.7 | 3.4 | 1.3 |
| 5. Pop. Est. ( , 000) | 24,787 | 24,978 | 25,165 | 25,353 | 25,617 | 25,912 | 26,219 |
| 6. $\chi$ change |  | 0.8 | 0.8 | 0.8 | 1.1 | 1.2 | 1.2 |
| 7. Coverage (3)/(5) | 95.3 | 95.0 | 94.3 | 94.0 | 93.7 | 95.9 | 95.9 |
| MUSBAND-WIFE FAMILIES (Excluding the Yukon and Northwest Territories) |  |  |  |  |  |  |  |
| 8. Fax Pop. ( , 000) | 5,510 | 5,524 | 5,570 | 5,528 | 5,592 | 5.753 | 5,866 |
| 9. \% change |  | 0.3 | 0.8 | -0.7 | 1.1 | 2.9 | 2.0 |
| 10. Pop. Est. ( , 000) | 5,722 | 5,773 | 5,824 | 5,875 | 5,932 | 5.987 | 5,905 |
| 11. X change |  | 0.9 | 0.9 | 0.9 | 1.0 | 0.9 | 0.1 |
| 12. Coverage (8)/(10) | 96.3 | 95.7 | 95.6 | 94.1 | 94.3 | 96.1 | 97.9 |
| SINGLE PARENT FAMILIES (Excluding the Yukon and Northwest Territories) |  |  |  |  |  |  |  |
| 13. Tax Pop. ( , 000) | 830 | 873 | 894 | 934 | 940 | 965 | 937 |
| 14. \% change |  | 5.1 | 2.4 | 4.5 | 0.6 | 2.7 | -2.9 |
| 15. Pop. Est. ( , 000) | 768 | 796 | 824 | 852 | 882 | 912 | 936 |
| 16. \% change |  | 3.6 | 3.5 | 3.5 | 3.5 | 3.4 | 2.6 |
| 17. Coverage (13)/(15) | 108.1 | 109.7 | 108.5 | 109.6 | 106.5 | 105.7 | 100.1 |
| COMMON LAW FAMILIES (Included | Line 8 | Includ | ng all | ovinces | d terri | ories.) |  |
| 18. Tax $(, 000)$ | 205 | 207 | 242 | 221 | 281 | 336 | 365 |
| 19. ${ }^{19}$ \% change |  | 0.8 | 16.8 | -8.5 | 27.3 | 19.4 | 8.6 |
| 120. Census (,000) |  |  |  | 487 |  |  |  |
| 121. Coverage (18)/(20) |  |  |  | 45.4 |  |  |  |
| (NON-FAMILY PERSONS (Excluding the Yukon and Northwest Territories) |  |  |  |  |  |  |  |
| 22. Tax (,000) | 2,973 | 3,071 | 3.153 | 3.269 | 3.540 | 3.794 | 4.096 |
| 23. $\quad$ x change |  | 3.3 | 2.7 | 3.7 | 8.3 | 7.2 | 8.0 |
| 24. SCF (,000) | 3,282 | 3,348 | 3.433 | 3.512 | 3.625 | 3.770 | 3,829 |
| 25. $\quad$ \% change |  | 2.0 | 2.5 | 2.3 | 3.2 | 4.0 | 1.6 |

Sources - T1FF: Preliminary, unpublished. 1982-87, 5\% sample. 1988, 100\% tabulations. Population Estimatcs. Catalogue Nos. 91-204, 91-210, 91-529.
SCF: 1982 to 1988: Family Incomes -- Census Families -- 1988, Cat. No. 13-208, Table 14, Jan.'90.

[^26]
### 3.1.2 Provincial Population Coverage Comparison: 1988 Taxfiler Family File (T1FF) to 1989 Population Estimates

A comparison between the T1FF estimates and the official population estimates is incomplete without an indication of the sub-national estimates. A comparison has, therefore, been included for the ten provinces and two territories for the 1988 tax year.

It can be noted in Table 2 that the provincial coverage ranged from 99.8 percent for Manitoba to 93.8 percent for Prince Edward Island.

Table 2.

Population Comparison by Province and Territory, 1988 T1FF to 1989 Population Estimates


## 32 Age Coverage Bias, 1988 TIFF to 1989 Population Estimates

In Table 3, some broad age range comparisons have been included. For the first age range (i.e., 0-18) since a considerable amount of imputing is required ( $92 \%$ imputed), it has not been possible to identify finer age breakdowns, and there is no gender information.?

It should also be noted that an assumption has been made in developing the family data regarding children living with their parents, namely, that the maximum age of a matched filing child would be 29 .

In reviewing column 4 of Table 3 (i.e., \% ratio), it can be noted that the coverage of the T1FF to the 1989 population estimates was $90 \%$ or higher for age ranges under $65 .{ }^{8}$ The T1FF coverage of the 1989 estimated population declined for the population $65+$.

[^27]There are several highlights in Table 3:

- The population under 19 is overcovered by $2.6 \%$.
* For the six age ranges between 19 and 59 , the lowest coverage was the $45-49$ age range ( $95.1 \%$ ) and the highest was for the $30-34$ age range ( $97.4 \%$ ).
- After the 55-59 age range, the age coverage declines to $70.2 \%$ for the 75 and over population. (Although not reported in this table, the overall coverage of the population aged 65 and over in 1988 was about 80 percent.)

Table 3.
Comparison by Age Group:
1988 T1FF to 1989 Population Estimates

| AGE GROUP | $\begin{gathered} 1988 \mathrm{T1FF} \\ (, 000) \end{gathered}$ | $\begin{gathered} 1989 \text { POP } \\ (, 000) \end{gathered}$ | $\begin{gathered} \% \\ \text { RATIO } \\ \text { (TIFF/POP.) } \end{gathered}$ | $\begin{gathered} 1988 \text { T1FF } \\ \text { IMPUTED } \\ (, 000) \end{gathered}$ | $\begin{gathered} \chi \text { OF } \\ \text { T1FF } \\ \text { IMPUTED } \end{gathered}$ |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 0-18 | 7,165.4 | 6,982.2 | 902.6 | 6,623.2 | 92.4 |
| 19-29 | 4,645.2 | 4,820,1 | 96.4 | 123.6 | 2.7 |
| $30=34$ | 2,277.9 | 2,338.4 | 97.4 | 86.6 | 3.8 |
| 35-39 | 2,038.2 | 2,116.5 | 96.3 | 93.8 | 4.6 |
| 40-44 | 1,809.3 | 1.898 .3 | 95.3 | 106.6 | 5.9 |
| 45-49 | 1,412.9 | 1,486.0 | 95.1 | 121.3 | 8.6 |
| 50-54 | 1,205.3 | 1.249.5 | 96.5 | 143.8 | 11.9 |
| 55-59 | 1,155.4 | 1,214.3 | 95.1 | 162.5 | 14.1 |
| 60-64 | 1,052.5 | 1.142.1 | 92.2 | 158.9 | 15.1 |
| 65-69 | 914.2 | 1,026.9 | 89.0 | 124.8 | 13.7 |
| 70-74 | 646.0 | 758.4 | 85.2 | 83.3 | 12.9 |
| $75+$ | 832.3 | 1,185.6 | 70.2 | 75.7 | 9.1 |
| TOTAL | $25,154.7$ | 26,218.5 | 95.9 | 7,904.1 | 31.4 |

Sources - 1988 T1FF: $100 \%$ T1FF tabulations (provinces + territories), unpublished.
1989 Pop.: (June 1) preliminary postcensal pop. estimates, 91-210, Table 2, Feb.'90 (provinces + territories).

### 3.3 Income Distribution Bias, T1FF to Survey of Consumer Finances

The third assumption noted on page 1 pertained to the expected bias that would be found in the income data derived from tax records. The general basis of this expectation can be stated as follows:
a. In the general population, some individuals have low incomes and do not file tax returns, therefore,
b. measures of central tendency from the tax data will be higher than similar measures from the general population.

To assess this assumption, it is possible to make comparisons with the T1FF data from either the Census of Population or with the annual Survey of Consumer Finances. Since there would only be one year for comparison with the Census (i.e., 1986), and since there are seven years for comparison with the SCF, a comparison was undertaken between the TIFF and the SCF data.

For this comparison, two different groups were compared: families and non-family persons (or unattached individuals).

### 33.1 Median Income Comparison for Families: T1FF to SCF, 1982-88

Table 4 includes a time series comparison of median incomes between the T1FF and the Survey of Consumer Finances (SCF) ${ }^{9}$ for the period 1982-88.

There are several highlights in Table 4:

- The TIFF medians were lower for all years.
- The T1FF medians were about $95 \%$ of the SCF for the period 1982.85.
- With the introduction of the Federal Sales Tax Credit in the 1986 tax ycar, the T1FF median fell to $92.0 \%$. (This tax credit was designed for low income individuals and those who received non-taxable incomes. Both of these groups are the ones that it was assumed would be poorly covered by the tax data.)
- In 1988, the income coverage ratio increased to $95.6 \%$, the highest coverage ratio that was found in any of the years in this time scries.

Table 4.
Income Comparison for Families, 1982-88:
T1FF to Survey of Consumer Finances.


Sources - T1FF: For 1982 to 1987, tabulations were from $5 \%$ samples. For 1988, $100 \%$ T1FF tabulations were used. (The Territories excluded for comparison purposes).
SCF: 1982 to 1988, Cat. No. 13-208, annual publication.

While a definitive conclusion cannot be drawn from this time scries, it does suggest that the tax data may cover the low income population better than one might expect. Of course, two programs have been introduced that have worked to increase the probability that low income individuals will file tax returns. First, the Child Tax Credit was introduced in 1978, a program directed at low income families with children. Second, there was the Federal Sales Tax Credit in 1986 that was directed at all individuals with low incomes. With the expected introduction of the Goods and Services Tax (GST), it is expected that there will be another surge in tax filing

[^28]by the low income population since this program is similar to the Federal Sales Tax Credit but with a higher refundable tax credit. ${ }^{10}$

### 3.3.2 Median Income Comparison for Non-Family Persons: TIFF to SCF, 1982-88

In Table 5, data are included from the two sources for the period 1982-88.
There are several highlights in Table 5:

- For the years $\mathbf{1 9 8 2 - 8 5}$ for non-family persons, the data support the assumption that data from the general population provide medians that are lower than medians derived from lax data. For these years, the percentage coverage ratios varied between $117.1 \%$ and $124.2 \%$.
- With the introduction of the Federal Sales Tax Credit in 1986, the coverage ratio declined to $104.8 \%$ in 1986, then to 99.4 percent in 1987, and in 1988 it rose to $105.8 \%$.

For non-family persons, the assumption appears to be supported. The assumption was reinforced with the predicted impact of the Federal Sales Tax Credit: low income individuals would file tax returns and the median would be lower when compared to the SCF.

Table 5.
Income Comparison for Non-Family Persons, 1982 - 1988 TifF to Survey of Consumer Finances


Sources T1FF: 1982 to 1987: tabulations drawn from 5\% samples; in 1988, 100\% T1FF tabulations (territories excluded for comparison purposes).
SCF: 1982 to 1988, Catalogue 13-208, Annual publication.

[^29]
## 4. MAJOR DIRECTIONS FOR 1989 +

Two new initiatives are in progress that utilize the taxfiler family data.

- A pilot Longitudinal Administrative Database (LAD) was under development during the period 1989-90. This database has been developed to enable research studies of poverty/welfare/income dynamics in Canada for the period 1982-86. The LAD was designed as a $10 \%$ sample to parallel the Panel Survey of Income Dynamics (PSID) that was begun by the Survey Research Center, University of Michigan about 20 years ago. (Duncan, 1984)
- Work is currently underway to develop an Administrative Record Consolidation File (ARC) through the linking of multiple records on a sample basis. This consolidated file will enable the taxfiler family data to be modelled to (a) improve the coverage of the population and (b) improve some of the variables on the taxfile (e.g., ages of children, reasons for which individuals received Unemployment Insurance benefits).


## 5. SUMMARY AND RELATED OBSERVATIONS

The TIFF data possess positive characteristics. The data are annual and small area estimates can be produced.
Regarding the overall quality of the T1FF data, if $95 \%$ is high coverage, the comparisons in this paper have indicated a high coverage of the population by the T1FF. Furthermore, the general trends indicate gradual improvements during the period 1982-88. There does, nonetheless, continue to be undercoverage of individuals with low incomes (typically the young and especially the elderly), although the introduction of the Refundable Federal Sales Tax Credit in 1986 led to a general improvement in the coverage of the low income population. Finally, and based on previous experience, it seems reasonable to expect that both the population coverage and the low income coverage will improve further over the next few years with the introduction of the GST Tax Credit.

## REFERENCES

Duncan, G.J. (1984). Years of Poverty, Years of Plenty. Survey Research Center, Institute for Social Research, The University of Michigan, Ann Arbor, Michigan.

Statistics Canada (1982). 1981 Census Dictionay. Statistics Canada, 1981 Census of Canada. Catalogue Number 99-901, Ottawa, Canada.

Statistics Canada (1988). Intencensal Estimates of Families, Canada, Provinces and Territories, 1981-1986. Catalogue Number 91-529, Ottawa, Canada.

Statistics Canada (1990). Postcensal Estimates of Families, Canada, Provinces and Territories, 1987, 1988, 1989. Catalogue Number 91-204, Ottawa, Canada.

Statistics Canada (1990). Postcensal Estimates of Population by Marital Status, Age, Sex and Components of Growth for Canada, Provinces and Terriories, 1989, Catalogue Number 91-210. Ottawa, Canada.

Statistics Canada (1990). (Annual). Family Incomes: Census Families, 1988. Survey of Consumer Finances. Catalogue No. 13-208 (Annual), Ottawa, Canada.

## SESSION 7

Improving the Quality of Frame Data and Their Use

# SAMPLING FLOWS OF MOBILE HUMAN POPULATIONS 

G. Kalton ${ }^{1}$


#### Abstract

Surveys are often conducted of flows of persons, such as: visitors to museums, libraries and parks; voters; shoppers; hospital outpatients; tourists; international travellers; and car occupants. The sample designs for such surveys usually involve sampling in time and space. Methods for sampling flows of human populations are reviewed and illustrated.


KEY WORDS: Mobile populations; Exit polls; Traffic surveys; Time and space sampling; Systematic sampling.

## 1. INTRODUCTION

Most surveys of human populations are household based, typically with a sample of households selected with a multi-stage sample design, and individuals sampled within the selected households. The household survey is a powerful method for collecting data on a wide range of characteristics about the population, such as social, demographic, economic and health characteristics and the population's opinions and attitudes. The method is, however, not so effective for studying the characteristics of mobile populations. Two types of mobile populations may be distinguished: those who do not reside regularly at a fuxed location, such as nomads and the homeless; and members of the general population who belong to the mobile population under study because they are in transit, such as visitors to libraries and parks, voters at polling booths, shoppers, hospital outpatients, travellers, and car occupants. This paper reviews sample design issues for this latter type of mobile population.

Although there are many surveys concerned with flows of mobile human populations, the general sampling literature contains little discussion of the sampling issues involved. The purpose of this paper is to describe the sample designs commonly adopted for surveys of flows of human populations, to discuss some of the special sampling issues faced, and to illustrate the range of applications for such surveys. The next section of the paper reviews the general time and space sample design used for sampling persons in transit and some of the issues involved in employing this design in particular situations. Section 3 then illustrates the application of the design in a range of different settings. Section 4 presents some concluding remarks.

## 2. SAMPLING IN TIME AND SPACE

It will be useful to consider a specific example in describing the general time and space sample design for sampling flows of human populations. Suppose that a survey of visitors to a summer sculpture exhibition in a city park is to be conducted to find out the visitors' socio-economic characteristics, how they heard about the exhibition, what means of transport they used to get to the park, and perhaps their views of the exhibition. Suppose that the exhibition is held from April 1 to September 30 in the year in question, that it is open from $10 \mathrm{a} . \mathrm{m}$. until 6 p.m. daily, and that there are three sites where visitors enter and leave the exhibition.

The sampling frame for a survey of this type is usually taken to be a list of time interval/site primary sampling units (PSUs). This frame is constructed by dividing the time period of the survey into a set of time intervals for each site. A simple construction of PSUs for the current example would be to divide each exhibition day at each

[^30]sitc into two time intervals, one from 10 a.m. until 2 p.m. and the other from 2 p.m. until 6 p.m. A more complex construction of PSUs could involve time intervals of different lengths on different days and/or at different sites. Once the PSUs are defined, a two-stage sample design is often employed. At the first stage a sample of PSUs is selected, and at the second stage a sample of visitors is drawn, usually by systematic sampling, in the sampled PSUs.

The actual specification of the sample design for a survey of persons in transit within the two-stage sampling framework depends on features of the mobile population under study and of the survey data collection procedures. A key feature is the nature of the flow of the mobile population. In particular, is there a predictable variability in the rate of flow across PSUs? For instance, is the flow at one site higher than at another site, or are the flows at some time intervals (say, Saturday afternoons) higher than others? Also, is the flow within a PSU a smooth one throughout the time interval or is it uneven, with visitors arriving (or leaving) in sizeable groups? Both these aspects of flow affect the sample design for the survey.

If the flow is fairly uniform across the PSUs, and if the PSU time intervals are the same, then the number of visitors per PSU is approximately constant. In this case, the PSUs may be sampled with equal probabilities, and a constant subsampling fraction can be applied within the selected PSUs to generate an equal probability, or epsem, sample of visits. The PSUs can be classified in two or more dimensions (e.g., day of week, time of day, and site), and a carefully balanced sample across these dimensions can be obtained using lattice sampling (Yates, 1981; Cochran, 1977; and Jessen, 1978).

In many cases, the level of flow varies across the PSUs in a manner that is partly predictable. For instance, the attendance at the sculpture exhibition may be known to be generally higher in the later shift each day and at the weekends, and particularly low on Mondays. Thus the PSUs comprise different numbers of visitors, that is, they are PSUs of unequal sizes. The usual procedure for handling PSUs of unequal sizes is to sample them with probabilities proportional to their sizes (PPS), or estimated sizes (PPES). In the current context, the actual PSU sizes are not known in advance, and estimated sizes must therefore be used. Sampling the PSUs with PPES works well provided that reasonable estimates of the sizes can be made. When PSUs are selected by PPES sampling, then the application within the selected PSUs of subsampling fractions that are inversely proportional to the estimated sizes of the PSUs produces an overall epsem sample of visits. In general, an attraction of PPES sampling (with reasonable estimates of size) is that the subsample sizes in the PSUs do not vary greatly from one PSU to another. This feature is of especial value for conducting the fieldwork in surveys of persons in transit. When time/site PSUs are sampled by PPES sampling, lattice sampling cannot be applied for deep stratification. Instead, controlled selection may be employed for this purpose (Goodman and Kish, 1950; Hess et al., 1975).

An important consideration in any two-stage sample design is the allocation of the sample between first- and second-stage units, that is, how many PSUs to select and how many elements to select per sampled PSU. In the case of surveys of persons in transit, that allocation is strongly affected by the fieldwork procedures to be used and the nature of the flow within the PSUs. The aim of the design is to make full use of the fieldworkers assigned to a sampled PSU while maintaining a probability sample of persons entering (or leaving) the site during the sampled time interval.

Many surveys of persons in transit use self-completion questionnaires, in which case the fieldwork process for the two-stage design described above consists of counting persons as they enter (or leave) the sampled site during the time interval, selecting every $k$ th person for a systematic sample, and asking the selected persons to complete the questionnaire. If the flow is light and evenly spread throughout the time interval, one fieldworker may be able to handle all the tasks involved. When this is so, the sampling interval $k$ can be chosen to give the fieldworker time to perform all the tasks in an unpressured way. If, however, the flow is heavy, either constantly or intermittently, two fieldworkers may be needed, one simply to count entrants (or leavers) and identify sampled persons, and the second to hand out the questionnaires and to instruct respondents on how they should be completed and returned. With this fieldwork arrangement, the sampling interval can be chosen to keep the second fieldworker as fully occupied as possible, while making sure that he or she is able to distribute questionnaires to all (or at least nearly all) of those sampled. Nonresponse can be a major concern with the selfcompletion mode of data collection. It is often possible to keep nonresponse to an acceptable level when sampled persons complete and return the questionnaire at the site. However, when they are handed the
questionnaire with the request to complete it later and return it by mail, the level of nonresponse can be very high and, moreover, there is generally no way of following up the nonrespondents.

When face-to-face interviewing is used for data collection, the fieldwork team for a PSU usually contains one counter and a small team of interviewers. The size of the interviewer team depends on the regularity of the flow and the length of the interview. Since persons in transit are likely to be unwilling to be delayed for long, interviews are necessarily mostly short. Longer interviews may, however, be possible if the sampled persons are in a waiting mode, such as waiting in line or in an airport departure lounge. The choice of sampling interval has to be such that there is always (or nearly always) an interviewer free to interview the next sampled person, and that the interviewers do not spend too much time waiting for the next sampled person to be selected. If the flow is irregular, allowance needs to be made to accommodate the peaks (for instance, the arrival of a coachload of visitors to the sculpture exhibition).

The PPES selection of the PSUs works to equate the subsample size for each sampled PSU. For face-to-face interview surveys, the interviewer load is thus roughly the same for each selected PSU, and hence the same-sized interview team can be used for each PSU. A problem occurs, however, when the PPES measure used in selecting the PSU at the first stage is seriously in crror. For example, a thunderstorm may substantially reduce the number of visitors to the sculpture exhibition on a particular Saturday afternoon, or an unforeseen holiday may substantially increase the number on another day. In the first case, applying in that PSU a sampling interval inversely proportional to its estimated size will leave the interviewers largely unoccupied, whereas in the second case it will result in a workload that the interviewers cannot handle. A modification that may be adopted in such cases is to change the sampling interval at the start of data collection to one that is more suitable for the flow actually encountered. Since this modification destroys the epsem property of the sample, weights are needed in the survey analysis.

A general limitation to the systematic sampling of visitors at selected PSUs is that if the sampling interval is made long enough to enable interviewers to cope with peak flows, they spend much of their time without work. On the other hand, if the sampling interval is reduced, the interviewers are more fully occupied, but they cannot cope with peak flows. Various methods have been proposed to circumvent these problems (Heady, 1985). One procedure is to take a systematic sample of times (say, every 10 minutes) and to select the next visitor to enter after each sampled time. This procedure might have fieldwork attractions, but it does not produce a probability sample of visitors. Persons arriving in busy periods are less likely to be chosen, as are those who travel in groups, and the walking habits of persons travelling in groups may affect the chances of selection in unknown ways. The sample generated by this procedure is clearly not an epsem sample. An attempt can be made to compensate for the selection bias that operates against visitors arriving in busy times by dividing the time interval for selected PSUs into a set of much shorter intervals, and keeping a log of arrivals in each such interval. Then weighting adjustments can be employed to compensate for the variation in the flow across the shorter intervals.

Another alternative procedure to systematic sampling of visitors is to take the next person to enter (or leave) after the last interview was completed. With this procedure, the first persons to arrive after gaps in the flow, perhaps the leaders of groups, clearly have greater chances of selection. Also interviewers may deliberately speed up or slow down their current interview in order to avoid or to select a particular individual. For these reasons, variants on this procedure that select the $n$th person after the completed interview, where $n$ might be set at 2 , 3,4 , or 5 , have been employed. Any version of this procedure produces, however, a nonprobability sample, with the risk of bias that entails. These alternatives to a straightforward systematic sampling of visitors may make much more effective use of fieldworkers' time, but there is a price to pay in moving from a probability to a nomprobability design.

Visitors may be sampled either as they enter or as they leave a location. If data about the visitors' activities in and opinions of the location are required, then leavers need to be sampled. In other cases, the choice between sampling entrants and leavers may depend on the nature of the flows. It may, for example, be difficult to sample and interview people leaving a theatre because they leave en masse and because they will not want to be delayed. On the other hand, they may be readily sampled and interviewed as they line up to enter the theatre.

In concluding this section, attention should be drawn to the fact that the samples described here are samples of visits not visitors. The standard two-stage design may produce an epsem sample of visits, but this is not the same
as an epsem sample of visitors unless each visitor visits the place under study (the sculpture exhibition) only once (or they all visit the same number of times). For most flow surveys, the visit, rather than the visitor, is the appropriate unit of analysis. There are, however, situations where the analytic unit is problematic. Using the visit as the unit of analysis, the researcher might readily accept visits to the sculpture exhibition on two separate days as distinct visits, but might not be willing to treat two entries on the same day (one, perhaps, after leaving briefly for refreshments) as two visits. The use of the visitor as the unit of analysis presents severe problems because of the problems of multiple visits, and the fact that visitors will not be able to report their multiplicities. They may be able to recall past visits reasonably accurately, but they will usually be unable to forecast future visits accurately.

## 3. SOME EXAMPLES

This section presents some examples of surveys of flows of buman populations in order to indicate the wide range of applications and to illustrate some of the special considerations that arise in particular settings.

### 3.1 A survey of library use

A survey of the use of the 18 libraries at the University of Michigan was conducted in 1984 (Heeringa, 1985). Each sampled person exiting the library was asked whether he or she had used the library's materials and services during that visit. If so, the person was asked to complete a short self-completion questionnaire of seven questions on the materials and services used. Most of the 5184 respondents completed the questionnaires on the spot and returned them to the survey fieldworkers; others sent them back by campus mail. A response rate of $96 \%$ was obtained.

The sample design followed the two-stage time/site sample design described in Section 2. The survey covered the full 1984 calendar year. Each day the libraries were open was divided into 10 two-hour time intervals, starting at $7.30 \mathrm{a} . \mathrm{m}$. and lasting until $3.30 \mathrm{a} . \mathrm{m}$. the next morning, the two-hour interval being chosen on the grounds that it was a suitable shift for the fieldworkers. The PSUs were then defined to be time interval/library combinations. The PSUs were selected by PPES sampling, where the estimated size for a PSU was the estimated number of persons exiting from that library in the specified time period. Rough estimates of these numbers were derived from average daily usage based on November, 1983, turnstile counts where available and on librarians' estimates where not, and on an assumption that library exit volume was twice as high between 9.30 a.m. and $5.30 \mathrm{p} . \mathrm{m}$. as at other times. The libraries were stratified into four types, and within each stratum controlled selection was employed to give a proportionate distribution of the sample across libraries, days of the week, and time intervals.

For each selected PSU, a systematic sample of persons exiting the library was selected for the survey, with the sampling interval being determined to yield an overall epsem sample of visits. Fieldworkers were provided with a record sheet of integers from 1 up to 430 , with the selected numbers marked on them. All they then needed to do was check off a number for each person exiting the library, and select the persons associated with the sample numbers. An advantage of this scheme is that fractional sampling intervals are readily handled. Where the exit volume for a sampled PSU was expected to be low, one fieldworker was assigned to perform both the counting and the contacting of sampled persons. Where the exit volume was high, two fieldworkers were assigned, one to count and one to contact sampled persons. There was also a need for more than one fieldworker for libraries with more than one exit.

## 32 A survey of museum visits

A face-to-face interview survey of visitors leaving the National Air and Space Museum in Washington, D.C. was conducted from mid-July until December, 1988 (Doering and Black, 1989). The interview, which took about four to six minutes to complete, collected data on the sampled person's socio-demographic background, place of residence, activities on the visit, exhibits of special interest, reason for visit, the size and type of group if part of a group visit, and mode of transport used. Children under 12 years old and persons working at the museum were excluded from the survey. Data were collected from 5,574 respondents, with a response rate of $86 \%$.

Each day in the survey period was divided into two half-days. Interviewing was conducted on one half-day every other day, alternating between mornings and afternoons. During the summer season, three public exits from the museum were in operation, while later in the year only two of them were open. During the selected balf-days, survey data collection was rotated on an hourly basis between the exits that were open.

The fieldwork team for an exit at a sampled hour comprised one, or two, counters and two interviewers. The lead counter used a mechanical counter and a stop watch to keep track of the number of persons exiting, and to maintain a record that gave the numbers of persons exiting in each 10 -minute interval in the hour. The lead counter also identified the persons to be interviewed. The selection of sample persons was made in order to keep the interviewers fully occupied. The lead counter noted when an interviewer had completed an interview and was ready to begin another one, and then chose the fifth person exating after that time as the next sampled person. The 10 -minute flow counts were used in the analysis to develop weights to compensate for the variation in the chance of selection associated with the variable flow of persons across time.

The distinction between the "visit" and the "visitor" is particularly salient for this survey. Persons could, of course, visit the museum on several days throughout the survey period, and also could visit the museum several times on a given day. This latter possibility is particularly likely with the National Air and Space Museum because entry to the Museum is free, and hence there is no incentive to enter only once. Given this situation, it may be appropriate to define multiple entries on one day as a single visit for some analytic purposes. For some purposes, this definition could be applied by restricting the analysis to those exiting for the first time on the sampled day.

## 33 Exit polls

A number of major news organizations conduct polls of voters on election days in the United States (Levy, 1983; Mitofsky, 1991). Voters are sampled as they leave polling places. Those selected are asked to complete a short and simple self-completion questionnaire, and to deposit the completed questionnaire in a ballot box. A typical questionnaire contains around 25 questions asking how the respondent voted, what the respondent's position is on key issues, what opinions the respondent has on various topics, and what are the respondent's demographic characteristics. Refusal rates for the CBS exit polls have averaged $25 \%$ for recent elections (Mitofsky and Waksberg, 1989).

The sampling of voters for election polls usually employs a straightforward two-stage sample design. At the first stage a stratified PPES sample of voting precincts is drawn, where the size measure is the number of voters in the precinct. At the second stage a systematic sample of voters leaving the polling place is selected, with a sampling interval chosen to produce an approximately epsem sample of voters within states. Usually only one interviewer is assigned to each selected precinct. The fieldwork is straightforward when a polling place has a single exit, and the interviewer is permitted to get close to it. When there are two (or more) exits, interviewers alternate between the exits, covering each one for set periods of time. When this applies, the sampling interval has to be modified accordingly. In some states interviewers are not allowed to approach within a certain distance of a polling place, and this can create problems if it results in voters departing in different directions before the interviewer can contact them.

### 3.4 Ambulatory medical care survey

The U.S. National Ambulatory Medical Care Survey (NAMCS) employs a flow survey design to collect data on visits to physicians' offices for physicians in office practice who direct patient care (Bryant and Shimizu, 1988). The NAMCS has been conducted a number of times since it was introduced in 1973. For each survey, data collection has been spread throughout the survey's calendar year in order to provide annual estimates of visit characteristics. Individual sampled physicians have, however, been asked to provide information for a sample of their visits occurring in only one week. The annual coverage is achieved by asking different sampled physicians to report on different weeks of the year.

The sample for the NAMCS is based on a complex three-stage design, which has varied over time. A broad overview of the design will serve for present purposes; for more details, the reader is referred to Bryant and Shimizu (1988). The first stage of the NAMCS sample design is the selection of a stratified PPES sample of area

PSUs, selected with probability proportional to population size. At the second stage, physicians are sampled from lists within the selected PSUs with different sampling intervals from PSU to PSU to take account of the unequal selection probabilities for the PSUs (in the more recent surveys, different specialty classes are sampled at different rates). Sampled physicians are then assigned at random in a balanced way to one of the 52 reporting weeks of the year. Each physician is asked to record information for a systematic sample of his or her patient visits occurring during the sampled week, with the sampling interval being chosen to yield about 30 sampled visits in the week. A sampling interval of $1,2,3$ or 5 is chosen for a particular physician on the basis of the number of office visits the physician expects during the week, and the number of days he or she expects to see patients. The fieldwork procedures consist of keeping a log of patient arrivals for sampling purposes, and then completing a short 16 -item record for each sampled visit.

The NAMCS is a survey of patient visits not patients. As such, it provides useful information about the nature of physicians' work on a visit basis - the frequency of use of diagnostic tests, the therapies provided, and the demographic characteristics of the patients seen. It does not, however, provide estimates on a patient basis, such as treatments and outcomes for patients' episodes of illness.

## 35 Surveys of international passengers

A number of countries conduct surveys of their international travellers, both those entering and those leaving the country by land, sea or air. This subsection will briefly describe the sample designs for a survey of international air passengers conducted by the United States, for surveys of international air and land travellers conducted by Canada, and for a survey of international air and sea passengers conducted by the U.K..

The United States Travel and Tourism Administration conducts an In-flight Survey of International Air Travelers to survey both foreign travellers to the U.S. and U.S. residents traveiling abroad (see, for instance, United States Travel and Tourist Administration, 1989). The survey is conducted through the voluntary cooperation of some thirty airlines. A stratified sample of scheduled flights is selected for the third week of each month and all passengers on those flights are included in the sample. Participating airlines are provided with a survey kit of instructions and questionnaires in appropriate languages for each sampled flight. The airline cabin personnel distribute the self-completion questionnaires in boarding areas or in flight to all adult passengers and collect them prior to debarkation. Nonresponse is a serious problem with these surveys. For the 1988 survey of visitors to the United States, one half of the flight kits issued resulted in no returned questionnaires. For flights for which questionnaires were returned, the estimated response rate for non-U.S. residents was $44 \%$ and for U.S. residents it was only $20 \%$.

The International Travel Section of Statistics Canada conducts international travel surveys at both airports and landports in Canada. The surveys are undertaken in cooperation with Canada Customs, with customs officers being responsible for distributing the self-completion mail-back questionnaires. The account here is based on the report by the International Travel Section, Statistics Canada (1979). It reflects the survey designs that applied prior to some changes that have recently been made. The sample designs for the landports and airports have been similar, and therefore only the design for the landports will be outlined here.

At one time the sampling scheme at landports for returning Canadian residents who had spent at least one night abroad was to distribute survey questionnaires to every travel party on every fourth day throughout the year, the days being chosen by systematic sampling. This scheme proved to be unworkable because the customs officers too often failed to apply it correctly. It was therefore replaced by a stint scheme in which a landport was assigned two periods, or stints, for each quarter of the year during which the questionnaires were to be distributed. The stints were expected to last from 6 to 10 days, with successive stints starting about $61 / 2$ weeks apart (Gough and Ghangurde, 1977). The number of questionnaires sent to a landport for a particular stint was determined from the expected traffic at that port. The customs officers were then instructed to start the distribution of the questionnaires on a given day, and to continue to distribute them until none were left. This sample design is geared to operational limitations resulting from the use of customs officers, for whom the survey is of only secondary concern, as survey fieldworkers. The design has some major drawbacks, but perhaps a more serious concern is a response rate of $20 \%$ or less.

The U.S. and Canadian surveys of international travellers both rely on cooperation from other agencies in conducting the fieldwork. This cooperation has notable benefits in costs, but a price is paid in terms of a lack of ability to apply rigorous controls to the fieldwork procedures. The U.K. surveys of air and sea travellers employ more costly face-to-face interviewing procedures.

The 1984 U.K. International Passenger Survey included the three Heathrow terminals, Gatwick and Manchester airports as strata (Griffiths and Elliot, 1987). Within each airport, days were divided into mornings and afternoons, and these periods constituted the PSUs. A stratified sample of PSUs was selected, and a systematic sample of passengers was chosen in selected PSUs. A sample of PSUs for other airports was also included. Two alternative data collection procedures were used at seaports. At some seaports, interviewers sampled and interviewed passengers at the quayside. At others, the interviewers travelled on the ship, interviewing passengers during the voyage. In the former case, they worked shifis that covered several sailings, and the shift became the PSU. In the latter case, the crossings were the PSUs.

### 3.6 Surveys at shopping centers

Surveys conducted at shopping centers are of two types. One type aims to describe the shoppers' socio-economic characteristics, their areas of residence, and their shopping activities in the center. The other type uses the shopping center as a convenient location to obtain samples of people from the general population of the area.

An example of a survey of the first type is a study that was conducted to examine the impact of the opening of a hypermarket on the outskirts of the city of Southampton, England (Wood, 1978). Surveys of shoppers were conducted in four neighbouring shopping centers both before and after the hypermarket opened (and also al the hypermarket). At each center, the first step in the survey process was the enumeration of all the retail outlets and their hours of opening. The second step was a counting of departures of groups of shoppers from sampled shops at sampled hours, with counting being conducted for 15 minutes within the hour. The counting operation was carried out over a period of one month. Based on the counts obtained, interviews were allocated between shop types and days of the week, and to specific shops and hours. Interviewers were then instructed to interview the given number of people leaving the shop, interviewing the next person to leave after they had completed the previous interview. The sample is one of shop visits, and shoppers could visit several shops on a particular trip to the shopping center. Respondents were asked about previous visits to shops in the center on this particular trip, and also about the number of extra shops they planned to visit. These data were used to develop weights for analysis of trips.

The second type of shopping center survey uses the selected persons at shopping centers as a convenience sample of the general population. Mall intercept surveys of this type are widely used in market research (Bush and Hair, 1985; Gates and Solomon, 1982). The procedures are often haphazard, and the samples are potentially biased. The issues involved are discussed by Sudman (1980), who discusses procedures for sampling shopping centers, locations at selected centers, and time periods to improve the sample designs, and by Blair (1983), Dupont (1987), and Murry et al. (1989).

### 3.7 Road traffic surveys

Surveys of occupants of passenger vehicles are conducted frequently to study seat belt usage and drivers' blood alcohol concentrations. A full discussion of the complex design issues involved in such surveys is outside the present scope; instead only a few general observations will be made.

The method of data collection to be employed exerts a strong influence on the sampling procedures for a road traffic survey. Seat belt usage is generally studied by observational methods, whereas the measurement of blood alcohol concentrations usually involves breathtesting. Shoulder belt usage of front-seat occupants can be observed in moving traffic, but lap belt usage and the seat belt usage of other occupants can be observed only when the vehicle has stopped briefly, for instance at traffic lights. Lack of street lights can preclude observation of seat belt usage at night at some sites. Breathtesting requires the vehicle to be stopped, and this can only be done safely in locations where the stopped vehicle does not hinder the other traffic. Unlike observational surveys, interview surveys that stop vehicles face a significant nonresponse problem.

An ingenious method of studying seat belt usage on interstate highways is described by Wells et al. (1990). For this study, an observer sat behind the driver in a passenger van that travelled at a slower speed than the prevailing traffic in the right hand lane of the highway. From that vantage point, the observer noted the shoulder belt usage of front-seat occupants of cars, light trucks, and vans that passed the observer's van in the adjacent lane.

A more usual approach to studying seat belt usage is to take observations at road intersections and freeway exits controlled by traffic lights, and sometimes at shopping centers and parking lots (Ziegler, 1983; Bowman and Rounds, 1989). O'Day and Wolfe (1984) describe an observational survey of seat belt use in Michigan applying this approach. They sampled a number of areal units, sampled a number of intersections with traffic signals within these areas, sampled days for observations to be taken at these intersections, and sampled five hours for observation on each selected day (based on an alternating one hour working, one hour free, scheme through the day). Observations of seat belt usage were taken at the selected intersections at the specified times for vehicles that stopped at the traffic lights. When more than one vehicle was stopped, observation began with the second vehicle, because of the bias associated with the first vehicle to stop at a light. In order to obtain more detailed information on the usage of child-restraints, observations were also made on vehicles entering shopping centers and rest areas.

The usual approach to analyzing observational data on seat belt usage is to calculate the proportion wearing seat belts among those observed. Brick and Lago (1988) propose an alternative measure, the proportion of estimated time front-seat occupants are belted in eligible vehicles to the total time in eligible vehicles. For their survey a probability sample of all roadway intersections, whether they had traffic signals or not, was selected. To avoid selection bias, observers were told the site they were to use for observation and the direction of the traffic to be observed in the specified 40 -minute interval of observation. The time occupants were on the road was estimated as the length of the road segment leading to the intersection divided by the estimated average speed of the traffic on that segment. This estimated time was used as a weighting factor in the analysis.

The sampling considerations for roadside breathtesting surveys are broadly similar to those for seat belt usage surveys, except that the locations for data collection need to be places where vehicles can be stopped safely. In the 1986 U.S. National Roadside Breathtesting Survey, local police officers cooperated in the survey by flagging down selected drivers and directing them to the survey interviewers (Wolfe, 1986). The interviews lasted about $5-6$ minutes. When an interviewer finished an interview and the respondent had taken the breath test, the interviewer would signal to the police officer to stop the next passing vehicle.

## 4. CONCLUDING REMARKS

As the examples in the previous section illustrate, fieldwork considerations and the economics of data collection play major roles in the choice of sample design for surveys of persons in transit. The length of the time interval used in defining the PSUs may, for instance, be dictated by the length of a suitable workshift for the fieldworkers, and this may result in PSUs with substantial internal variation in the rate of flow. For example, in a survey of passengers arriving at a railway station, a morning interviewer workshift may include a peak flow of early morning commuters and a low rate of flow later on. If it were not for the need to make the PSU time interval conform to the fieldworkers' workshift, it would be preferable to avoid such variation in flow within PSUs since it leads to problems in how to subsample in the selected PSUs.

When the flow of persons within a PSU is uneven, the use of systematic sampling, or any epsem sampling scheme, for selecting persons creates a variable workload over time. If this variability in workload is substantial, there are difficulties in deciding how to staff the PSU for the survey fieldwork, particularly for a face-to-face interview survey. The assignment of sufficient staff to cope with peak flows is uneconomic since interviewers will then often be inactive at off-peak times. Sometimes staffing for somewhat below peak flow may be preferable. This will introduce some nonresponse at times of peak flow because no interviewer is available to conduct an interview with some sampled persons, but it will more fully use the interviewers' time.

The most effective use of the interviewers' time is to assign them to interview the first person to arrive (or leave) after they have completed their current interview. Schemes of this type suffer the disadvantage of not producing
probability samples, and hence there is a risk of bias in the survey estimates. Where cost effective probability sampling designs can be devised, they are to be preferred. However, the choice of a sampling scheme in which the first (or second, or third) person is selected after an interviewer becomes free is understandably attractive for face-to-face interview surveys when the flow is very variable and unpredictable.

## ACKNOWLEDGEMENT

I would like to record my thanks to many researchers who generously provided me with information about the flow surveys with which they have been associated.

## REFERENCES

Blair, E. (1983). Sampling issues in trade area maps drawn from shopper surveys, Joumal of Marketing, 47, 98 106.

Bowman, B.L., and Rounds, D.A. (1989). Restraint System Usage in the Traffic Population, 1988 Annual Report, Washington, D.C. U.S. Department of Transportation.

Brick, M., and Lago, J. (1988). The design and implementation of an observational safety belt use survey, Journal of Safety Research, 19, 87-98.

Bryant, E., and Shimizu, I. (1988). Sample design, sampling variance, and estimation procedures for the National Ambulatory Medical Care Survey, Vital and Health Statistics, Series 2, 108, Washington D.C. U.S. Government Printing Office.

Bush, A.J., and Hair, J.F. (1985). An assessment of the mall intercept as a data collection method, Joumal of Marketing Research, 22, 158-67.

Cochran, W.G. (1977). Sampling Techniques (3rd ed.), New York: John Wiley \& Sons.
Doering, Z.D., and Black, KJ. (1989). Visits to the National Air and Space Museum (NASM): Demographic Characteristics, Working Paper 89-1, Institutional Studies, Smithsonian Institution.

Dupont, T.D. (1987). Do frequent mall shoppers distort mall-intercept survey results?, Joumal of Adverising Research, Aug.-Sept., 45-51.

Gates, R., and Solomon, P.J. (1982). Research using the mall intercept: state of the art, Joumal of Adverising Research, 22, 4, 43-49.

Goodman, R., and Kish, L. (1950). Controlled selection - a technique in probability sampling, Journal of the American Statistical Association, 45, 350-372.

Gough, J.H., and Ghangurde, P.D. (1977). Survey of Canadian residents returning by land, Survey Methodology, 3, 215-231.

Griffiths, D., and Elliot, D. (1987). Sampling errors on the International Passenger Survey, unpublished paper, London: Social Survey Division, U.K. Office of Population Censuses and Surveys.

Heady, P. (1985). Note on some sampling methods for visitor surveys, Survey Methodology Bulletin (U.K. Office of Population Censuses and Surveys), 17, 10-17.

Heeringa, S.G. (1985). The University of Michigan 1984 Library Cost Study: Final Report, Ann Arbor, Michigan: Institute for Social Research, University of Michigan.

Hess, I., Riedel, D.C., and Fitzpatrick, T.B. (1975). Probability Sampling of Hospitals and Patients, Ann Arbor, Michigan: Health Administration Press.

International Travel Section, Statistics Canada (1979). Data Collection and Dissemination Methods for Intemational Travel Statistics in Canada, Ottawa, Statistics Canada.

Jessen, R.J. (1978). Statistical Survey Techniques, New York: John Wiley \& Sons.
Levy, M.R. (1983). The methodology and performance of election day polls, Public Opinion Quarterly, 47, 54-67.
Mitofsky, W.J. (1991). A short history of exit polls. In Polling in Presidential Election Coverage, P. Lavrakas and J. Holley (eds.), Newbury Park, California: Sage.

Mitofsky, W.J., and Waksberg, J. (1989). CBS models for election night estimation. Paper presented at American Statistical Association San Diego Winter Conference.

Murry, J.P., Lastovicka, J.L., and Bhalla, G. (1989). Demographic and lifestyle selection error in mall-intercept data, Journal of Advertising Research, Feb.-March, 46-52.

O'Day, J., and Wolfe, A.C. (1984). Seat Belt Observations in Michigan - August/September 1983, Ann Arbor, Michigan: University of Michigan Transportation Research Institute.

Sudman, S. (1980). Improving the quality of shopping center sampling, Joumal of Marketing Research, 17, 423-31.
United States Travel and Tourism Administration (1989). In-flight Survey: Overseas and Mexican Visitors to the United States. Survey Period: January-December 1988, Washington D.C. United States Travel and Tourism Administration.

Wells, J.K., Williams, A.F., and Lund, A.K. (1990). Seat belt use on interstate highways, American Joumal of Public Health, 80, 741-742.

Wolfe, A.C. (1986). 1986 U.S. National Roadside Breathtesting Sunvey: Pnocedures and Results, Ann Arbor, Michigan: Mid-America Research Institute.

Wood, D. (1978). The Eastleigh Carrefour: A hypermarket and its effects, London: U.K. Department of the Environment.

Yates, F. (1981). Sampling Methods for Censuses and Surveys (4th ed.), London: Charles Griffin.
Ziegler, P.N. (1983). Guidelines for Conducting a Survey of the Use of Safety Belts and Child Safety Seats, Washington, D.C.: U.S. Department of Transportation.

# IMPROVING THE QUALITY OF THE US. CENSUS OF AGRICULTURE LIST 

C.Z.F. Clark ${ }^{1}$


#### Abstract

The U.S. Census of Agriculture mail list is compiled by linking agricultural, statistical, administrative, and commodity lists of establishments. Background information on the list compilation process is presented while focusing on five quality aspects-content of list, accuracy of the address information, uniqueness of list operations, universe coverage of the list, and cost-efficiency of the list compilation. List procedures that are relevant to these quality attributes are discussed. Results of formal evaluations of record sources, of census coverage, of the discriminant model's impact on content and coverage, of undeliverable report forms, and of duplicate list records are presented. Future plans and initiatives for improving the quality of the list are discussed in relation to their potential effect.


KEY WORDS: List frame; Mail data collection; Agriculture.

## 1. DEVELOPMENT OF THE CENSUS OF AGRICULTURE MAIL LIST

The objective of the census of agriculture program is to enumerate all places from which $\$ 1,000$ or more of agricultural products were sold or normally would have been sold during the census year. Since 1969 the United States Census of Agriculture has used a mail-out/mail-back data collection procedure instead of a personal enumeration of farm operations. The census mail list then provides the encompassing frame of potential farm operations. Extensive efforts are made to get response from the list addresses by mail. Telephone data collection is used only as a last resort after all mail efforts fail and only for potentially large or unique operations or in counties with less than 75 percent response.

The census list frame is comprised of individuals, businesses, and organizations that have been identified as having some association with agriculture. A new list is compiled prior to each periodic census. The 1987 list was assembled from the 1982 census records, the administrative records of the Internal Revenue Service and the Social Security Administration, and the statistical records of the National Agricultural Statistical Service in the U.S. Department of Agriculture. In addition, lists were obtained for large or specialized operations (e.g., nurseries and greenhouses, specialty crop farms, poultry farms, fish farms, livestock farms, cattle feedlot operations, grazing permittees) from State and Federal government agencies, trade associations, and similar organizations. Lists of companies having one or more establishments (or locations) producing agricultural products were obtained from the 1982 census and updated using the information from the Standard Statistical Establishment List maintained by the Census Bureau.

The periodic mail list development process consists of acquiring various source record lists, applying a common format to all list addresses, linking list records, removing duplicate records, applying a screening process to list records, identifying records to receive each type of census form, and preparing mail labels for each record. For the past three censuses, the process has included two similar phases of record linkage and the screening procedure. In the 1978 and 1982 censuses, the Farm and Ranch Identification Survey was used between the two phases to screen potential nonfarm or duplicate addresses from the preliminary list. In the 1987 census, a statistical discriminant model was used to remove less probable farm records from the list and to identify less likely farm operations to receive a much shorter census report form. The model was tested for application to

[^31]list development between the two linkage phases. Using either list screening procedure, the two phase linkage process permitted incorporating more current source records than would have been available using a single linkage.

In 1987, general parameters were set for the size and composition of the census mail list to reduce respondent burden and to control costs. The census mail list was limited to a maximum of 4.2 million addresses, of which no more than 3.2 million addresses were to receive either the four-page regular or the six-page sample report form. Up to 1.2 million addresses would be permitted to receive the much shorter two-page report form. These limits were in sharp contrast to the 1982 census where 3.0 million addressees received the short farm and ranch identification survey, 3.65 million addresses received the census report form (including 1.7 million who received the survey) for a total of 5.35 million census contacts. The final 1987 mail list of 4.1 million addresses met the data collection requirements cited above. Of the $4,098,693$ total addresses $-3,192,287$ addressees received either the four-page regular form or the six-page sample form and 906,406 received the two-page short form.

## 2. CONTENT OF LIST

## 2.I Sources of Mail List Records

The census of agriculture mail list forms the basis for the census enumeration. Thus, the content and coverage of the list are important to the potential completeness of the enumeration. Concerted efforts are made to include in the list compilation all important sources of agricultural record information. In the United States this includes tax records of filers with farm income, employment records of farm employers, statistical records of federal agencies collecting agricultural data, administrative records of federal and state agencies conducting farm and related programs, and private records of agricultural trade associations and other organizations. The major record sources used for the census mail list have been basically the same for the past three censuses. However, some notable changes in list content were made for the 1987 census.

In 1987 the Census Bureau was able to use the agricultural farm operation list of the National Agricultural Statistical Service (NASS) for all 50 states as contrasted with the 31 states available for use in the 1982 census list compilation. The NASS list is maintained on an ongoing basis. New name and address information is incorporated into the list upon receipt. Agricultural and commodity information contained on the individual records is updated from current survey and administrative data. However, the NASS does not have access to tax records for farm operators. Data values on the list are classified and categorized once a year. The NASS list was developed to provide good coverage of large and commodity specific operations for use in making dual frame estimates at the state level. Studies conducted by the NASS using its area frame sample estimate that the list includes approximately 55 percent of farm operations in the U.S. meeting the census definition.

The mail list compilation did not use directly the list of the Agricultural Stabilization and Conservation Service (ASCS) as had been done in the 1978 and 1982 censuses. This decision was made for several reasons. Important ASCS record information was thought to be included in the current NASS list since NASS incorporated ASCS record information directly in its list compilation process. The ASCS list was very large--5.0 million records. Name and address information had not been as reliable as other source records in past censuses. Thus, in the linkage process, many ASCS records did not match other source records and required screening ( $706,000-25$ percent - of the 3.0 million records in the 1982 Farm and Ranch Identification Survey were ASCS records only). Many of those addressees were ultimately found not to have a farm operation. Of the 1982 census farm respondents, only 1.4 percent came uniquely from the ASCS list.

The 1987 mail list incorporated the records of the Massachusetts Department of Agriculture at the request of the State of Massachusetts. The Massachuselts Department of Agriculture thought that the 1982 Census of Agriculture data for Massachusetts did not provide very complete state coverage. The Department felt that incorporating its list of 10,600 farm records into the census list would improve census coverage. Codes were used in mail list development and data processing that will permit us to assess any improvement in coverage from the use of this list. If this evaluation proves to be favorable, we will investigate other such lists in states with relatively large census undercoverage.

### 2.2 Evaluation of Mail List Record Sources

Following the 1982 Census of Agriculture, an extensive evaluation of record sources used in the final mail list was conducted based on a weighted sample of census records. The record sources were evaluated with respect to: the response rate for names on the list, the proportion of names on the list representing farms, the proportion of the census farm universe covered by the list, the number and proportion of farms contributed uniquely by the list, and characteristics of the uniquely contributed farms by list. This evaluation showed that 11.5 percent of all enumerated census farms were unique to the IRS list as contrasted with 2.9 percent unique from the previous census mail list, and 2.0 percent from the NASS list. It also showed that the proportions of farms on the IRS list and the special agricultural production lists that were unique was much higher than for the other source lists. Table 1 summarizes this information for the 1982 mail list (Gaulden, 1985). A similar evaluation is being conducted for the 1987 census mail list, but results are not yet available.

Table 1: 1982 Record Source Evaluation

| Source <br> List | Records in <br> Final List | Response <br> Rate | Farm <br> Rate | Coverage <br> Rate | Unique Farm <br> Rate | Unique <br> Coverage Rate |
| :--- | ---: | :---: | :---: | :---: | :---: | :---: |
| Any Source | $3,629,425$ | 87.4 | 61.9 | N/A | N/A | N/A |
| IRS | $2.688,850$ | 90.5 | 69.1 | 85.6 | 9.3 | 11.5 |
| 78 Census | $1,913,800$ | 95.2 | 75.9 | 70.4 | 3.2 | 2.9 |
| Farms |  |  |  |  |  | 2.6 |
| NASS Farms | $1,638,219$ | 88.8 | 69.7 | 51.6 | 1.5 | 2.0 |
| ASCS | $2.204,275$ | 87.6 | 71.1 | 69.9 | 9.1 | 11.4 |
| Special Ag Lists | 139.075 | 91.5 | 59.4 | 3.8 | 0.6 |  |

Response Rate - Percentage of all records on final list that responded to the census.
Farm Rate - Percentage of respondents that represented farms.
Coverage Rate - Percentage of all farms that are on this list.
Unique Farm Rate - Percentage of respondents that represented unique farms.
Unique Coverage Rate - Percentage of all farms unique to this list.

In addition, a small study was jointly conducted by the Census Bureau and the NASS in 1985 to measure and compare the percent coverage of all farms provided by the census and NASS lists, separately and combined. The personal enumeration in the 1984 June Enumerative Survey (JES) area frame sample in four states (Washington, New York, Iowa, and Georgia) was considered to represent complete coverage in each state for the sampled segments. The NASS designated each area frame sample case as either a match, a probable match, or a nonmatch in relation to the NASS list. The Census Bureau specified whether each JES area sample case matched a 1982 census farm, nonfarm, or nonrespondent.

The Census Bureau estimated that the coverage of all farms across the four states was 75 percent using the census farm list only, 53 percent using the NASS list only, and 80 percent using both census and NASS lists. The coverage of farms with sales greater than $\$ 2,500$ was better with 80 percent using the census list, 61 percent using the NASS list, and 86 percent using both lists. The state estimates varied considerably where farm coverage by the census farm list and the NASS list ranged from 71 percent in Washington to 88 percent in Iowa. These estimates are somewhat lower than we would expect for the census, due to the two year lag in the currency of the census list. The NASS estimates differed slightly from the Census Bureau estimates due to the differences in match rules. On the basis of the evaluations of the 1982 mail list and the match study with the census and NASS lists, the IRS lists were considered necessary sources for good coverage for the 1987 census (Gaulden, 1986, and Davie, 1986).

## 23 Screening of Mail List Records

Although all records used to compile the census of agriculture mail list have some association with agriculture, this does not ensure that each record represents a current unique operation meeting the census definition of a farm. The complexity of agricultural operating arrangements makes it difficult to determine uniqueness of operations. The identification and agricultural data available for each list record does not generally provide adequate information to determine whether the record represents a farm operation. In 1978 and 1982, the Farm and Ranch Identification Survey was mailed to addresses with questionable farm status or that were potentially duplicate addresses, thus reducing the number of final census mail list records. The survey instrument was designed to determine the agricultural production of the addressee and to identify multiple operations of the addressee. Even using this survey, the number of records on the final census mail list has been nearly double the number of responding farm operations in the United States over the past few censuses.

The discriminant model applied to the mail list in 1987 served to accomplish the first objective of removing records with questionable farm status, but not that of identifying duplicate records. An evaluation of the classification tree model (Schmehl, 1990), demonstrated that the model was effective in separating the records into groups according to the proportion of expected census farms in each group. Record characteristics such as the source of the mail list record, number of source lists on which the record appeared, expected value of agricultural sales, geographic location, and 1982 farm status were used to separate 1982 mail list records into model groups. The proportion of 1982 census farm records in each model group was calculated to provide an estimate of the probability that a 1987 mail list addressee with that group's characteristics operated a farm (Clark, 1989, and Owens, 1989). Refinements to the model for 1992 will attempt to improve the accuracy and predictive power of the model by including record variables not previously available and changing terminal model group size constraints. This may provide sufficient information to eliminate more records from the final mail list.

### 2.4 Size of the Mail List

The changes in the source records for the 1987 census were expected to reduce the number of distinct records coming out of the linkage and matching process and to enable us to compile a list of potential farm operations whose size approached the 4.2 million record limit. Prior to implementation, it was not known how effective the statistical discriminant model would be in separating list records into groups by probability of farm status. As no prescreening survey was authorized, we did not have the luxury of compiling a large list and screening potential nonfarm records afterwards. The best strategy for list compilation with these constraints appeared to be to limit the number of less reliable input records.

The total records used in the 1987 census list compilation was 13.5 million as contrasted with 18.9 million in 1982 . After linking records identifiable as duplicates by name, address, Social Security Number (SSN), or Employer Identification Number (EIN) 6.0 million records remained at the end of the second linkage phase. Of the linked records, approximately 1.6 million came exclusively from nonfarm sources and were deleted from the mail list. (Nonfarm records from both NASS and the 1982 census were included in the initial linkage process to separate less likely farm operations from current nonfarms by whether or not the previous nonfarm record linked to a current agricultural list record.) A total of 180,000 records that were previous census nonrespondents with expected sales of less than $\$ 2,500$ were also deleted, leaving 4,275,000 records. An additional 175,000 records were deleted based on group probabilities developed by the discriminant model, leaving 4.1 million records on the final census mail list. From that list, the census sample was selected and then the remaining 900,000 records least likely to represent farms were designated to receive the census short form.

### 2.5 Results of Census Data Collectlon

Data were obtained as a result of the census data collection for $3,404,036$ ( 86.2 percent) of the deliverable addresses on the census mail list. The 1982 and 1978 censuses had comparable response rates of 85.1 and 87.3 percent, respectively. The respondents included $1,826,702$ farm operator addresses, $1,533,776$ nonfarm operator addresses, and 43,558 nonclassified report forms. There were a total of 542,248 final nonrespondents and 148,252 final post master return addresses. A nonresponse weighting procedure (based on individual state surveys of nonrespondents) was used to account for the nonrespondent farm operations in the final census data. Using this
procedure, we estimated that 263,057 of the nonrespondent addressees had farm operations for a total of 2,087,759 census farm operations from the total census mailout of $4,095,061$.

## 3. ACCURACY OF RECORD INFORMATION

### 3.1 Procedures Used to Increase Accuracy of Record Information

Specific procedures were used in mail list development to enhance the accuracy of the final record information (Sandusky, 1984, and Gauiden, 1990). Each input record was placed into a standard format for name (establishing name control), street or rural address, place name, and processing code fields. Such standardization was necessary to implement all computer linkage and duplicate identification processes. Records were linked by computer first on identification number (Social Security and Employment Identification Number). All remaining unmatched and potential duplicate records were coded by geographic area and processed through a name part identification and name recoding prior to conducting a computer match by name and address within ZIP code or area. An additional linkage of historical duplicate agricultural records was conducted. Duplicate records were deleted, retaining the record whose source has the highest quality address and information. All potential duplicate records (identified by pair numbers) were reviewed clerically. The final address and identification information for list records was based on that available from all linked records.

Name control was essential in determining positive or potential duplicate status during the initial linkage on identification number. A special "skip list" dictionary was developed to use in establishing an appropriate name control for agricultural records. This dictionary contained over 1,000 words and abbreviations (such as Farm, Dairy, Bros) which could conceivably appear in the name field but were not likely to be the surname. Name control was established by reading the primary name field from right to left until a non-numeric word with three or more characters was encountered which did not appear in the skip list dictionary and by specifying the first four characters of this word as the name control. A surname locator (indicator code) was placed in the record to identify the field position of the word used to derive the name control. A small number of records (one half of one percent) had no name control (e.g., A \& B FARMS INC).

Processing code fields were assigned in the initial record format and standardization to link records and to facilitate the use of the most reliable information in the final record. In particular each record was assigned a name and address record priority code. This code was used in the linkage process to determine which source record to retain in the case of duplicates. The priority code was based on the expected currency of the record source address information.

Each record was also assigned a measure of size derived from size indicators present in the source record. This is expected to be an estimate of the total value of agricultural products (TVP) that might be sold in the census year. Each source had a separate field for this size code so that during record linkage, the size code could be retained for all sources on which a name appeared by transferring data from the deleted duplicate record to the retained record. This allowed the derivation of both a "source combination code" indicating all the sources for the final record, and a "final size code" based on the reliability of size information for each source. Both source and size codes were important variables used in the discriminant model. The final size code was used in census processing to determine the type of report form to mail, sampling rate, and type of follow-up procedure for nonrespondents. Both codes were used to evaluate the census mail list.

The geographic coding system was designed to ensure that all records entering the record linkage system contained standardized and edited geographic codes - agricultural census state and county codes, county alphabetic codes, and ZIP codes. This coding information was critical for mailing and processing as well as name and address linkage. The procedure used geographical information from the Geography Division's Reference File. Validity checks of ZIP code to post office name were performed using full post office name and soundexed names. During the subsequent name and address linkage, ZIP code was used as the blocking factor such that only records within a ZIP code were compared for possible matching.

### 3.2 Evaluation of Accuracy of Record Information

Several measures from the census data collection permit an assessment of the accuracy of the name and address information on each record. Those report forms that the post office can not deliver are returned to the census processing office in Jeffersonville, Indiana as post master returns (PMRs). The PMRs were sorted by deceased addressee, name or address change, and other. All PMRs were remailed once with address changes made, and deceased remailed to "the estate of." A listing was generated for all second time PMRs that had been preidentified as potentially large cases for later telephone follow-up. In 1987 there were a final total of 148,252 PMRs ( 3.6 percent of the census mailout). This compared with 82,792 PMRs in 1982 ( 2.3 percent). The increase in PMRs was likely due to the lack of a farm and ranch identification survey. Through this survey, 446,000 PMRs were removed from the preliminary 1982 list; $1,373,000$ presumably verified addresses in the farm and ranch survey were also in the census although 522,000 of these were survey nonrespondents.

An indication of the accuracy of the individual record information is obtained from results from the Nonresponse Survey. Prior to completion of the census data collection for each state a Nonresponse Survey is conducted. This survey provides an estimate of the percent of farms among the nonrespondents. The survey cases are mailed a questionnaire and later telephoned. In 1987, we were able to contact and classify 63.5 percent of the sample of 27,096 cases; in 1982 we contacted and classified 70.5 percent of the sample of 13,489 cases. The difference in response and classification is attributable to changes in the sample sizes, the data collection methodology from one conducted by analysts in 1982 to a clerical procedure in 1987, and the quality of list addresses.

Additional information on the accuracy of addresses in the Nonresponse Survey sample was obtained through a study of 1,263 nonclassified survey cases in selected stratum in six states. In early 1989, we conducted intensive mail and field follow-up on these cases to obtain further information about the nonrespondents. First we mailed a certified report form to each case in the sample. We obtained response from 488 ( 38.6 percent) of the cases. The Post Office returned 382 forms ( 30.2 percent) as PMRs (deceased, address changes, or other undeliverable mail). We reviewed all cases and completed data collection by telephone where possible. We conducted field interviews for the remaining cases, completing data collection for 880 survey cases ( 70 percent). Of these cases, we discovered through field work that a large number had incorrect addresses. The high percent of the survey cases that were PMRs in the certified mailing indicated to us that we either did not receive corrected address information from the Post Office in our initial census mailings or that we did not incorporate Post Office address changes into our subsequent mailings.

Table 2: 1982 Relationship of Expected to Actual TVP

| Mail Size Code <br> Lower Bound | Percent with <br> Same Code | Percent Plus or <br> Minus One Code |
| ---: | :---: | :---: |
| All Records | 34.9 | 72.7 |
| $\$ 1,000,000$ | 3.1 | 80.0 |
| 500,000 | 37.7 | 82.2 |
| 200,000 | 51.9 | 84.3 |
| 100,000 | 49.1 | 74.0 |
| 80,000 | 22.6 | 68.0 |
| 60,000 | 24.9 | 60.5 |
| 40,000 | 31.9 | 73.0 |
| 20,000 | 39.8 | 74.6 |
| 10,000 | 34.2 | 73.0 |
| 5,000 | 27.9 | 64.0 |
| 2,000 | 32.8 | 77.4 |
| 1,000 | 34.5 | 87.5 |
| Less than $\$ 1,000$ | 42.4 | 70.0 |

Each mail list record contains important information for conducting the census in addition to the name and address for the record. Section 3.1 described the mail list source code and the mail size code and indicated the use of the codes in the census processing. After the census, the mail list record file was matched to the census processing file and the census farm file. Tabulations from this matched file provide important insights for
evaluating the quality of the census mail list. Information from such tabulations on the record source were provided in Section 2.2. Tabulations of mail size by actual total value of agricultural products sold (TVP) in the census year provides an indication of the accuracy of this code. Approximately 35 percent of all list records have a value for TVP in the same size code as their mail size; 73 percent have a value within one size code in either direction.

## 33 Potential Improvements In Accuracy of Address Information

The increase in PMRs in the 1987 census, the increase in nonclassified cases in the Nonresponse Survey, and the high number of bad addresses in the small study of census nonrespondents indicate that there is the potential for improvements in the accuracy of address information. We are presently investigating the use of precensus address verification services offered by the United States Post Office where the Post Office updates address lists provided by the Census Bureau. We would use this service for potential duplicate addresses, in particular. We expect to have an increase in potential duplicates because rural addresses are being changed to street addresses. Our system identifies as potential duplicates records with the same name and different addresses.

## 4. UNIQUENESS OF LIST OPERATIONS

### 4.1 Procedures Affecting Uniqueness of List Addresses and Operations

The list matching procedures used in the census of agriculture are designed to link records that have a high probability of representing the same name and address. Computer and clerical procedures are designed to identify almost exact matches. The rules are intended to provide a high level of census coverage by retaining possible separate operations associated with minor name or address variations. In the agricultural universe, farm or ranch operators often have multiple operations such that an operator may be involved with one or more parinerships as well as an individual operation. To address one aspect of these operating arrangements, the census mail list compilation includes a linkage of historically identified operations with different names. The source record format program also inserts a possible partnership or corporation (PPC) flag to assist in linkage operations. The PPC flag is used to prevent automatic computer deletion of records as duplicates, causing all records linking to a PPC record to be manually reviewed.

### 4.2 Evaluation of Uniqueness of List Addresses and Operations

The coverage evaluation program measures error in report form farm classification and in list duplication as well as farms not on the mail list. Nonfarms classified as farms and duplicate operations contribute to overcounted farms in the census. The total number of 1987 estimated overcounted farms $(135,600)$ was very similar to the 1982 estimated number ( 113,623 ). However, the proportion of the overcounted farms that represented duplicate operations changed from 17 percent in 1982 ( 19,062 farms) to 47 percent in 1987 ( 63,290 farms). This increase in 1987 was primarily attributed to the lack of a precensus screening survey.

The initial format program standardized the name and address information for all records. The address standardizer for the census of agriculture mail hist was especially designed for rural addresses. We are in the process of evaluating the census of agriculture standardizer by comparing it with the standardizer used for the decennial census and one developed for the 1990 Post Enumerative Survey. This evaluation will indicate if the accuracy of our matching process can be improved by using another address standardizer.

We are also evaluating the use of a probability based linkage system for name and address matching. This methodology would permit specifying the degree of certainty desired for matched records within the system. We will compare the probability based matcher used for matching the 1990 Post Enumerative Survey records to the 1990 decennial census with the census of agriculture name and address matching procedures. We are in the process of defining appropriate files for this evaluation. Improvements in the accuracy of the address standardizer and the matcher should increase the proportion of unique operations on the list, reduce the number of duplicate enumerations, and reduce the cost of clerical review.

### 4.3 Plans to Reduce Duplicate Census Enumerations

We are presently reviewing several procedures to reduce the duplicate enumerations in the census of agriculture. We are considering changes in the mail list development, the census data collection, and the census data processing. We will incorporate the results of the evaluation of the computer and clerical list matching rules and the probability based matcher into mail list compilation. We are also investigating a Post Office address verification. Additionally, we are considering a small scale mail or telephone survey of list addresses similar to the farm and ranch identification survey. The Post Office address verification and the survey would additionally increase the accuracy of list addresses.

The primary objective of the evaluation of the list computer and clerical rules is to increase the accuracy of the matching procedure. However, we do not want to increase false matches as that would adversely affect census coverage. It may be preferable to retain more potential duplicates in the census mail list and to use the census data collection and processing to reduce the number of duplicate enumerations. A precensus mail test followed by a telephone reinterview will evaluate three variants in instructions for the respondent to mail back duplicate report forms together. These procedures are designed to encourage respondent identification of separate and duplicate operations.

We are investigating the combination of several procedures to use during data review to identify potentially duplicate enumerations. They include a telephone number match, a universe listing with important data variables, and an alphabetic sort within counties during data review. The universe listing was used in the 1982 census but not in the 1987 census. Telephone numbers were first available in the 1987 keyed data, but telephone number matching during census processing was only used in a few states.

## 5. UNIVERSE COVERAGE OF LIST

### 5.1 Evaluation of Census List Coverage

The objective of any census is to provide a complete enumeration of the target universe. This is an extremely difficult goal for the census of agriculture because of the impossibility of identifying operations in the target universe for inclusion in the census. Data from formal coverage evaluation programs conducted for each census of agriculture indicate that coverage of the census farm universe has always been less than complete even when the census was conducted by personal enumerations as it was prior to 1969. Data from the census coverage evaluation programs estimate that net farm coverage has ranged from 85.0 percent to 96.6 percent. (The net farm coverage is defined as the total number of farms in the census divided by the coverage estimate of farms in the universe. Prior to 1982 the numerator was estimated by the area frame survey; for 1982 and 1987 the numerator was the census farm count. Prior to 1982 the coverage estimate of farms in the universe (denominator) was based on the area frame and classification samples; for 1982 and 1987, the coverage estimate of farms was based on the capture-recapture estimator.) Although, the coverage of farm operations has been approximately 90 percent for these censuses, the census has enumerated approximately 98 percent of the agricultural production.

Table 3: Estimated Census Farm Coverage: 1954-1987

| CENSUS <br> YEAR | PERCENT NET <br> FARM COVERAGE |
| :---: | :---: |
|  |  |
| 1987 | 92.8 |
| 1982 | 90.9 |
| 1978 | 96.6 |
| 1974 | 89.3 |
| 1969 | 85.0 |
| 1964 | 88.7 |
| 1959 | 91.6 |
| 1954 | 91.9 |

The best coverage of the farm universe occurred in the 1978 census. This census was conducted as a dual frame census data collection using an area segment sample in conjunction with the mail enumeration. This methodology substantially improved the state and U.S. level coverage of the census, particularly for farms with sales of less than $\$ 2,500$ where the census enumeration is least complete. In 1978 , the percent of these farms not included in the census was 6.5 percent compared with 28.6 percent in 1982 and 32.3 percent in 1987. Budget constraints have not permitted the use of dual frame methodology in subsequent censuses.

The coverage evaluation program for the 1987 census of agriculture was enhanced to provide estimates of farms not on the census mail list at the state level as well as more reliable estimates of incorrectly classified operations and duplicate operations on the census mail list. The percent of estimated farms on the census mail list was approximately the same in 1987 as in 1982--89.2 percent contrasted with 89.4 percent. The reduction in size of the total mail list and the lack of a screening survey did not adversely affect the coverage of the mail list. The changes in the source lists for the 1987 census, improvements in the quality of the source records, and the effectiveness of the discriminant model contributed to maintaining the previous level of list coverage despite the drastic reduction in total list records included in the census. The estimate of duplicate operations on the list, however, approximately doubled from the previous census. No 1987 census procedures proved as effective in removing list duplication as the precensus farm and ranch identification survey.

## 52. Effect of List Compilation Procedures on Census List Coverage

The philosophy for compiling the census list from the input source records has been to design procedures that enhance the potential coverage of farm operations in the census enumeration. In particular, the precensus screening survey and specific computer and clerical record linkage rules were designed to ensure that the final list was comprehensive within given size and cost constraints. However, the much more rigid size constraints and restrictions on precensus screening in 1987 necessitated that the resulting list have a high proportion of farm operations to ensure good coverage. The classification tree model was developed to accomplish this objective in lieu of precensus screening.

The computer and clerical linkage procedures used to identify duplicates affect list coverage. Less rigid rules for determining duplicate records risk removing potentially different operations from the list. The present computer and clerical rules for identifying duplicates err on the side of retaining the potential duplicate records in the mail list. For example, three records with the same address and different names may represent zero to three different operations; if the linkage rules specify that they represent one operation and there are really three operations, then two operations will potentially be missed. Although these procedures may increase coverage, they also increase the potential for enumeration of duplicate operations in the census. An evaluation of these rules is planned prior to the 1992 census.

An evaluation of the classification tree discriminant model was conducted to determine how effective the model was in defining a mail list with a high proportion of farm operations. The evaluation compared the expected proportion of farm records with the actual proportion of farm records by model group for records retained on the mail list. The evaluation determined that, for most model groups, the actual proportion was slightly less than the expected proportion (Schmehl, 1990). The model for the 1992 census of agriculture will be able to include several additional variables-a reason for nonfarm status on past census and NASS nonfarm records, and two additional codes on IRS records--a material participation and an agricultural activity code.

As part of the evaluation of the model, a sample survey of the records dropped from the census mail list (exclusive of the nonfarm only source records) was conducted. The records selected to be dropped from the mail list using the discriminant model were in groups whose proportion of farms was expected to be 11.7 percent or iess. The survey estimated that 14.6 percent of these dropped records represented farm operations. An upward bias is expected in this estimate as only 46.4 percent of the sample cases were classifiable, and farm addressees are generally more likely to respond to the census than nonfarm addressees. From this result, we estimated that 25,500 of the 175,000 records dropped from the mail list had farm operations.

The coverage evaluation estimated a total of 242,850 farms not on the final census list. Thus about 10 percent of the farms not on the final census list were on the preliminary list and the remaining 90 percent were either on the list of nonfarm addresses and dropped or not on either list. The discriminant model, thus, provided a
reasonably good estimate of the impact of reducing the size of the final mail list on census coverage. For future censuses, this estimate of census undercount can be balanced with the cost and burden implications of a larger census mail list.

## 53 Potential Improvements in List Coverage

In 1978 the census of agriculture published estimates of farms not on the census mail list for each state obtained through the use of a personal enumeration area sample. These state estimates indicated that census list coverage varied considerably by state. However, state estimates were not available from the 1982 census of agriculture or coverage evaluation program. Without such estimates in 1982, it was difficult to identify geographic inadequacies of the list to isolate areas for improvement. An important rationale in 1987 for the expansion of the census coverage evaluation program was to provide state estimates of farms not on the census mail list in order to assess census coverage in individual states. This information was particularly important for the 1987 census mail list since substantial changes were made in input list record sources (NASS list was expanded from 31 to 50 states and the ASCS list was not used).

The state estimates of farms not on the 1987 census mail list varied from a low of 1.4 percent in North Dakota to a high of 26.6 percent in West Virginia. Estimates of percent of total estimated U.S. farms not on the mail list for the census divisions are provided in Table 4. An analysis of these estimates indicates that the census mail list provides lower coverage for the New England and South Atlantic states, and higher coverage for the West North Central and East North Central states. Based on these data and the individual state data (see Table 3, 1987 Census of Agriculture Coverage Evaluation report), the Census Bureau will investigate different procedures to improve the coverage of the list in specified states. If the evaluation of the data from the State of Massachusetts Department of Agriculture indicates that use of this list substantially improved Massachusetts census coverage, then the Census Bureau will pursue acquiring such lists in other states.

Table 4: 1987 Divisional Estimates of Undercoverage

|  | Census <br> Published <br> Farms | Estimated <br> Farms Not on <br> Mail List | Percent <br> Not on <br> Mail List |
| :--- | ---: | ---: | :---: |
| United States | $2,087,759$ | 249,306 | 10.7 |
| New England | 25,158 | 7,767 | 23.6 |
| Middle Atlantic | 98,324 | 17,330 | 15.0 |
| E. North Central | 364,872 | 33,593 | 8.4 |
| W. North Central | 497,110 | 23,423 | 4.5 |
| South Atlantic | 239,687 | 56,565 | 19.1 |
| E. South Central | 249,556 | 27,767 | 10.0 |
| W. South Central | 334,608 | 42,987 | 11.4 |
| Mountain | 124,210 | 17,142 | 12.1 |
| Pacific | 154,234 | 22,732 | 13.3 |

## 6. COST OF LIST COMPILATION

### 6.1 Cost Factors in List Compilation

The actual cost of compiling the mail list for the census of agriculture program is difficult to separate from the overall program cost of $\$ 65$ million in 1987. Much of the cost is associated with salaries for professional staff.

A high level of work occurs for projects associated with the development, implementation, and evaluation of the mail list during threc years of the census cycle. Planning and research activities occur in the remaining two years. Professional work on the mail list requires a minimum of three statisticians during the cycle and three computer programmers during development, implementation, and evaluation.

Other costs of compiling the 1987 census of agriculture mail list included procurement of source records, computer processing and geocoding of those records, salaries for clerical staff reviewing potential duplicates, and report form label preparation. Source records are obtained at minimal cost to the Census Bureau as they are output of other statistical and administrative data file preparations rather than separate data collection costs. For example, in 1987 we paid the NASS $\$ 30,000$ for its 2.4 million records and IRS less than $\$ 125,000$ for its 6.0 million records. The number of records used in the mail list linkage affects the overall cost of computer linkage and staff costs for clerical review. That number was 5.4 million ( 30 percent) less in 1987 than in 1982. Computer processing for the list is conducted on the Census Bureau's mainframe computer with its associated costs and overheads. Our estimates of these costs are not complete due to inaccuracies in charging and changes in charging algorithms. An indication of staff costs for clerical work is provided by the number of potential duplicate sets of records prepared for review- 767,448 sets in 1987 compared with 1,332,000 sets in 1982 for the first phase of linkage.

The discriminant model with its associated computer and professional salary costs was an inexpensive substitute for the much more costly farm and ranch identification surveys conducted in 1978 and 1982. However, the discriminant model application was only designed to remove nonfarm addresses from the list. It did not accomplish the other two objectives of the farm and ranch identification survey of obtaining more current address information and identifying duplicate operations. The costs of such an independent data collection are relatively high. Although the costs are affected to some degree by the size of the survey mailing ( $\mathbf{3 . 0}$ million in 1982), the marginal cost of additional survey cases is small with such a large scale data collection.

### 6.2 Cost Implications of Potential Improvements

Potential improvements in the development of the census of agriculture mail list have been addressed in the preceding sections of the paper. These suggested changes were discussed in relation to quality characteristics of a list--content, accuracy of addresses, uniqueness of records, and universe coverage. The potential changes included procurement of state or other lists of agricultural operations, post office address verification, selective precensus screening of list addresses, different computer or clerical match rules, refinements in the address standardizer and name and address matcher, and enhancements to the statistical discriminant model. Cost implications of these potential improvements can only be addressed in a general way until more specific plans are made.

Generally, the cost of procuring state or other lists of agricultural operations has been minimal for census list development. The costs associated with additional lists would be the additional record processing costs and the cost of any additional programming resources required to standardize the formats of a number of different lists. As our programming resources are scarce, we would be inclined to add additional lists selectively if we were not able to receive these lists in a standardized format to our specifications.

The post office address verification would provide one of the functions of a precensus screening survey-that of verifying and correcting address information. The cost of the post office verification is likely to be substantially less than survey screening. We could then mail or telephone the addressees that were thought to have potentially duplicate operations (e.g., individual name linked to operation name). This two-step approach could provide a less costly alternative to the large scale farm and ranch identification surveys conducted in the past.

The computer record linkage rules are presently designed to avoid computer deletion of potential duplicates without a high degree of certainty for matches. Less stringent matching rules could decrease the number of potential duplicate sets provided for clerical review, thus reducing clerical staff costs. The use of less stringent rules though would need to be balanced with a potential loss in census coverage. The cost of clerical staff review may increase in 1992 independent of any change in linkage rules. This is due to postal changes in rural addresses from rural route and box numbers to street addresses. This would result in more records with the same name but different addresses-thus more potential duplicates.

It is difficult to assess any cost implications of changes in the address standardizer, the name and address matcher, or the discriminant model at this stage. Research is currently progressing on the standardizer and matcher. This research is being conducted in the minicomputer programming environment, although it is unlikely that we will be able to move the entire mail list compilation process from the mainframe computer to the minicomputer in 1992. The more important issue may be the cost of resources to develop the desired procedure for the applicable programming environment. Changes in the model controls to improve the accuracy of estimates of expected group farm proportions would have very minor cost implications. Enhancing the model to incorporate new variables would require professional staff resources and computer costs.

## 7. SUMMARY

Building a high quality mail list for a large data collection is a difficult task. It requires an ongoing program of research, evaluation, and development whether or not the list is maintained and updated periodically or recreated cyclically as is the census of agriculture list. Changes in source records and postal delivery procedures affect the list compilation process, requiring new list techniques. The purpose for which the list is intended is an important factor in determining the quality requirements for the list.

Compiling a list of agricultural operations to conduct a census has a unique set of challenges. Yet a census of farm operations conducted exclusively from a list is unlikely to ever have a near complete list of farm operations meeting the current census definition. Coverage of all farms has been at the 90 percent level over several censuses. However, the list coverage of larger farms has generally been above 95 percent and agricultural and economic activity approximately 98 percent. Even maintaining this level of list coverage requires continual attention to improvements.

Since the census is the only source of detailed county-level data and data on rare agricultural commodity items, it is necessary to have as complete a mail list as can be obtained using records from many diverse sources. From the United States experience we have determined that tax records are needed to achieve a high level of coverage as this source uniquely provided approximately 11.5 percent of all records in 1982. This is due in part to the high turnover in operations from one census to the next (only 71 percent of farms in 1978 were farms in 1982). Thus, we continually have the problem of identifying farms that have gone out of business and finding new operations.

The many different arrangements under which farms and ranches operate will invariably affect duplication in the list. As we discovered in 1987, we need to have controls in our overall census processing system to address the unavoidable duplicate mailings to list addresses. We had not expected such a large effect on duplicate enumerations in the census from eliminating the precensus screening survey. We must continually assess changes in methodology intended to increase quality of one aspect of the list in relation to their effect on other quality attributes.

## ACKNOWLEDGEMENTS

The basic record linkage system for the census of agriculture mail list was developed in 1978. The effort was directed by D. Dean Prochaska with major contributions from Ralph Graham, Jane Sandusky, Tommy W. Gaulden, Billy Stark, Steve Schobel, and Steve Hess. Modifications and improvements were proposed and implemented for the 1982 and 1987 censuses by these same individuals. Dedrick Owens, under the direction of Ruth Ann Killion, developed the classification tree discriminant model used in the 1987 census in lieu of a precensus screening survey. Formal evaluations of various aspects of the mail data collection by staff of the Agriculture Division are referenced in the body of this paper. Charles P. Pautler, Jr., Tommy Gaulden, and Jane D. Sandusky provided information, critique, and helpful suggestions for this paper. Melody Atkinson, Diana Mars, and Antoinette Wooten prepared graphics for the paper. Sharon Lewis and Rosemary Vespoint prepared the manuscript for the proceedings.

## REFERENCES

1987 Census of Agriculture, U.S. Summary and State Data, 1, 71, Appendices A and C.
1987 Census of Agriculture Coverage Evaluation Report, 2, 2.
Clark, C.Z.F. (1989). Mail Data Collection Methodology and Research for the U.S. Census of Agriculture, Proceedings of the 47th Session of the International Statistical Institute.

Clark, C.Z.F. (1985). Mail Enumeration in the United States Census of Agriculture, Proceedings of the 1991 Census Planning Conference, Statistics Canada.

Clark, C.Z.F. (1984). Comparability of Data from the Censuses of Agriculture, 1984 Proceedings of the Section on Survey Research Methods, American Slatistical Association.

Davie, W.C. (1986). Special Study on Farms Undercounted in the 1982 Census of Agriculture Due to Misclassification, U.S. Census Bureau internal report.

Davie, W.C., Marz, D., and Wooten, A. (1986). 1986 Census/NASS List Study Matching Procedure, U.S. Census Bureau internal report.

Davie, W.C., Lorenzen, E., and Prochaska, D.D. (1984). Coverage Evaluation for the 1982 Census of Agriculture, 1984 Proceedings of the Section on Survey Reseanch Methods, American Statistical Association.

Dea, J.Y., Gaulden, T.W., and Prochaska, D.D. (1984). Record Linkage for the 1982 Census of Agriculture Mail List Development Using Multiple Sources, 1984 Proceedings of the Section on Survey Research Methods, American Statistical Association.

Gaulden, T.W. (1990). Development of the 1987 Census of Agriculture Mail List, U.S. Census Bureau internal report.

Gaulden, T.W. (1985). Evaluation of Mail List Sources Using the Mail List Research Sample, U.S. Census Bureau internal report.

Gaulden, T.W. (1986). Census/NASS List Match Study, U.S. Census Bureau internal report.
Lorenzen, E.B. (1986). Special Study of Overcounted Farms in the 1982 Census of Agriculture Coverage Evaluation, U.S. Census Bureau internal report.

Owens, D., Killion, R.A., Ramos, M., and Schmehl, R., (1989). Classification Tree Methodology for Census Mail List Development, 1989 Proceedings of the Section on Survey Research Methods, American Statistical Association.

Ruggles, D.R., Dea, J.Y., Kwok, F.K., and Carman, C.A. (1984). Evaluation of the Effectiveness of Data Collection Procedures for the 1982 Census of Agriculture, 1984 Proceedings of the Section on Survey Research Methods, American Statistical Association.

Schmehl, R., and Ramos, M. (1990). Evaluation of Classification Tree Methodology for Census Mail List Development, 1990 Proceedings of the Section on Survey Research Methods, American Statistical Association.

Wright, K.E., Davie, W.C., and Sandusky, J.D. (1989). 1987 Coverage Evaluation Estimation, 1989 Proceedings of the Section on Survey Research Methods, American Statistical Association.


# IMPROVING THE QUALITY OF STANDARD STATISTICAL establishment list data 

P.S. Hanczaryk and M.L. Trager ${ }^{1}$


#### Abstract

The Standard Statistical Establishment List (SSEL) is a standardized computer file, or register, of all U. S. business firms and their establishments. The Bureau of the Census compiles, maintains, and uses the SSEL on a continuous basis. It is assembled from administrative record information of other government agencies together with Census Bureau-collected data. The incoming data are processed through a series of quality assurance and edit routines to ensure the accuracy of the data. In recent years, we developed several innovative processes that have improved the quality of the SSEL without increasing the analytical staff. These processes include an automated computer program that searches the business name of unclassified SSEL records and assigns appropriate classification codes; an establishment-edit routine that computerizes the principles and judgments used in the analytical-correction process; and a microcomputer network system that permits on-line access to the 7 million establishment records on the SSEL.


KEY WORDS: Accuracy; Automation, Editing

## 1. INTRODUCTION

A major initiative in the Census Bureau's economic area is to improve the quality of the SSEL data, and to do so within the limits of a restrictive budget. In past years, qualitative improvements to the SSEL oftentimes were accompanied by increases in spending. One would expect that the quality of the data would improve with increased funding.

With the prospects of substantial federal budget deficits in the immediate future, we do not foresee increases in SSEL funding in real terms over the next several years. The Census Bureau, however, is committed to the continued improvement in the quality of the SSEL. In keeping with the overall governmental policy on spending, these improvements must result from more efficient uses of available resources, rather than increases in funding.

Recently, the Census Bureau has attained considerable gains by automating certain routine operations that previously were performed by analysts or clerks. These automated processes, which were implemented initially as cost-saving measures, have served to improve the quality of the SSEL data as well. This paper presents the details of three analytical operations that have been automated. These and other similar types of operations will permit us to improve the quality of our business list within current budgetary limitations.

The SSEL is a computerized list of all U. S. business firms and their establishments. It is maintained and used by the Census Burcau on a continuous basis. The SSEL contains basic economic information for approximately 7 million businesses with paid employees. ${ }^{2}$ Information for another 7.5 million nonemployer businesses is compiled once every five years, as part of the economic censuses. The separate farm portion of the list also is

[^32]assembled once in each 5 -year period, using results from the Census of Agriculture. There are nearly 3 million farm businesses on the SSEL.

The SSEL systematically integrates and controls the information needed for the Census Bureau's economic and agriculture programs. For each establishment, it contains data on the primary and secondary business name, mailing and physical location addresses, Standard Industrial Classification (SIC) codes, employment and payroll data, sales and receipts, and numerous other data items. The SSEL is updated continuously to reflect the latest information available from the economic and agriculture censuses and the many current surveys conducted by the Census Bureau. The list also is updated regularly from administrative data received from the Internal Revenue Service (IRS) and the Social Security Administration (SSA).

Various analytical and clerical editing procedures are applied to the data to help ensure completeness and reliability, including: editing the contents of the data files received from administrative sources, evaluating the completeness and coverage of multiestablishment companies in our annual Company Organization Survey, editing year-to-year relationships of data aggregated to industry and county levels, assigning SIC codes to establishments based on the business name, and several other analytical and clerical operations. These ongoing manual operations, together with our standard computer edits and imputations, help produce a complete and accurate universe list of business establishments.

The Census Bureau has begun to closely examine SSEL-related analytical and clerical operations in an effort to determine processes that can be automated. We have made progress in three areas-the assignment of industrial classification codes from written business descriptions or business names, the correction of data values in analytical-edit processes, and the development of a microcomputer-based system that permits on-line access to SSEL data records. Although we are still enhancing these processes, considerable quality enhancements and cost savings have been realized by automating these tasks.

The first two operations-the assignment of classification codes and the correction of data values-involve processes that historically have been performed by skilled employees. We have attempted to replace manual processes with computerized routines that incorporate the methods and techniques of analysts. The third operation, the microcomputer network system, permits on-line access to the 7 million establishment records on the SSEL. This system, which utilizes many new techniques in microcomputer technology, has eliminated several tedious manual operations in one of our major edits. This paper presents some features of these three operations.

## 2. AUTOMATED ASSIGNMENT OF INDUSTRIAL CLASSIFICATION CODES

### 2.1 Overview

The SIC system and related industrial classification systems group business statistics into a wide variety of categories for analysis and publication. The Census Bureau conducts certain operations that convert descriptions of business activity or business names to industrial classification codes. Previously, these operations were accomplished by a clerical coding staff. A clerk oftentimes can recognize the meaningful words in the written response and determine the proper code. It is not straightforward for a computer program to assign accurate codes because of the many possible variations for each industrial activity.

In an effort to overcome this restraint, we attempted to program the computer to emulate the techniques of a coding clerk. To accomplish this, we first developed a reference file of classification activity responses that occurred frequently and could be coded with a high degree of accuracy; then, we determined the key words of those responses. By eliminating nonessential words and matching the incoming responses on key words only, we were able to standardize the responses, and assign accurate industrial classification codes in many instances.

### 2.2 Automated Coding of Business Descriptions

In the 1987 Economic Censuses, it was essential for the Census Bureau to assign industrial classification codes to over 2.5 million nonemployer businesses. The projected cost for the Census Bureau to clerically assign codes
to these returns was estimated to be $\$ 500,000$. To reduce these large coding costs and to accelerate the coding process, we developed an automated computer program that searched the written description field and assigned appropriate classification codes. In total, we processed $2,563,537$ unclassified records through the automated program. Of these, $1,675,505$ records (approximately 66.4 percent) were assigned industry codes through the automated coding. The computer cost for the automated-coding process was under $\$ 500$.

In addition to description-coding operations, the Census Bureau needed to develop a routine to assign industry codes to SSEL establishment records based on the legal name or trade name of the business. Although the name may not reflect the industrial activity of the business, we will achieve important benefits by automating the business name-coding procedure.

## 23 Automated Coding of Business Names

In economic census years, we expend considerable effort to assign codes to all employer firms on the SSEL. If industrial classification codes are not available from census report forms or other reliable sources, we obtain the proper industrial activity from a classification mailing. The unclassified rate in an economic census year naturally is quite low.

In noncensus years, we rely more heavily on administrative record sources to classify establishments. Specifically, we use industrial classification codes from the SSA and the IRS to update SIC codes of many new establishments on the SSEL. However, many administrative codes are not received in time to update the SSEL. Our SSEL files may contain up to 500,000 unclassified establishments in a given noncensus year. It is important to reduce the number of unclassified establishment records on the SSEL-the SSEL serves as the input source for our County Business Patterns (CBP) publications and various other annual economic programs.

In an effort to lower the unclassified rate, clerical staff conduct a name-coding operation annually--SIC codes are assigned to unclassified records based on the business name. Because the business name frequently does not contain an indication of the industrial activity, many cases cannot be coded. Typically, the clerical staff assigns codes to approximately 40 percent of the unclassified SSEL universe. Assuming 500,000 unclassified records, we would expect the clerical staff to code 200,000 records--the other 300,000 SSEL records would remain uncoded.

The Census Bureau recently conducted an evaluation of the clerical name-coding operation. We mailed classification forms to 13,285 cases that had industry codes assigned from a clerical name-coding process. The code assigned in the name-coding process was compared to the code obtained from the classification-form mailing (which was assumed to be correct). The results showed that 58.9 percent of the name-coded cases were accurate at the 4 -digit level, 64.4 percent were accurate at the 3 -digit level, 70.2 percent at the 2 -digit level, and 81.2 percent at the industrial division level.

We were disappointed with these results. However, the evaluation enabled us to identify groups of cases that yielded inconsistent coding. For example, we discovered that the accuracy for cases coded to Mining and Wholesale Trade was far less reliable than the accuracy for the other industrial divisions. The Census Bureau has implemented a strategy to eliminate Mining and Wholesalc Trade codes assigned in the name-coding process.

Additionally, the evaluation provided an ideal test file for automated name-coding research. From the evaluation results, Census Bureau staff developed a reference list of key words and combinations of key words that could be used to determine the industrial activity of an establishment with a high degree of accuracy. A computer program was written to access this reference list--when the business name on the SSEL record contains the appropriate combination of key words, a corresponding SIC is assigned.

With the coded test file, we were able to determine the accuracy of specific combinations of key words on the reference list. Combinations that did not meet minimum quality standards were culled from the list. To accomplish this, we sorted the evaluation records by name-coded SIC by final SIC (i.e., the SIC from the classification-form mailing). Then, we systematically evaluated the correctness of each 4 -digit SIC group. In general, only those SIC groups with name-coded accuracy rates exceeding 67 percent at the 4 -digit level and 80 percent at the industrial division level were considered for the automated system.

### 2.4 Results

We were pleased with the results of the automated name-coding test--approximately 83.9 percent of the records were accurate at the 4 -digit level, 86.2 percent were accurate at the 3 -digit level, 89.1 percent were accurate at the 2 -digit level, and 94.3 percent were accurate at the industrial division level. Although the automated results do not match the SIC reliability from Census Bureau mailings, the results exceed the reliability of clerically name-coded classifications.

In addition to automating the SIC assignment, the program also segregates the uncoded cases into two groups-those likely to be coded clerically, and those unlikely to be coded clerically. We will provide the clerks only with those records that are likely to be coded, thereby reducing the number of cases the clerks will need to review.

We expect that this systematic approach to name coding--the computerized assignment to certain unclassified records, coupled with the elimination of uncodable records from the clerical-coding operation--will provide major cost and quality benefits to the Census Bureau. Over 50 percent of the clerical name-coding burden will be reduced using this approach.

### 2.5 Methodology

As mentioned above, we developed a reference list of key words and combinations of key words that are used to assign SIC in the automated name-coding program. Certain business names require only one key word to produce a match. For example, a business name containing "Accounting" would be coded to accounting and bookkeeping, without the need for another qualifying word. However, if two conflicting key words are found in the name (e.g., Lawyer/Accounting; or Accounting Books), the case will not be coded.

Most business names, however, require more than one word to constitute a match. For example, "Appliance" is a key word, but is not coded without the presence of another key word; "Appliance Store" and "Appliance Center" are coded to household appliance stores, whereas "Appliance Repair" and "Appliance Repairs" are coded to electrical and electronic repair shops.

The first process of the name-coding program separates the incoming business name into distinct words. Words are defined as consecutive alphabetic characters delimited by spaces or other nonalphabetic characters. Each distinct word is then compared to a key word list. The key word list consists of individual words from the reference list. All words in the business name that match the key word list are concatenated and compared to the predetermined list of key-word combinations. Cases matching that list are assigned the appropriate SIC code.

The following example illustrates the various computer processes of the coding program for the business name, "ABC TERMITE \& PEST CONTROL."

Incoming description

## "ABC TERMITE \& PEST CONTROL"

Process 1: Convert nonalphabet characters to spaces (ampersand becomes a space).

Process 2: Separate the words of the incoming business description.

Process 3: Compare each separate word to a key word list.
"ABC TERMITE PEST CONTROL"

Wd 1 ABC
Wd 2 TERMITE
Wd 3 PEST
Wd 4 CONTROL

Wd 1 ABC - No Match
Wd 2 TERMITE - Match
Wd 3 PEST - Match
Wd 4 CONTROL - Match

Process 4: Concatenate the words that match the key word list:

## "TERMITEPESTCONTROL"

Process 5: Match the concatenated string to the reference list of descriptions (also concatenated). Assign the appropriate 4 -digit industry code to matched records:
"ABC TERMITE \& PEST CONTROL" SIC Code 7342 assigned.

This technique allows us to match on key words without checking each character string for the possible occurrence of these words. For example, the business names "ABC TERMITE \& PEST CONTROL," "JOHNSONS TERMITE AND PEST CONTROL," and TERMITE AND PEST CONTROL OF DADE COUNTY" will be coded identically because the same key words are present.

This illustration presents a basic overview of the program; I have omitted several aspects of the program for the sake of clarity. One feature of the program that I would like to discuss is the mechanism used to identify cases that are unlikely to be coded clerically. As mentioned above, we will provide the clerks only with the records that are legitimate candidates for coding. This will reduce the magnitude of the clerical coding operation.

Many unclassified establishments cannot be coded clerically because the name does not indicate the type of industrial activity. For example, businesses with such names as "ABC Corporation" and "JJ Enterprises" clearly cannot be coded. To improve the efficiency of the clerical coding operation, it was essential to remove these types from the coding files.

Other cases can be coded clerically. However, these were not coded through the automated system, primarily because the reference list was not exhaustive. With the countless possibilities of key words in the business names and limited staff time, we could not provide for all combinations. The following business names are typical of cases not coded through the automated system, but which may be clerically coded:

SUITLAND HOSPITAL PROFIT SHARING PLAN CAROL'S GIFT AND NOVELTY SHOPPE JEFFERSONVILLE MOPED SALES CANDY STORE - ABC CANDY etc.

Although each of these descriptions contain a key word, the combinations of these key words did not match the reference list. For example, "SUITLAND HOSPITAL PROFIT SHARING PLAN," matches on four key words, "HOSPITAL," "PROFIT," "SHARING," and "PLAN." Although our reference list contains the concatenated phrase "PROFITSHARINGPLAN," it does not contain the more specific phrase "HOSPITALPROFITSHARINGPLAN," so the business name was not coded. A capable clerk certainly would realize that this business name indicates profit-sharing activity, and would assign the appropriate code. ${ }^{3}$

Until a comprehensive reference file can be developed, we sought to develop a practical method to isolate business names with a legitimate chance of clerical coding. We concluded that business names without a key word had an extremely low chance of clerical coding; and those names that contained at least one key word had a legitimate chance to be clerically coded. By segregating the uncoded cases into two separate output files, depending on whether a key word is present, we established a mechanism that met our objectives splendidly. Only a small number of cases without a key word can be coded clerically.

[^33]The Census Bureau will use the automated name-coding system this November to assign industry codes to unclassified establishments on the SSEL. We expect to assign codes to nearly 25 percent of the records, and to remove approximately 30 percent from further research. Accordingly, nearly 55 percent of the unclassified records will be eliminated from the clerical name-coding operation. Assuming an initial unclassified universe of 500,000 records, we will exclude approximately 275,000 cases from this year's clerical-coding operation.

Overall, the Census Bureau will assign valid classifications to approximately 200,000 cases $-125,000$ in the automated-coding system, and 75,000 in the clerical-coding operation. The resulting SSEL unclassified rate will be reduced from approximately 7.1 percent to 4.3 percent.

### 2.6 Future Enhancements

Although we are pleased with our progress to date, many refinements will be made in the future to enhance the utility of this system. First, we will supplement the current reference list with many additional descriptions. This year's clerical coding operation will provide the basis for these enhancements. Specifically, business names that clerks routinely code this year will be added to the reference list. With a more complete reference file, a higher percentage of cases will be coded in subsequent operations. Second, we will research the use of other data items available on the SSEL records, such as legal form of organization codes or employment and payroll figures, to aid in the coding operation. Third, we intend to standardize the key words by eliminating plurals and suffixes. For example, "Radiology," "Radiologist," and "Radiologists" could all be represented by the base derivative, "Radiolog." This technique would not only reduce the number of required key words, it may increase the successful match rate by including all the alternate forms of a word.

We intend to incorporate these and other improvements into the automated name-coding system within a couple of years. Our goal is to have a completely automated name-coding system, without any clerical intervention, in place by 1993. Moreover, we would like to adapt this automated technique to other coding operations at the Census Bureau. For example, it may be possible to develop a similar program to assign product-level codes in the 1992 Census of Manufactures.

## 3. AUTOMATED EDIT RESOLUTION

The Census Bureau also has made progress in automating certain analytical edit operations. The economic areas produce numerous edit listings for analytical resolution. Analysts oftentimes resolve the establishment outliers on the listings by examining the relationship of the various data components and indicators on the SSEL record. By formalizing the resolution rules and judgments of the analysts, and incorporating these rules and judgments into a computer routine, we have successfully replaced certain analytical operations with computer processes. These computer processes are not innovative or complex, but simply quantify the knowledge of human analysts in an effort to reproduce their judgments.

As part of the SSEL's imputation routine, reject records are provided to the CBP analytical staff for review and possible correction. Several categories of reject records are outputted for review, depending on specific data comparisons. Analytical staff from CBP correct the reject cases based on data relationships and other indicators on the SSEL record. However, CBP analytical staff has limited time to work the edit listings. Over 125,000 establishment records are included on the listings, and the results must update the SSEL within a few weeks, before year-end processing is completed. Because of these constraints, CBP analytical staff only review the extreme outliers, and many reject records remain uncorrected. As a result, subsequent CBP analytical edits become much more difficult to process.

We were determined to develop an automated procedure to properly correct the reject records-it was crucial to review all of the reject records in the edit, not just the extreme outliers. As a first step in the developmental process, three of our best analysts worked a representative sample of cases from the previous year's edit listing. As is typical with most analytical review processes, many of their correction decisions required judgments in addition to certain accepted principles.

From the corrections made, the analysts attempted to quantify the overall set of rules used in the correction process. This phase of the development process was the most tedious. As it turned out, the analysts developed the rules through an iterative process. Fundamental rules initially were formulated and built into a computer program. We then processed the sample file of reject records through the program to evaluate the reliability of the rules. Several exceptions and additions to the rules were identified and incrementaliy incorporated into the program. This process was repeated several times until our complete set of rules, with all the necessary variations, was established.

The results of this automated process have been outstanding. We evaluated the corrections applied by the automated procedure to the corrections of the same cases by CBP analysts. For cases with corrections from both processes, the results were remarkably similar. However, considerably more cases were coded through the automated process. To illustrate, the analytical-correction process for data year 1987 generated under 750 corrections. The automated correction routine generated 6,635 corrections for data year 1988 and 7,402 for data year 1989.

We realize that conclusions reached from an automated system that is based on the judgments and techniques of analysts will not be one hundred percent accurate. However, the system does provide for nearly optimal decisions based on the available information. Further, this systematic process ensures consistent corrections, whereas analytical corrections may vary according to the individual methodologies used.

By formalizing and automating the analytical methodology from this one operation, the Census Bureau has realized significant qualitative and cost benefits. This approach does not provide for any new imputation methodologies or editing strategies, but merely replaces analytical corrections with computer corrections.

## 4. ON-LINE MICROCOMPUTER SYSTEM

### 4.1 Introduction

In addition, the Census Bureau has developed a microcomputer network system that permits on-line access to the 7 million establishment records on the SSEL. We created this system to eliminate several tedious manual operations in the CBP analytical edits. Foremost, the on-line system permits instantaneous access to establishment records, replacing the cumbersome microfilm research that analysts previously used. Further, the system enables the analysts to enter corrections directly into the system, bypassing the need for a clerical data transcription process.

The system has significantly improved the overall efficiency of our analytical-edit processing in CBP. We have realized major benefits from the on-line system, including enhanced data quality and more timely publications. In addition, the system has reduced processing costs by shifting processes from the mainframe to the microcomputer environment, and by eliminating a large portion of the clerical keying. Among the many other benefits, the network has provided the capability to prepare special tabulations of the business universe quickly and at minimal costs.

### 4.2 CBP Edit Processing

The CBP statistics provide detailed information on the industrial makeup of the United States. Specifically, the CBP publications present aggregate data on number of establishments, employment, and first quarter and annual payroll. Data are tabulated by industry for all counties and states in the Nation. Before publication, we process the data through various analytical edits. The major CBP edit, the cell edit, compares current-year summary totals for establishments, employment, and payroll to prior-year summary totals at the state and county levels. If the current-year to prior-year totals exceed specified tolerance limits, appropriate establishment records in the cell are selected for review.

In past years, analysts researched microfilm in an effort to resolve data discrepancies. The microfilm contained comprehensive information on each business establishment, whereas the data on the edit listings provided only basic information. Two versions of the microfilm were created-the first provided the business universe in
identification-number order, and the second provided the universe in state/county/SIC order. Because of space constraints, however, the second version contained information only for the larger establishments in the universe-those with annual payroll of $\$ 125,000$ or over. This hindered our analytical research in certain instances.

Approximately 10,000 corrections are identified annually in the cell edits. Analysts entered the corrections directly on the cell-edit listings. After all the corrections for a state were identified, clerks transcribed the corrections onto posting documents, and then keyed these corrections into our mainframe computer system. As with most clerical-keying operations, numerous errors in transcription as well as inadvertent omissions in the keying occurred.

To make better use of our staff time and improve data quality, we developed an on-line microcomputer system. This system has enabled CBP to reduce the analytical-edit review time by 3 months-from 8 months under the manual system to 5 months under the on-line system. More importantly, we bave improved the quality of the published data-many additional establishment outliers are reviewed, certain illogical data corrections are prohibited, and numerous keying errors are eliminated.

### 4.3 Benefits from an On-line System

In 1989, we replaced the microfilm files previously used for analytical research with a single reference file containing essential information for every case in the business universe. The file was loaded onto an optical disk device, and was accessible to each CBP analyst through a networked microcomputer system. A multiple-access mechanism was developed to permit more than one analyst to view the same cases simultaneously.

The reference file is indexed to permit random access in two different sequences--in identification number order, to view a specific case, or in state/county/SIC order. This second sequence is used widely during cell-edit research, as the cell-edit listings also are arranged in state/county/SIC order.

The on-line establishment record contains over 50 data fields, which correspond to analytical requirements in CBP. Further, we established a series of cell-edit commands to expedite the review process. These commands permit immediate viewing of different categories of establishment records-establishments in the next SIC within County and State, establishments in the first cell of the next State, establishments that have become inactive, and many other categories of establishment records.

Among the advantages of the on-line reference file:

- On-line access is virtually instantaneous. Microfilm research was a tedious and time-consuming operation-microfilm reels were loaded on readers, and it could take several minutes to locate the establishment record in question.

The on-line system permits many users to view establishment records concurrently. In the past, multiple access was inconvenient as different analysts often required the same microfilm tape.

- All establishment records in a state/county/SIC cell can be accessed through the on-line system. This represents a distinct advantage over microfilm, which included only the larger establishment records.

The cost to prepare the microfilm has been eliminated. In past years, the Census Bureau paid a private contractor approximately $\$ 15,000$ for this job. Before providing the files to the coniractor, we sorted the records on the mainframe, at a significant expense. Although records still must be sorted for the on-line system, this operation is performed at the microcomputer level at minimal costs.

Further, the on-line system permits analysts to enter edit corrections directly into the microcomputer network, which eliminates the clerical-transcription process. This has significantly reduced the number of keying errors. In addition, we introduced edits that prevent illegitimate entries. As an example, the system does not allow invalid identification number entries to update the database. Also, data values must pass internal consistency checks before updates are accepted. These edits, although modest, have identified numerous errors. We will introduce more comprehensive industry-and geographic-specific edits in future applications.

Our microcomputer system also satisfies the requirement in another area--special tabulations. The Census Bureau has encountered some difficulties in satisfying special tabulation requests. In the past, special requests were produced exclusively on the mainframe computer. Not only is mainframe processing expensive, the staff responsible for programming the tabulations is assigned to a separate division at the Census Bureau. Interdivisional specifications are required, which contribute to long delays in the preparation of the tabulations.

Because the business universe is now maintained on the microcomputer network, we can produce special tabulations quickly and efficiently without crossing divisional boundaries. Moreover, we have developed a general-purpose routine that offers non-programmers as well as programmers the capability to prepare tabulations easily.

Although we are pleased with the efficiencies provided by the microcomputer system, we are planning for enhancements in the coming years. To begin with, we will supplement the on-line record with additional data elements to assist with analytical decisions. More importantly, we intend to provide interactive edit capabilities. Corrections will be made directly to the data files, and the revised summary totals and edit-tolerance values will be recomputed and displayed on the screen. Instantly, analysts can examine the impact of the correction on summary totals. If the correction does not alter the totals sufficiently to ameliorate the problem, additional changes may be warranted.

### 4.4 Developing the On-line Microcomputer System

The first step in preparing the on-line reference file is to download the entire SSEL universe from our UNIVAC mainframe. This entails moving a 7 million record database, comprising close to 2.5 gigabytes of data, through several processing steps. First, the file is transferred to 9 -track magnetic tape (starting next year, with the move of the SSEL database to a Digital Equipment Corporation (DEC) minicomputer, we will transfer the file directly from the DEC to our microcomputer system via the Bureau's Wide Area Network, thus eliminating the tedious conversion to tape). Next, the files are stored from tape to optical disks, which serve both as a backup medium and as input to the remaining processing phases.

Optical disk technology was the clear choice as a storage medium for several reasons: it offers the most practical way to store the immense amount of information involved on one physical piece of hardware (although recent developments in hard disk technology now make possible storage capacities in excess of 1 gigabyte); and, because it is a removable media, it provides for easily accessing data sets from previous years--simply by removing one optical disk and inserting another.

Even with the advantages offered by optical disks and microcomputer systems in general, one additional step was essential to the success of this project-data compression. Storing and accessing 2.5 gigabytes of data is relatively expensive and time-consuming; we were determined to minimize the hardware (and thus cost) required to build the system, and to provide the fastest database accesses and record retrievals possible under the circumstances. Compressing the data set yielded several auxiliary benefits as well, such as dramatically faster sorting, merging, indexing, and other preparatory computer processing steps.

In evaluating data compression techniques, we had several objectives that we felt were mandatory if the system was to succeed. First, we wanted to achieve a reasonable degree of compression-at least 50 percent; second, the algorithm used had to be fairly straightforward, so that minimum processing time would be spent decompressing data and so that the program code itself could be easily maintained or modified if necessary; third, compression had to be attainable on a record by record basis, as opposed to an entire file; and, last, the algorithm had to produce lossless data compression.

Commercial data compression packages proved not to be feasible for our purpose, as they invariably operate on an entire file only. In a random access database environment, a single record can be retrieved from any location in the file-obviously, data compression techniques that are dependent on the entire file to build upon repeating character patterns would not be able to dynamically decompress a single record at a time. Further, commercial packages could not be integrated easily with our custom database software and certainly could not be modified or maintained as we saw fit.

We turned then to writing our own compression/decompression software, but first had to settle on an appropriate algorithm. As mentioned, the algorithm had to produce a lossless data set after compressing and then decompressing. A number of statistical compression techniques, especially those used in image compression, can not guarantee literal restoration of the original data-some bits are lost, although in graphical images the loss is not fatal, or even noticeable in some cases. For our business universe, however, the loss of even a single bit could have drastic consequences. We narrowed the field to several promising algorithms, probably the most well-known of which is the Lempel-Ziv Welch technique (LZW). However, LZW as applied to our needs has two substantial drawbacks: it is a complex algorithm and would have made software maintenance and modification difficult at best, especially if one is not well-versed in the intricacies of LZW; also, because LZW works by finding and encoding repeating character patterns, it is much more efficient when applied to long strings of data where the probability of repeating sub-strings is higher. We could achieve only 45 percent compression when using LZW on the SSEL a record at a time. Although this compression ratio can be improved with additional steps, such as converting all numeric data types to binary representation (a form of compression in itself), the drawbacks argued against using LZW.

Instead, we settled on a simplistic yet very efficient and easily implemented compression algorithm called variable length Huffman bit encoding. Huffman codes produce a unique bit pattern for each original character to be compressed, with the most frequently occurring characters requiring the fewest number of bits for encoding. Typically, characters that occur often in a file (or record) require only 1 or 2 bits for encoding (as opposed to the nominal 8 bits comprising each byte, or character, in the ASCII character set). Less frequently occurring characters might require more than 8 bits for encoding, but these are more than offset by the high frequency characters. Unlike LZW and other techniques, Huffman codes can be instantly built or decompressed as the bit stream is read-there is no interdependency from one character to the next; thus compression efficiency does not require long strings of data to find repeating patterns-each byte is treated separately. Writing a Huffman compressor/decompressor does require bit level access to the data-this was most easily accomplished in Assembly Language, although other languages that provide bit manipulation capability, such as C , could be used as well.

The simplicity of the Huffman process allowed us to supplement it with one additional technique--run length encoding, which we applied to one character only-the space character. By its nature, the SSEL universe contains many spaces, especially in the name and address fields (we further capitalized on the compressibility of spaces by ensuring that all numeric fields were filled with leading spaces, rather than zeros). As the original SSEL data are read into the compressor, we examine the stream for consecutive spaces and encode them by appending 4 additional bits to the Huffman code for a space. These 4 bits represent a count of the number of additional spaces present in the string. Thus, up to 16 consecutive spaces ( 128 bits) can be represented by 5 bits-a 1 bit Huffman code indicating a space, and an additional 4 bit nibble containing a repeat count value up to 15 .

Overall, we achieved 65 percent compression for the SSEL, reducing storage requirements from over 2 gigabytes to 800 megabytes.

Huffman codes are constructed using a Huffman tree, as illustrated in the attachment (a complete explanation of the process can be found in the book, Data Compression, by Gilbert Held, published by John Wiley \& Sons, Lid.).

Building a Huffman tree requires that the original file be pre-processed to produce a frequency distribution of all characters contained therein. The distribution is then arranged so that the most frequently occurring characters are positioned at the top of the tree, with their actual frequency of occurrence notated. Then, starting at the bottom of the tree, the branches of the two least frequently occurring characters are combined and their frequencies added to produce a new branch. Taking this new branch into account, the process of combining the lowest frequencies continues up the tree until all branches have been merged. Each node or intersection of branches is then assigned a 0 bit on one side and a 1 bit on the other. The Huffman code (or bit string) for each character is then determined by starting at the right most branch and accumulating the 0 or 1 notations found at each node. The resulting bit patterns for each character can then be entered into a look-up table contained in the actual software compressor.

Having achieved compression, we were left with determining the best access strategy for retrieving records from the database. Several techniques are possible, the most common of which involves storing the actual data records
(compressed) in one file, and the associated access keys in separate files. Thus, as the original file is compressed and loaded onto optical disk, a separate file containing index records is built. In the index file are separate records for each ID Number in the database, along with the location in the compressed database (byte offset) of the corresponding data record. A record is retrieved by looking up the ID in the index file, moving the read/write pointer in the main file to the appropriate offset, and finally reading and decompressing the record. Accessing multiple records with the same key, such as State/County/SIC, requires a different index file containing one record for each State/County/SIC and the byte offset in the main database of the first record with that key. Blocks of records with the same access key can then be read and decompressed. One disadvantage to this technique is that accessing blocks of records with the same key can be rather inefficient unless the file is sorted in order of that particular key combination. Otherwise, the individual records comprising the group to be retrieved are not contiguous in the file; many physical accesses to the disk are necessary to gather up all records. Obviously, the file can not be sorted in many different orders and the fastest, most efficient database accesses are limited to single record retrievals by ID and multiple record retrievals by State/County/SIC.

Although the advantage to this technique is that it can be programmed in virtually any language, other file structures are possible, depending on the language used, that can simplify the building of access strategies and auxiliary index files. For example, some COBOL compilers offer support of a file type called VSAM, which is a derivative of ISAM (Indexed Sequential Access Method) with the added benefit of allowing variable length records (recall that variable length records are a must when using data compression). Using VSAM, one compressed database can be built with many different access keys or key combinations and without the overbead of building, maintaining, and accessing separate index files.

Our first implementation of an access strategy used the former (non-VSAM) approach, and has given excellent results from a performance point of view. We are currently rewriting the software, however, using a particular version of a COBOL compiler that provides VSAM capabilities. The primary goal of the rewrite is not to improve performance, but to expand the number of access keys used to retrieve records (i.e., to broaden the potential "views" of the database) without an overly complicated system of index files and custom coding.

The on-line database system, although quite large even by mainframe standards, performs as well as any interactive system we have seen. Instantaneous accesses, flexibility, minimal cost, and ease of use all argue convincingly for increased use of microcomputer systems as worthy alternatives to the more traditional large-scale mainframe or minicomputer plafforms.

## 5. SUMMARY

The Census Bureau has begun to examine our analytical and clerical operations in an effort to determine processes that can be automated. By developing computer processes to replace manual operations, we can achieve tremendous processing efficiencies.

We have made progress in automating three manual operations-the assignment of industrial classification codes from written business descriptions or business names, the correction of data values in our establishment reject edit, and the development of an innovalive microcomputer network. We will continue to monitor and enhance these three automated processes, and will seek to automate other manual processes as well. These types of automated routines will permit us to improve the quality of the SSEL within budgetary limitations.

## REFERENCE

Held, G. (1988). Data Compression (2nd ed.), New York: John Wiley \& Sons, Ld.

Huffman Tree


## SESSION 8

## Improving Data Processing and Estimation

# A REVIEW OF SOME MACRO-EDITING METHODS FOR RATIONALIZING THE EDITING PROCESS 

L. Granquist ${ }^{\text {t }}$


#### Abstract

The paper presents descriptions, studies, results and conclusions (including recommendations) on: The Aggregate Method, The Hidiroglou-Berthelot Method (Statistical Edits), The TopDown Method, The Box-Plot Method and the graphical Box Method. By simulations on survey data in a production environment it has been found that these methods can reduce the manual verifying work of suspected data by $35-80 \%$ as compared to corresponding traditional microediting methods without any loss in quality. They are easy to understand and can easily be implemented in existing computer assisted error detecting systems by just adding programs to the existing system. The particular features of the methods are all compared and the main finding is that the methods are very similar and that they all aim at finding efficient boundaries to generally used micro edits. Boundaries should only be based on statistics on the weighted keyed-in data.


KEY WORDS: Simulation; Data quality; Interactive system; Micro-editing

## 1. OBJECTIVE

The objective of this paper is to present macro-editing methods as a solution to the over-editing problem. In experimental studies in data processing environment and in statistical production they have proved superior to micro-editing methods in editing quantitative data. Savings of the manual verifying work from 35 up to 80 per cent are reported.

The methods are described in a brief and schematical form to make them easy to understand and to clarify the essential features of the treated type of macro-editing methods. The rational aspects of macro-editing as compared to micro-editing are emphasized.

Detailed descriptions of the methods and studies underlying this paper are found in the references given in the text and in the reference list.

The paper ends with a summing-up discussion on macro-editing versus micro-editing methods.

## 2. INTRODUCTION

In Granquist (1984) macro-editing is introduced as a means for tackling unknown misunderstanding errors. (Misunderstanding errors are defined as errors that arise due to ignorance or misapprehension of questions, concepts or definitions, but they also comprise tactical errors).

At Statistics Sweden we started experimenting with macro-editing methods by performing edits on aggregates. When a suspect aggregate was found, the first step was to find out whether there were one or two individual

[^34]observations which caused the aggregate to be classified as "suspect". When working on means for identifying such data, we found that the macro-editing methods were more efficient than traditional micro-editing methods in detecting suspect data. Then we decided to work on this first step of the macro-editing concept only, until we managed to get such methods extensively implemented at Statistics Sweden. We are now completing this first part of the macro-editing project and plan to continue with the second part next year.

From our experiences, we have found it appropriate to define macro-editing (this type of macro-editing methods) as "statistical edits" applied to expanded data. The methods which are reported in this paper may be characterized as methods to provide micro-editing methods with more efficient acceptance bounds. They bring a priority thinking to the verifying work.

## 3. THE OVER-EDITING PROBLEM

The essential problem of traditionally applied micro-editing procedures might be formulated as follows: Too many checks with too narrow bounds cause many error-detecting systems to produce too many error messages to be verified manually by clerks. These clerks are not able to assess the importance of a suspected error. Every flagged data has the same weight and claims the same amount of resources, but many errors have a negligable impact on the estimates, as they are small or cancel out. Generally, the bounds of the checks are set subjectively on the principle "safety first" which means that only those data are accepted for which there are no reasons to suspect any error. For example, a very generally used check in business surveys in Statistics Sweden is to flag every data where the relative change since the previous survey exceeds $\pm 10$ per cent. A considerable amount of over-editing is the general consequence of such micro-editing procedures.

## 4. "DEFINITIONS"

This paper deals with checks on quantitative data which flag "suspicious" data for a manual review. This type of checks may be considered the opposite of validating checks which indicate data that are erroneous. The first type Ferguson (1989) calls "Statistical Edits". They use the distributions of current data from many or all questionnaires, or historic data of the statistical unit, to generate feasible limits for the current survey data.

In this paper macro-editing implies a procedure for pointing out suspicious data by applying statistical checks/edits based on the weighted keyed-in data. The upper and lower limits of a macro-editing check (macro edit) should be based only on the data to be edited and on the importance of the data on the total level.

## 5. THE EVALUATION TECHNIQUE

The methods have been evaluated by simulation studies on real survey data. The simulations were carried out by prototypes of a complete editing process for the survey. The results of the studied method were then compared with the results of the micro-editing methods applied when the survey was processed. The changes made as a result of the editing process of that survey were entered in a change file and the study consisted in finding out (by calculating a few measures) which data in the change file were flagged by the macro-editing method and which were not. The rationalizing effect was measured as the reduction of the number of flagged data, and the "quality loss" as the impact of the remaining errors (the errors found in the processing of the survey, but not flagged by the macro-editing method).

## 6. MACRO-EDITING METHODS

Below, descriptions and a few results are given on "The Aggregate Method", "The Hidiroglou-Berhelot Method" ("Statistical Edits") and "The Top-Down Method". "The Box-Plot Method" and "The Box-Method" are also treated, but as modifications or developments of the Aggregate Method and the Top-Down Method respectively. The Box-method is under development.

### 6.1 The Aggregate Method

### 6.1.1 References

The original Aggregate Method is described in detail in Granquist (1988b). It was developed in SAS as a prototype of a complete editing system for a survey on Employment and Wages (SEW) to be run on the mainframe. A modified version of the method is reported in detail in Lindström (1990). It was developed in PC-SAS as a prototype for the SEW to be run on a personal computer.

### 6.12 Description of the method

The basic idea is to carry out checks first on aggregates and then on the individual records of the suspicious (flagged) aggregates. All records belonging to a flagged aggregate of any of the variables form the error file. The checks on the individual level are carried out on that error file.

The most essential feature of the aggregate method is that the acceptance limits are set manually by reviewing lists of sorted observations of the check functions. Only the "s" largest observations and the " m " smallest observations of the check functions are printed on the lists.

Both the check function $A$ on the aggregate level and the check function $F$ on the individual level have to be functions of the weighted (according to the sample design) value(s) of the keyed-in data of the variable to be edited. By using the weighted values in function A there are no problems to calculate the checks on the aggregate level in the same way as is done in traditional micro-editing. The macro-editing process can be run as smoothly as a micro-editing process.

The lists of the sorted observations can be used either directly as a basis for reviewing the observations manually (when identifiers are printed out together with the observations), or indirectly as a basis for setting acceptance limits for an error detecting program, which then produces error messages of suspect data for manual reviewing. The advantage of using an error detecting program is that the process can be made to look identical to the old one. To implement the Aggregate Method, only programs for printing the lists of the sorted observations are needed. Such programs can easily be added to the old system.

Obvious improvements are to provide the lists of the tails of the distributions with such statistics as the median, the quartiles, the range, the interquartile range or graphs. Here, Box-plots (see Tukey (1975)) are recommended.

### 6.1.3 Experiences

The checks used in our studies on the Aggregate Method consisted of a ratio and a difference check. Both checks had to be rejected to label a value as suspicious. Several simulation studies were carried out on SEW data from different months. The main finding is:

The Aggregate Method can supply the main element of the editing process in the production processing of a survey. The method reduces the verifying work by $35-80$ per cent without losses in quality or timeliness. The macro-editing concept is a realistic alternative or complement to micro-editing methods, and can be applied during the processing of the data under the same conditions as computer-assisted micro-editing methods, which reduces the manual verifying work to a considerable extent.

Other findings were:
Micro-editing may not always detect even serious errors.
The acceptance intervals of the checks should be much wider than what is current practice to-day and should not be symmetric around 1 for the ratios and around zero for the differences.

A good strategy is to set the limits as close as possible to the first outlier on both sides.

In a study with the SAS-PC prototype (see Lindström (1990)), a very important finding caused us to widen the "definition" of this type of macroediting methods:

The result of this study was a reduction in the number of flagged data by nearly 80 percent. However, the loss in quality was slightly higher than in the other studies. It was found that this loss in quality was caused by a few large errors, which did not cause the aggregates to be flagged. These errors were very easy to detect. One interesting solution was to let all data pass a ratio check with very wide acceptance intervals. This could be done in the micro-editing part of the procedure or as a final check.

The aggregate checks were then questioned. The only advantage of the aggregate checks is that they can save storage or computer time. However, when there are no problems with either the storage capacity or the computer cost, the aggregate checks can be skipped.

This finding caused us to widen the definition on macro-editing to the definition quoted above. Thus this method is still a macro-editing method but it should not be called the Aggregate Method. When the tails of the distribution of the check function is provided with Box-Plots (which is recommended), it is called the Box-Plot Method.

One conclusion may be that the Aggregate Method should not be used for small surveys, where the Box-Plot Method is a better alternative.

### 6.2 The Top-Down Method

### 6.2.1 References

The Top-Down Method is described in Granquist (1987) and in Lindblom (1990). The method has been implemented in the main-frame production system of "The Survey of Delivery and Orderbook Situation" (DOS). The production system is written in APL. A prototype written in PC-SAS for running on a micro computer is developed and reported in Lindblom (1990).

### 6.2.2 Description of the method

The idea behind the method is to sort the values of the check functions (which are functions of the weighted keyed-in values) and start the manual review from the top or from the bottom of the list and continue until there is no noticeable effect on the estimates.

The method is described as it is applied in the DOS production system. The generalization is obvious.
The procedure is governed by an inter-active menu program, written in APL. The in-data file is successively built up by the records of the three batches from the data-entry stage. There are three functions to select the records to be studied, i.e.
i) the 15 highest positive changes
ii) the 15 highest negative changes
iii) the 15 greatest contributions
which for every variable can be applied to the total, and to the 38 main domains of study. For a selected function and domain of study, the screen shows the following list for the 15 records of the in-data file sorted topdown:

## IDENTITY DATA WEIGHT WEIGHTED VALUE TOTAL

The operator selects a record and immediately gets the entire content of the record on the screen. If an error is identified he can up-date the record on the screen and immediately see the effects. The record usually looses its position on the top 15 list and the total is changed. The operator goes on until further editing does not change the total.

### 6.2.3 Experiences

The Top-Down procedure, was developed as a complementary or reserve procedure to a traditional micro-editing procedure. The two procedures were developed at the same time but the micro-editing procedure, intended as the basic procedure for DOS, was at once replaced by the Top-Down procedure.

The Top-Down Method can be used as an editing method during the processing of a survey without losses in quality and timeliness. It can reduce the verifying work by $50-75$ per cent and it is continuously educating the staff on the subject-matter and problems of the survey. The clerical personnel are very satisfied because they dominate the editing task and can see the effects of their work.

Since the first production with the macro-editing process, the number of records for manual review has decreased slowly. The DOS statisticians have become convinced that: The editing should only be carried out at the "total manufacturing industry" level.

Though there seems to be a certain amount of over-editing, it is doubtless the most rational editing procedure at Statistics Sweden.

According to Anderson (1989a) the method is also used as an out-put editing method and considered as the most efficient out-put editing method in use at the Australian Bureau of Statistics.

### 6.3 The Midiroglou-Berthelot Method (Statistical Edits)

### 63.1 References

The Hidiroglou-Berthelot Method (the HB-Method) is described as a micro-editing method in HidiroglouBerthelot (1986). The method is a ratio check inspired by Tukey's Explorative Data Analysis (EDA) methods (see Tukey (1977)), and in the paper it is considered as a solution to some problems connected with the traditional ratio-check method. It is in use at Statistics Canada and known as "Statistical Edits".

As a macro-editing method it is reported in Höglund (1989). At Statistics Sweden the HB-Method has been studied on different surveys. Only the study on "The Survey of Delivery and Orderbook Situation" (DOS) has been reported in English (Höglund (1989)).

### 6.3.2 Description of the method

The HB-Method is a ratio method, for which the bounds are automatically calculated from the data to be edited. The method uses the robust parameters median, quartiles and interquartile ranges because the bounds should not be influenced by single outlier. The drawbacks of the ordinary application of the ratio method are that the outlier on the left tail may be difficult to detect and that the method does not take into account that the variability of ratios for small values is larger than the variability for large values. The HB-Method solves these drawbacks by a symmetric transformation followed by a size transformation.

$$
\begin{aligned}
& S_{i}=\left\{\begin{array}{l}
1-R_{\text {MEDIAN }} / R_{i}, O<R_{i}<R_{\text {MEDIAN }} \\
R_{i} / R_{\text {MEDLAN }}-1, R_{i} \geq R_{\text {MEDIUN }}
\end{array}\right. \\
& E_{i}=S_{i}=\left(\operatorname{MAX}\left(X_{i}(t), X_{i}(t+1)\right)\right)^{U}, 0 \leq U \leq 1
\end{aligned}
$$

$X_{i}(t)$ and $X_{i}(t+1)$ are the expanded values of the item from period $t$ and $t-1$ respectively. $E_{Q}, E_{Q}$ are the first and third quartiles of the transformation $E$.

$$
\begin{aligned}
& D_{Q}=\operatorname{MAX}\left(E_{\text {MEDON }}-E_{Q l},\left|A * E_{\text {MEDLN }}\right|\right) \\
& D_{Q^{3}}=\operatorname{MAX}\left(E_{Q 3}-E_{\text {MBDLN }},\left|A * E_{\text {MBDLN }}\right|\right),
\end{aligned}
$$

which gives the lower and upper limits of the checks:

$$
\begin{gathered}
l=E_{\text {MBDAN }}-C * 10 * D_{Q 1} \\
u=E_{\text {MBDIAN }}+C * D_{Q^{3}}
\end{gathered}
$$

$A$ is considered as a constant, equal to 0.05 , which means that there are only two parameters, $U$ and $C$, which have to be set in advance to get the method to run.

Figure 1 may explain the transformations and the method. The figure is produced by the prototype of the Box Method on data from the Survey of Employment and Wages. The acceptance limits have been calculated for two pairs of values of the parameters $U$ and $C$. This method can be applied when the implementation of the HB-Method forms an operation in the production process. The acceptance bounds would then be completely determined by the data to be edited.

Figure 1


### 6.3.3 Experiences

This is an excellent method. It can well compete with the Top-Down Method. In a comparison with the aggregate method on data from the Survey of Employment and Wages, the HB-method was found to be superior. We have found that the parameters $U$ and $C$ are not very sensitive. The same values can be used for many variables of a survey which indeed makes the method easy to use.

### 6.4 The Box-Plot Method

### 6.4.1 Introduction

The concept of Box-Plots was introduced by Tukey (1977). The Box-Plot Method as a micro-editing method is reported in Anderson (1989b).

Anderson (1989a) reports an experiment carried out on data from the Australian Bureau of Statistics (ABS) survey "Average Weekly Earnings" (AWE). ABS extended the bounds of every ratio check used in that survey
 interquartile range. The study indicates that 75 per cent of the resources for the manual reviewing of flagged data from the whole editing process could be saved by limiting the manual review to "extreme outlier". Anderson used the same evaluation technique as we have utilized in our studies at Statistics Sweden. The remaining errors in data had no significance at all on the estimates.

In the latter version of the report, Anderson suggests that the lower and upper bounds of the checks should be set on the basis of a manual analysis of box-plots of the check functions. Then the bounds can be modified by taking into account outher near the bounds for extreme outlier. Above all, the survey staff will get full control of the data editing.

### 6.4.2 Description of the method

The distributions of the check functions of the expanded (according to the sample design) values of the items are displayed as box-plots. Then the acceptance intervals for the checks are set by the staff on the basis of these graphs and put into the regular error-detecting program. By Anderson (1989a) and the studies on the Aggregate Mcthod reported in this paper, the definition of extreme outlier may serve as guide-lines for efficient limits.

### 6.5 The Box Method

### 6.5.1 Introduction

The Box Method is a graphical macro-editing method under development at Statistics Sweden. A first version of a prototype for the Survey of Employment and Wages is expected by January 1991. The basic principles are to utilize computer graphics to visualize the distribution of the check function of the weighted data and the interactivity of a computer to get indications when to stop the manual verifying work. The method may be considered as a combination of a generalized Box-Plot method and the Top-Down Method.

### 6.52 Description of the method

The values of the items are weighted and then put into the check function. Any mathematical expression may be used as a check function. The values of the function are plotted on the screen and acceptance regions of any shape can also be provided. The reviewer draws a box around the observations he wants to review. On the screen, the data then appear on in advance selected items of the records belonging to the data points inside the box. For every check function the user can select the items of the records to be displayed. A change is entered inter-actively and some data (statistics) of the impact of the change will be displayed.

The method may also be used as a tool to find appropriate values of acceptance regions for other editing methods (e.g. the HB-Method).

## 7. SUMMARY

### 7.1 Description of the methods

The basic principle of all the reported macro-editing methods is that the acceptance regions are determined solely by the distributions of the received observations of the check functions. The keyed-in values of the variable to be checked are weighted by the inflation factor before it is put into the check function.

In the Aggregate, Box-Plot, HB and Top-Down methods all values of the check function are sorted by size.
The tails of the distributions are displayed and analyzed in the Aggregate and the Box-Plot methods in order to set the acceptance bounds for the checks. In the HB-Method this setting of the acceptance limits is done automatically.

In the Top-Down and the Box methods the effects of the detected errors on the estimates "determine" how far the manual review should go. In the Top-Down Method the manual review work starts with the extreme values and goes towards the median value, while in the Box Method the records for manual reviewing are selected by the reviewer from a graphical display of the values of the check functions. This selection may be supported by guide-lines (acceptance regions) displayed in the same graph.

The choice of method should be based on the number of variables to be edited by the macro-editing method and on how the staff wishes to work.

## 72 Macro-editing versus micro-editing methods

Macro-editing is not a new concept. It has always been used in data editing, but only as a final checking. Another term is out-put checking. What is new is that such out-put check methods can be used in data editing procedures in the same way as micro-editing methods, and that they have proved to be much more efficient than traditionally applied micro-editing procedures.

Macro-editing methods of the type described in this paper may be considered as a statistical way of providing micro-editing checks with efficient acceptance limits. The limits are based only on the data to be edited. The methods bring a priority thinking to the verifying work i.e. data are edited according to their impact on the estimates. The macro-editing methods solve the general problem inherent in micro-editing methods, i.e. that they produce 100 many error signals without giving any guidance as to how the resources of the verifying work are to be allocated. With micro-editing procedures even very large errors are not always detected, due to the large number of flagged data.

All published studies on the impact of editing show that only a few of the detected errors influence the estimates. For example, in Greenberg et al (1982) it is reported that for all items studied and all domains of study approximately 5 per cent of the cases contributed over 90 per cent of the total change. Many of these large changes were due to reporting in units rather than thousands. A study on the Swedish 1987 Survey on Financial Accounts yielded similar results as did the studies of the editing process in three Australian Bureau of Statistics surveys reported in Linacre \& Trewin.

Thus, where only a few of the errors influence the estimates, the macro-editing methods offer a possibility to reduce the reviewing resources to a considerable extent. Here reported studies indicate that the work is reduced by $35-80$ per cent.

However, there are no limits to the number of cases that can be selected for a manual review. The reviewer can select all the cases he deems necessary. The difference to micro-editing methods is that the cases are selected in a priority order, i.e, according to the impact they may have on the estimates. The selection is mainly done by the reviewer, which means that he governs and has full control of the whole scrutinizing process. In microediting procedures, the reviewer is dominated by the computer and cannot see the effects of his work.

Both procedures focus on randomly appearing negligence errors and utilize the same principle to flag outlier or extreme observations. Micro-editing procedures flag data by criteria fixed in advance, based on historical data, while macro-editing procedures focus on data, which at that very moment and relative to the estimates are the most extreme ones.

Systematic errors, i.e. that many respondents misunderstand a question in the same way or deliberately give wrong answers cannot (in principle) be detected by either macro-editing or micro-editing methods.

It should be noted that neither traditional micro-editing procedures nor any of the procedures discussed here will essentially improve the quality of the survey estimates. It is true that certain (kinds of) errors are removed, but it is not clear whether the estimates are improved. The study of the machine editing of the World Fertility Survey found that the editing did not improve the estimates, but delayed the results by one year (see Pullum et al (1986) or Granquist (1988a))

Below a few references are given to methods which focus on misunderstanding errors. They show that there may be errors even in carefully micro-edited data which have a considerable impact on the estimates. This impact is of greater magnitude than the impact of the "remaining" errors in the simulation studies carried out on the macro-editing methods discussed in this paper.

Werking et al (1988) present the Response Analysis Survey approach, which implies that in an on-going survey, a special survey on how the respondents have answered certain questions is conducted by a self-response touchtone technique. By this technique, errors which cannot be found by traditional editing methods are detected, which reduces the bias in the estimates.

Mazur (1990) presents a method to find so called in-liers in survey data. In repetitive surveys there are respondents which for every period report the same figures also to questions which certainly require the answers to change between periods. They are termed in-liers, because they always lie between the bounds of traditional edits. In her experiments, Mazur has found that these in-liers may cause considerable bias in the estimates.

The conclusion is that we do not know whether any essential improvement is gained by either of the two procedures! But macro-editing certainly is a more efficient way of reaching the same "quality" standard and may release resources for editing the misunderstanding errors.

## REFERENCES

Anderson, K. (1989a). Draft, Output Edit Study, Average Weekly Eamings, Statistical Services Branch, Australian Bureau of Statistics.

Anderson, K. (1989b). Enhancing Clerical Cost-Effectiveness in the Average Weekly Earnings, Draft, Australian Bureau of Statistics, Statistical Services Branch.

Ferguson, D.P. (1989). Review of Methods and Software Used in Data Editing, SCP2/DE/WP. 33 (U.S. Department of Agriculture, National Agricultural Statistics Service).

Granquist, L. (1984a). On the Role of Editing, Statistisk Tidskrift 1984:2.
Granquist, L. (1987). Macroediting - The Top-Down Method, Statistics Sweden, Report 1987-04-09.
Granquist, L. (1988a). On the Need for Generalized Numeric and Imputation Systems, The Seminar on Statistical Methodology, Geneva.

Granquist, L. (1988b). Macroediting - The Aggregate Method, Statistics Sweden, Report 1988-08-18.

Greenberg, B., and Petkunas, T. (1982). 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.

Hidiroglou, M.A., and Berthelot, J.-M. (1986). Statistical Editing and Imputation for Periodic Business Surveys, Survey Methodology, 12, 1, 73-83.

Höglund D.E. (1989). Macroediting - The Hidiroglou-Berthelot Method (Statistical Edits), Statistics Sweden, Report 1989-03-28.

Linacre, S.J., and Trewin, DJ. Evaluation of Errors and Appropriate Resource Allocation in Economic Collections, Undated, Internal paper, Australian Bureau of Statistics.

Lindblom, A. (1990). A review of the macro-editing procedure Top-Down, Data Editing Joint Group Product NR SCP2/D.12/f.

Lindström, K. (1990). A macroediting method application developed in PC-SAS, Data Editing Joint Group Product NR SCP2/D.11/.

Mazur, C., (1990). A Statistical Edit for Livestock Slaughter Data, USDA-NASS, Washington DC 20250.
Pierzchala, M. (1988). A Review of the State of the Art in Automated Data Editing and Imputation, Staff report for National Agricultural Statistics Service, U.S.D.A. No. SRB-88-10.

Pullum, T.W., Harpham, T., and Ozsever, N. (1986). The Machine Editing of Large Sample Surveys: The Experience of the World Ferility Survey, International Statistical Review, 54, 3.

Tukey, J.W. (1977). Exploratory Data Analysis, Addison-Wesley Publishing Company.
Werking G., Tupek A., and Clayton R. (1988). CATI and Touchtone Self-Response Applications for Establishment Surveys, Journal of Official Statistics, 1988-4.

# ON-LINE CAPABILITIES IN SPEER <br> (Structured Programs for Economic Editing and Referrals) 

L. Draper, B. Greenberg and T. Petkunas'


#### Abstract

The SPEER System is a multipurpose edit and imputation system developed at the Census Bureau for use with continuous data under ratio edits. SPEER has been employed to edit the Auxiliary Establishment Report and the Enterprise Summary Report of the 1982 and 1987 Economic Censuses. It was also used to edit the 1987 Census of Construction Industries and an adaptation of SPEER was used to edit the 1987 Censuses of Manufactures and subsequent Annual Survey of Manufactures. This report discusses the on-line, interactive capabilities of the SPEER System. These features have been employed in the analyst review of edit referral cases and as a data entry device for responses to mail questionnaires.


KEY WORDS: Edit; Imputation; SPEER.

## 1. INTRODUCTION

All survey data must be edited to detect improbable response combinations on a questionnaire, to make changes to keyed reported data, and to impute for missing items. The objective is to detect and correct errors caused by misunderstanding a survey question, faulty reporting, or problems in data entry. The flow of a data record through all stages of error detection and correction is a combination of automated procedures, manual review, and an interactive combination of the two. We describe the sequence of activities at the Census Bureau for the typical large-scale economic establishment survey or census.

The data collection instrument is a printed questionnaire mailed to the survey universe to be returned to the Census Bureau Jeffersonville, Indiana facility to be keyed. During data entry there are rudimentary checks to detect data entry errors. Keyed data are sent to headquarters in Suitland, Maryland, where they are run through an automated batch edit program to detect inconsistencies, make changes, and impute for missing responses.

Within the automated edit and imputation routines, selected records are targeted as referral cases for analyst review. Criteria typically are: (1) large change to reported data, (2) imputations for large establishments, and (3) unsuccessful imputation of a value that will pass tolerance checks. During the review process, an analyst will have the respondent's questionnaire, will be able to call respondents by telephone, and will have the use of alternative data sources to determine a reasonable impute for nonresponse or to adjust an assumed erroneous field value. The cycle of automated routines followed by analyst review involves (1) processing records at headquarters, (2) sending referral listings to Jeffersonville, (3) hand corrections to referral documents, (4) keying of corrections, (5) sending corrections back to headquarters, and (6) a subsequent cycle of processing at headquarters. There are ample opportunities for delays and new errors, and these multiply as the cycles grow.

The need for an on-line, interactive analyst review capability has been evident for some time. The objective is for an analyst to key corrections directly onto the data record during the review process and have changes edited as they are entered. Although SPEER was originally designed for batch processing, the system has evolved to

[^35]include the capability to perform interactive review of referral records and interactive data entry. The system is sufficiently flexible to accommodate a wide variety of user-specific, user-specified requirements.

In Section 2, we describe SPEER capabilities, structure, and basic methodology. This paper does not go into depth in these areas and the reader is referred to Greenberg ( $1981,1982,1987$ ) and Greenberg and Surdi (1984) for more detail. In Greenberg and Petkunas (1990) we describe actual implementations of SPEER and discuss how working with users led to the evolution of the SPEER system. That paper is, in some sense, a companion to this one. In Section 3, we describe the interactive capabilities in SPEER and discuss how they have been employed for edit referrals and data entry. Two examples are presented to provide a sense of system flexibility and options.

The batch versions of SPEER and the on-line, interactive versions have the same underlying program. The interactive version of the program pauses at selected points in the code to wait for input from the keyboard as it is in conversational mode with the user. The system is menu-driven and allows a user to interact with all SPEER subroutines. The exact interactive capabilities as well as screen displays shown here were dictated by subject specialists in Economic Surveys Division and Industry Division, with whom we worked closely throughout.

## 2. METHODOLOGY IN SPEER

SPEER is a multipurpose edit and imputation system for numeric data under ratio edits. An establishment record consists of a vector with numeric data fields and a ratio edit is the requirement that the quotient of two fields lies between prescribed bounds which are read into the system as parameters. A typical ratio edit is of the form $L_{i j} \leq x_{i} / x_{j} \leq U_{i j}$, where $L_{i j}$ and $U_{i j}$ are the lower and upper allowed limits for the ratio of $x_{i}$ to $x_{j}$. For example, the ratio of the annual total salaries paid to construction workers divided by annual total hours worked by construction workers must be within reasonable limits. If $L_{12} \leq x_{1} / x_{2} \leq U_{12}$ and $L_{23} \leq x_{2} / x_{3} \leq U_{23}$ are two ratio edits, the implied edit is $L_{12} L_{23} \leq x_{1} / x_{3} \leq U_{12} U_{23}$. Starting with a set of user supplied explicit edits an edit generation subroutine first derives all implied edits.

Edit checking subroutines determine which edits (from the full set of implied edits) pass or fail for a given record. If all edits pass and no data values are missing, the record is considered acceptable. In addition to the use of current reported data for edit checking, data can also be checked against prior year data or against other comparable data values -- see Section 3. If no edits fail but some data items are missing the record is sent for imputation of the missing fields. If one or more edits are failed by a record, the record is sent to error localization to determine a set of fields to delete so that the remaining fields will be mutually consistent. The objective is typically to delete a weighted minimal set of fields.

Before imputing for a missing field (either reported and deleted or not reported at all), the system derives a region for imputation. Suppose there are $n$ fields on a record and (after re-ordering if necessary) fields $x_{1}, \ldots, x_{k}$ for $k \leq n$ were reported and not targeted for change--they are mutually consistent. In particular, for all $i, j \leq k$, the ratios $L_{i j} \leq x_{i} \mid x_{j} \leq U_{i j}$, are satisfied. If $k=n$ the record is consistent.

If $k<n$, we establish an imputation range for $x_{k+1}$. Note that for all $j<k$, we have the ratio $L_{k+1, j} \leq x_{k+1} / x_{j} \leq U_{k+1, j}$ and by multiplying through by $x_{j}$ we also have the pair of inequalities $x_{j} L_{k+1, j} \leq x_{k+1} \leq x_{j} U_{k+1, j}$ where $x_{j} L_{k+1, j}$ and $U_{k+1, j}$ are known. Each $j=1, \ldots, k$ determines an interval in which $x_{k+1}$ must reside to be consistent with $x_{j}$. If $x_{k+1}$ lies in the intersection of the 1 intervals, it will be consistent with each of the fields $x_{j}$. If the edits are consistent, the intersection is not empty, and it is referred to as the feasible region for field $x_{k+1}$

For each field we have an imputation module which contains a sequence of imputation rules provided by survey staff. When a field is selected for imputation, the feasible region for the missing field is computed and the program reaches into the imputation module to obtain a feasible imputation. The SPEER variables are typically divided into basic and secondary items. Basic variables are those fundamental to the operations of an establishment and these are edited jointly in a core program as described above. The secondary items are
grouped into satellites of related items which are edited against one another. First the basic items are edited against each other and then the satellite items are made to conform to the basic items as well as each other within the same satellite.

## 3. ON-LINE, INTERACTIVE SPEER

### 3.1 Review of Refertal Cases

The Enterprise Statistics program consists of a series of publications at the company level. Two enterprise reports include: the Large Companies publication and the Auxiliary Establishments publication.

The Large Companies publication is based on responses to questionnaires sent to companies with 500 or more employees and published tables show selected financial statistics of large companies. The Auxiliary Establishments publication presents data on auxiliary units of multi-establishment firms. Auxiliary establishments support activities of the other components of a company. Examples are research and development centers, warehouses, and administrative offices. The Enterprise Summary Report form (ES-9100) collects data for the Large Companies publication and the Auxiliary Establishment Report form (ES-9200) collects data for the Auxiliary Establishments publication.

The first implementations of SPEER were for the 1982 Enterprise Summary Report and 1982 Auxiliary Establishment Report. SPEER was also used to edit these two report forms for the 1987 Economic Censuses and we took advantage of the newly developed on-line, interactive capabilities for review of referral cases.

When using the interactive version of SPEER to review a referral record, the analyst will access a record and indicate which field he/she wishes to examine and possibly revise. If a second field is to be reviewed on this record, the program displays information as it did for the first field. Values for the second field are based, in part, on the new value of the revised first field. After all fields have been examined and adjusted, the review is complete and no further batch processing is required. The analyst converses with the automated program in order to augment the system expertise and override batch processing rules. Screen design and variations in the system were based on requests from survey staff.

When working referrals for the ES-9100, a lead-in screen prompts the user for a record request. Requests include (1) access next record, (2) access specific record, (3) access next unresolved case, (4) return to first record in the file, (5) bring back current record, (6) enter a Company record, (7) enter an Auxiliary Establishment, and (8) exit SPEER.

Screen One covers the basic items on the ES-9100 report form and satellite items are treated on subsequent screens. There are a total of 3 screens, 9 basic items, and 46 satellite items in the ES-9100 edit programs. The header contains the Census File Number, name of establishment, 1982 category code, 1987 category code, and microfilm reference number.

Figure 1: Display for ES-9100 Screen One

| 9999999901 | The American Weigh |  | CAT82:999A | CAT8 | 7:9999 | 99991 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Mnem | Current | Reported | ST | Low |  | Upper |
| EMP | 1,000 | 1,000 | R |  | 71 | 2963 |
| APR | 32,000 | 32,000 | R | 17,0 |  | 32947 |
| QPR | 7,111 | 3,000 | X | 4,2 |  | 11250 |
| FBR | 2,843 | 0 |  | 1,5 |  | 5525 |
| SLS | 120,000 | 120,000 | R | 74,8 |  | 180415 |
| AET | 66,945 | 100,000 | X | 65,63 |  | 69804 |
| TOT | 60,000 | 60,000 | R | 57,5 |  | 203031 |
| RPT | 1,200 | 0 | I |  | 64 | 2160 |
| ADE | 50,500 | 50,500 | R | 14,202 |  | 51510 |
| Action taken: | Mult: 1.0 | Analyst: TFP |  | 2/14 |  | Rank:17 |
| Flags: AETDET | AETDIMP | TOTDET | TOTDIMP | ABTDET | ABTDIMP | P FGCET |
| ACTIONS: | ccept 1. | 1. Delete | 2.Run SPE |  | 3.Restore | reported |
|  | pute 6.R | Restore com | plx 7.Ne | screen | 8.Return |  |
|  | w reported | C.View | complex | M.Change | mult |  |

The first column displays the mnemonics for each basic item: Number of Employees (EMP), Annual Payroll (APR), First Quarter Payroll (QPR), Fringe Benefits Required (FBR), Sales (SLS), Total Ending Assets (AET), Total Assets (TOT), Total Rental Payments (RPT), and Accumulated Depreciation for the End of Year (ADE). The next two columns display data values for each of the basic items with all dollar values in thousands. Column two shows data values as they appear after the batch edit and column three shows reported values for each basic item. For those items whose reported value differs from the current value, the current value is highlighted.

The fourth column displays the current status flag for each basic item: reported greater than zero and passed edits (R), reported greater than zero and changed by SPEER (X), reported greater than zero and set equal to zero ( $Z$ ), imputed to a positive value from zero (I), and nonresponse set to zero ( $N$ ). The final two columns display the lower and upper limits of each basic item's feasible region. The first line following the data displays the record multiplier, analyst identification, date, and company rank. Typically, companies with a higher rank will be given more attention from the analyst.

Record multipliers allow a user to expand the feasible region to permit acceptance of data values outside the usual range. One can think of ratio edits to be of the form $(1 / m) L_{i j} \leq x_{i} / x_{j} \leq m U_{i j}$ where $m \geq I$ is the multiplier. When $m=1$, we have the initial explicit ratios and when $m$ is greater than one the lower and upper ratio limits are extended outward. Accordingly, the lower limit of the feasible region is divided by $m$ and the upper limit is multiplied by $m$, thus increasing the size of the region.

Flags describe changes to the entire record to allow an analyst to see a snapshot of the whole record without scrolling through screens. For example, the flag FGCET indicates that Capital Expenditures, which is on Screen Two, has been changed by a substantial amount. The definition of "substantial amount" is a user-defined parameter. The menu contains actions designed by the subject-matter specialists.

0 . Accept: This option indicates the present status of the record is acceptable--directly after the batch run or after analyst changes.

1. Delete: This option is designed to remove a record from the database.
2. Run SPEER: Invoking SPEER allows the analyst to immediately see how changes made will affect the rest of the record.
3. Restore reported: Reinstates the originally reported data with one key stroke.
4. Impute: Blanks out all values to allow the analyst to impute an entire record from only a few specified data values. An analyst will use this option to impute from administrative records.
5. Restore complex: Reinstates values as they were at start of session.
6. Next Screen: Displays the next screen which contains other data items, typically satellite items.
7. Return: Returns record to database to be reviewed further at a later time.
8. View reported: Displays all originally reported values on one screen to give analyst a picture of entire establishment without paging through all screens.
C. View complex: Displays all current edit values on one screen.
M. Change mult: Allows analyst to change the multiplier for this record to override the multiplier currently in place.

Actions 1, 3,5, and 6 have safeguards and it takes two keystrokes to invoke them. After selecting one of these options, warnings are provided to guard against an analyst overwriting the current data by mistake or deleting a record from the universe by accident.

The SPEER programs can generate a large quantity of diagnostic information and the choice of diagnostics displayed on the screen is an option given survey staff. In addition to a large amount of diagnostic information at time of record review, information is also available to monitor the review process. One can monitor how often analysts employ a multiplier, accept the automated system actions, override the batch programs, and so on. These capabilities can help monitor the activities of analysts and evaluate their performance, and program managers can use this information to understand and perhaps improve this highly subjective process.

If an analyst value is not consistent with other basic items, a bell sounds, the menu disappears, and a message is displayed. For example, if an analyst entered $\$ 72,000$ for AET in Figure 1, the following message would appear: " $L_{0}=65,632, \mathrm{AET}=72,000, \mathrm{Up}=69,804 ; 72,000$ is not within current bounds. Should this value be $\operatorname{accepted}(\mathrm{Y} / \mathrm{N})$ ? $==>{ }^{\text {" }}$.

If the analyst wants this value accepted, he answers YES and a new message inquires about adjusting the record multiplier. Note that a new multiplier is displayed and bounds for each basic item have been adjusted by multiplying the upper bounds by the new multiplier and dividing the lower bounds by it.

Figure 2: Effects of Recalculating Multiplier (abbreviated display)

| EMP | 1,000 | 1,000 | R | 898 | 3,204 |
| :--- | ---: | ---: | :--- | ---: | ---: |
| APR | 32,000 | 32,000 | R | 15,781 | 35,631 |
| QPR | 7,111 | 3,000 | X | 3,945 | 12,166 |
| FBR | 2,843 | 0 |  | 1,429 | 5,965 |
| SLS | 120,000 | 12,000 | R | 74,454 | 195,112 |
| AET | 72,000 | 100,000 | X | 60,689 | 75,490 |
| TOT | 60,000 | 60,000 | R | 57,226 | 219,570 |
| RPT | 1,200 | 0 | I | 59 | 2,336 |
| ADE | 50,500 | 50500 | R | 14,124 | 58,085 |

Action taken: Mult: 1.1 Analyst: TFP 2/14/88 Rank: 17
Enterprise information is also obtained by aggregating establishment information reported in other segments of the economic censuses. For all establishments belonging to a single enterprise, reported data are collected from other sources including the censuses of Wholesale Trade, Retail Trade, Service Industries, Manufactures,

Minerals Industries, Construction Industries, and selected Transportation Industries. Data from these diverse sources are added together to form a Summary Record for the given enterprise. This summary information is employed to edit and impute enterprise information collected on the ES-9100. To use the summary information to impute for missing or erroneous ES-9100 data values, one selects Option 4 on the menu below as for item FBR.

Figure 3: ES-9100 Screen One with Summary (abbreviated screen)

| 9999999901 | 1 The A | can Weigh | CAT82:999A |  | 7:9999 |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Mnem | Current | Summary | Reported | ST | Lower | Upper |
| EMP | 1,000 | 987 | 1,000 | R | 971 | 2,963 |
| APR | 32,000 | 30,117 | 32,000 | R | 17,066 | 32,947 |
| QPR | 7,111 | 7,111 | 3,000 | X | 4,267 | 11,250 |
| FBR | 2,843 | 2,843 | 0 |  | 1,545 | 5,525 |
| SLS | 120,000 | 101,056 | 120,000 | R | 74,866 | 180,415 |
| AET | 66,945 | 66,945 | 100,000 | X | 65,632 | 69,804 |
| TOT | 60,000 | 0 | 60,000 | R | 57,543 | 203,031 |
| RPT | 1,200 | 1,200 | 0 | I | 64 | 2,160 |
| ADE | 50,500 | 0 | 50,500 | R | 14,202 | 51,510 |
| ACTIONS: 0.Accept 1.Delete |  |  | 2.Run SPEER 3.Restore reported |  |  |  |
| 4.Sum replace 5.Impute 6.Restore complx 7 .Next screen |  |  |  |  |  |  |
| 8.Return 9.View reported C.View complex M. Change mult |  |  |  |  |  |  |

SPEER is written in FORTRAN and is easy to transfer from one operating system to another and programs were adapted to microcomputers with no difficulty. The batch versions of SPEER for the ES-9100 and ES-9200 questionnaires were run on the UNISYS operating system primarily due to communication lines established between Suitland and Jeffersonville. After records were run in batch mode on the UNISYS mainframe, referral cases were down-loaded and analysts performed their review of referral cases using IBM microcomputers connected through the local area network sharing a single database.

### 3.2 Data Entry

The Annual Survey of Manufactures (ASM) provides intercensal yearly estimates of key measures of manufacturing activity. Data for these key measures, as well as other detailed statistics for manufacturing, are collected in the Censuses of Manufactures. An annual survey has been taken each year between censuses starting with 1949 to provide a continuous series of basic statistics and furnish benchmarks for current business indicators and for measures of industrial production and productivity. Industry Division has designed an edit based on SPEER which has been used for the 1987 Census of Manufactures and each Annual Survey of Manufactures since 1986.

The Annual Survey of Manufactures consists of a sample of manufacturing establishments taken from the Census of Manufactures. Sample cases are in the survey for four consecutive intercensal years and are sent questionnaires to be returned by mail. Response forms received after data entry facilities are closed at Jeffersonville are referred to as late adds and are entered into the database by analysts at headquarters. Typically, 2,000 to 2,500 late adds are received yearly for the ASM, and up to 15,000 for the Census of Manufactures. In the past, late adds entered into the database were not run through the automated batch edit programs. In some cases, large establishment reports have been included in the database without machine edit.

Staff responsible for the Annual Survey of Manufactures requested an interactive version of SPEER for data entry for late adds. Using this system, data are edited at the time of entry with no need for batch editing. Programs are menu driven and follow the basic SPEER structure with specialized screens and options designed according to the needs of the Annual Survey of Manufactures staff.

We describe the interactive data entry system for late adds for the 1988 Annual Survey of Manufactures. This system was run on microcomputers and data entered through this system were later up-loaded to the UNISYS
mainframe. The program begins with a prompt for the analyst's initials. Edit parameter values are loaded for 529 Standard Industrial Classification (SIC) codes consisting of a lower bound, upper bound, and the industry average for each explicit ratio. The parameter values contain a set of current and prior-year edit limits which differ by SIC. Both edit and imputation rules make use of prior-year information when it is available. Items are edited against current-year and prior-year information and each item must pass both sets of edits to be considered acceptable.

Prior-year data are used to edit current reported values through a year-to-year ratio edit. If $x_{i}, x_{p} y_{i}$ and $y_{j}$ are current-year and prior-year data values, respectively, a year-to-year ratio edit is $l_{i j} \leq\left[x_{i} / x_{j}\right] /\left[y_{i} / y_{j}\right] \leq u_{i j}$ where $l_{i}$ and $u_{i j}$ are the lower and upper bounds for this ratio. See Greenberg (1981) for a discussion of year-to-year ratio edits. Feasible regions are calculated based on current-year values and based on prior-year values. The intersection of these two becomes the final feasible region. If prior-year information is not to be used for a particular record, the program operates using current-year values.

The first screen is for data entry of the core items. Initially, a 10 -digit permanent plant number (PPN) is entered, followed by a 6 -digit SIC code which must match one of 529 SIC's contained in a separate list. If the PPN is not 10 digits or the SIC does not match one on the list, appropriate error messages are displayed and the user is asked to re-enter the information. When a valid SIC is entered, the current and prior-year parameters are loaded for that SIC through a direct access file. In this manner, the program's working storage is minimized.

Figure 4: Interactive Data Entry Screen for ASM

| PPN: 1234567890 |  |  |  |  | SIC: 201112 |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Mnem | Reported | Prior | ST | Lower | Upper |
| SW | 1,000 | 21,000 | R | 600 | 3,076 |
| VS | 25,000 | 30,000 | R | 11,029 | 54,546 |
| OW | 600 | 700 | R | 35 | 1,000 |
| WW | 500 | 1,300 | R | 533 | 1,000 |
| PW | 50 | 100 | R | 46 | 70 |
| PH | 125 | 175 | R | 53 | 117 |
| TIE | 2,500 | 1,700 | R | 0 | 5,666 |
| TE | -1 | -1 | N | 50 | 200 |
| TIB | 1,700 | -1 | R | 250 | 4,999,995 |
| LE | -1 | -1 | N | 10 | 243 |
| OE | -1 | -1 | N | 10 | 120 |
| CM | -1 | -1 | N | 1,250 | 15,000 |
| VP | -1 | -1 | N | 10 | 400 |
| Mult: 1.0 |  | itor: Lisa |  | Date: D | C 3, 1990 |

The first column displays the mnemonic for each basic item. They are: Salary and wages (SW), Value of shipments (VS), Other employee wages (OE) (non-production workers), Production worker wages (WW), Number of production workers (PW), Plant hours worked by production workers (PH), Total inventory end of year (TIE), Total employment (TE), Total inventory beginning of year (TIB), Legally required expenditures (LE) (such as social security, etc.), Other employees (OE) (non-production workers), Cost of materials (CM), and Voluntary payments (VP) (medical benefits, retirement, etc.).

The second column is the value keyed from the report form. The next column is prior-year information which the analyst may optionally key in. The fourth column, labeled ST, refers to the status flag. The status flag indicates the source of a number. Lastly, each field's feasible region is displayed. For a record to be consistent, each field must lie within its respective bounds. The value "-1" is a place-bolder for a blank field.

An analyst entering a record can observe bounds changing on the screen as each item is keyed. New lower and upper bounds of the feasible region are computed as each new piece of information is entered. In effect, each current-year field is being evaluated for consistency with other items as it is entered-but no changes are made at this time.

After all information on the form has been keyed, data are run through the SPEER edit program and a new screen is displayed. The reported values are shown as well as the edited (current) fields and prior-year values. The SIC is split into an Industry and Sub-industry code, the record number is displayed next to the date, and a menu appears at the bottom of the screen with available actions. An analyst can make subsequent adjustments to the record or accept it in its current state. The options are quite similar to those for the ES-9100 forms.

Figure 5: Data Edit Screen One for ASM Interactive Programs

| PPN: 12 | 567890 |  |  |  | IND: 2011 | SUB: 12 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Mnem | Current | Prior | Reported | ST | Lower | Upper |
| sw | 1,100 | 2,000 | 1,000 | X | 600 | 3,076 |
| Vs | 25,000 | 30,000 | 25,000 | R | 14,583 | 60,000 |
| OW | 600 | 700 | 600 | R | 115 | 1,100 |
| ww | 500 | 1,300 | 500 | R | 375 | 1,100 |
| PW | 50 | 100 | 50 | R | 33 | 70 |
| PH | 88 | 175 | 125 | X | 53 | 117 |
| TIE | 2,500 | 1,700 | 2,500 | R | 0 | 5,666 |
| TE | 73 | -1 | -1 | I | 50 | 220 |
| TIB | 1,700 | -1 | 1,700 | R | 250 | 2,299,997 |
| LE | 121 | -1 | -1 | I | 11 | 267 |
| OE | 23 | -1 | -1 | 1 | 10 | 73 |
| CM | 8,750 | -1 | -1 | 1 | 1,250 | 15,000 |
| VP | 101 | -1 | -1 | 1 | 11 | 440 |
| LC | 222 |  |  |  | N/A | N/A |
| Action ta ACTION | : Mult: 1.0 |  | Editor: Lisa | Date: DEC 3, 1990 |  | 1 |
|  | 1.Accept record 5.Change status flag |  | 2.Run thru SPEER 6.Next screen 9.Print record |  | 3. Set mult | tiplier |
|  |  |  | 7. Print aud | t trail |
|  | 8.Start record over |  |  |  |  |  |

This interactive version of SPEER has proven very successful for data entry of late adds for the 1988 Annual Survey of Manufactures. Because of success of the programs described above, we were asked to develop an online version of these programs on the VAX which will directly access the ASM database. We are now working on Version 2.0 which currently consists of 3 data screens: 2 detail screens and a core item screen. These screens contain basically the same information as in Version 1.0 , but have more detail items and different menu options.

Version 2.0 will access the database directly, eliminating the need for the analyst to key prior-year data since this information will be available from the database. The user will key a tab number upon which the record will be matched from the database. This matched record will be displayed on the screen. An internal check will be performed to determine whether the edit should include prior-year data based on the ratio of current-year salary and wages and prior-year salary and wages. If this ratio is not within an acceptable range, prior-year data will not be included in the edit. In such a case, or if no record in the database is matched with the keyed tab number, the system will edit using current-year data only.

If a record passes all current-year edits but fails some prior-year edits, the prior-year edit constraints will not be enforced either in editing or in imputation. If a record passes all prior-year edits but fails some current-year edits, current-year internal consistency constraints will not be enforced for either edit or imputation. In the first case we allow an internally consistent record to differ year-to-year and in the second instance we make allowance for a typical current year performance. In either case we allow more reported data to pass edits and attempt
to avoid over-editing. Another new feature in Version 2.0 is that after error localization, the program pauses before deleting fields and highlights candidate fields to delete. This allows the analyst to select fields to change without using the automated error localization subroutines. Of course, automated error localization is available.

Any discussion of Version 2.0 at this stage must be very incomplete as programs are currently being developed. This version of the on-line, interactive SPEER programs will be used to enter and edit late adds for the Annual Survey of Manufactures in 1989 and later, as well as for the 1992 Census of Manufactures.

## 4. CONCLUDING REMARKS

Work started in 1980 on what evolved into the SPEER system. The original objective was to design edit programs for numeric data under ratio edits which incorporated the advances in methodology made by Fellegi and Holt (1976) and by Gordon Sande $(1976,1981)$ at Statistics Canada. We worked very closely with the staff in Industry Division to design an Annual Survey of Manufactures prototype and had no intention of developing a multi-user system; however, the SPEER programs evolved from this work.

The current system can be thought of as a shell which we customize for individual survey-staff users. Each time we implemented this system enhancements were introduced into the core programs. We did not wait until we had a full-blown system with all desirable features before venturing to use it and it can be thought of as under continual development. In earlier sections are examples to indicate system flexibility and we discussed use of summary information and prior-year data to highlight system adaptability. The direction for changes to SPEER stems from needs and requests of users; and desire for on-line, interactive programs have dictated the work described in this report.

## ACKNOWLEDGMENT

We take pleasure in highlighting the work of John Monaco and Cindy Schott in the Economic Surveys Division for their design of the edit and imputation programs for the ES-9100 and ES-9200. John and Cindy directed the work by describing system requirements, outlining needs and requesting capabilities. Their efforts run throughout these programs. Bob Rosati and Rich Sterner in Industry Division were responsible for the design of the ASM data entry programs for 1988 and they are directing the work on the 1989 ASM program in Version 2.0. Once again, their efforts are part of every aspect of these programs.

## REFERENCES

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

Greenberg, B. (1981). Developing an Edit System for Industry Statistics, Compuser Science and Statistics: Proceedings of the 13th Symposium on the Interface, Springer-Verlag, New York, 11-16.

Greenberg, B. (1982). Using an Edit System to Develop Editing Specifications, Proceedings of the Section on Survey Research Methods, Joumal of the American Statistical Association, 366-371.

Greenberg, B. (1987). Edit and Imputation as an Expert System, in Statistical Policy Working Paper Number 14: Statistical Uses of Microcomputers in Federal Agencies, Statistical Policy Office, Office of Information and Regulatory Affairs, Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget, Washington, D.C., 85-92.

Greenberg, B., and Petkunas, T. (1990). SPEER (Structured Programs for Economic Editing and Referrals), Proceedings of the Section on Survey Research Methods, American Statistical Association, to appear.

# M-ESTIMATORS AND OUTLIER RESISTANT ALTERNATIVES TO THE RATIO ESTIMATOR 

L.-P. Rivest and E. Rouillard'


#### Abstract

Outlier resistant alternatives to the ratio estimator are proposed. They are constructed using robust Mestimators of the slope of the regression model through the origin which underlies the ratio estimator. Resistant estimators downweight sample units which are outliers with respect to the regression model. Several resistant estimators are compared to the ratio estimator in a Monte Carlo study. One resistant estimator is selected for further studies. Its properties, with respect to the sampling plan, are reviewed. It is not consistent; expressions are given for its large sample bias and its large sample variance. Estimators for the mean squared error of the outlier resistant estimator are proposed. The coverages of outlier resistant confidence intervals built using these mean squared error estimators are investigated by Monte Carlo. The performances of the ratio estimator and of its outlier resistant competitor are also compared conditionally, given a fixed number of outliers in the sample.


KEY WORDS: Conditional analysis; Confidence intervals; M-estimators; Monte Carlo simulations; Regression; Sample design.

## 1. INTRODUCTION

Editing is an important step in the processing of survey data. It ensures the quality of the data and makes survey estimates representative of the populations under study. From a statistical point of view, an important role of editing is to investigate unusual responses, i.e. responses that are markedly different from what was expected for a particular sample unit. We will call these unusual responses outliers. In some instances outliers are caused by clerical errors which can be corrected in the editing process. The unusual response is then replaced by the correct one. Such outliers are called unrepresentative. In other instances outliers correspond to the true states of sample units. For example, population units experiencing sudden changes over short periods of time can become outliers. Such units are called representative outliers. Representative outliers are components of the populations under study, they should be accounted for in the estimation of population characteristics of interest. In other words, part of the sampling variance of estimated characteristics should be related to the treatment given to representative outliers. Replacing, during editing processes, representative outliers by imputed values and treating, in estimation procedures, these values as true responses leads to underestimations of the sampling variances of survey estimators.

Statistical methods for outliers in survey data generally have two steps. First come detection of outliers according to some data dependent criteria. Once they are detected, outliers are given special treatment. Usually, one reduces their sampling weights or the magnitude of their $y$-values. Winzorization, that is replacing a $y$-value by a fixed or a data dependent constant, is often used. These methods have been investigated by several authors (Searl, 1966, Fuller, 1970, Rao, 1971, Ernest, 1980, Tambay, 1988). These authors show that outlier treatment can yield interesting reductions in mean squared errors of survey estimators.

This paper considers estimators incorporating special procedures for outliers when auxiliary information is used in the estimation of population characteristics. Alternatives to the ratio estimators are investigated. The proposed estimators are based on the M-estimators of "Robust Statistics". Such estimators have been considered

[^36]in the context of survey sampling by Chambers (1986), Gwet and Rivest (1990) and Lee (1990). The outhier resistant estimators are constructed in Section 2. Monte Carlo comparisons with the ratio estimator are presented and one outlier resistant estimator is selected for further investigations. Section 3 considers to problem of estimating the mean squared error of the resistant estimator and of constructing outlier resistant confidence intervals for the unknown population mean. Section 4 presents a conditional investigation, given the percentage of outliers in the sample of the properties of the resistant estimator and of the ratio estimator.

## 2. CONSTRUCTION OF OUTLIER RESISTANT ESTIMATORS

A strategy for constructing outlier resistant estimators is suggested. The usual set up for the ratio estimator is first reviewed. A population $U$ of size $N$ is sampled according to a without replacement simple random sampling plan. The size of $s$, the sample, is $n$. The aim of the investigation is to estimate $\overline{\boldsymbol{Y}}$, the population mean of variable $\boldsymbol{Y}$, using the data in the sample and the auxiliary variable $X$ which is known for all the units in $U$. Often probability model ( $M$ ), which is given by

$$
Y_{i}=\beta X_{i}+\varepsilon_{i}
$$

where $E\left(\varepsilon_{i}\right)=0$ and $V\left(\varepsilon_{i}\right)=\sigma^{2} X_{i}$, provides a good description of the relationship between $\boldsymbol{Y}$ and $\boldsymbol{X}$. Then the ratio estimator, $\bar{y}_{r}=\bar{X} \bar{y} / \bar{x}$, where $\bar{y}$ and $\bar{x}$ are the sample means of the $y$ and of the $x$ variable respectively, is a good estimator of $\bar{Y}$. Estimator $\bar{y}_{r}$ can be described as $\bar{X}$ times the least squares estimator of $\beta$ under model $(M)$. This suggest to construct outlier resistant alternatives to $\bar{y}_{r}$ as

$$
\begin{equation*}
\bar{y}_{g}=\bar{x} \hat{\beta}_{g} \tag{1}
\end{equation*}
$$

where $\hat{\beta}_{g}$ are outlier resistant estimators of parameter $\beta$ of model ( $M$ ).
Outlier resistant estimators of the slope of model ( $M$ ) are usually defined by implicit equations. Several proposals have been made. They all involve odd increasing bounded functions, labelled $\boldsymbol{\psi}$, and robustness constants, called $k$. Popular choices of $\psi$ are Huber function which is given by

$$
\psi_{k}(t)=\operatorname{sgn}(t) \min (|t|, 1),
$$

where $\operatorname{sgn}($.$) denotes the sign function, modified Huber function which is equal to$

$$
\psi_{A_{m}(t)}=\left\{\begin{array}{l}
\sin \left(\frac{\pi}{2} t\right) \text { for }|t| \leq 1 \\
\operatorname{sgn}(t) \text { for }|t|>1),
\end{array}\right.
$$

where $\sin ($.$) denotes the sine function, and function fair which is equal to$

$$
\psi_{f}(t)=\frac{8}{1+|t|} .
$$

These three functions are compared by Rey (1983), section 6.4.10. They have similar shapes and they yield estimators with similar properties. One advantage of $\psi_{H m}(t)$ and $\Psi_{f}(t)$ over $\psi_{H}(t)$ is that they have continuous first derivatives. Three types of robust estimators for parameter $\beta$ of model $(\mathbf{M})$ are considered. First come $M$-estimators defined as the solution of

$$
\sum_{z} \sqrt{x_{i}} \psi\left(\frac{y_{i}-\beta x_{i}}{k \sqrt{x_{i}}}\right)=0 .
$$

where $k$ is a robustness constant. With $M$-estimators, outliers with large $x$-values still have an important impact
on the estimator. This had led researchers to construct GM-estimators of regression parameters reducing further the impact of units with large values of $\boldsymbol{x}_{\boldsymbol{i}}$. These estimators are of two types (see Hampel, Ronchetti, Rousseeuw, and Stahel, 1986). The Hampel-Krasker class is defined as the solution of

$$
\sum_{s} \neq\left(\frac{y_{i}-\beta x_{i}}{k}\right)=0,
$$

and Mallows class of estimators is obtained by solving

$$
\sum_{s} \psi_{1}\left(\frac{\sqrt{x_{i}}}{k_{1}}\right) \phi_{2}\left(\frac{y_{1}-\beta x_{i}}{k_{2} \sqrt{x_{i}}}\right)=0 .
$$

The least squares estimator $\bar{y} / \bar{x}$ belongs to each of these three classes. It can be obtained by taking arbitrarily large robustness constant $k$, or $k_{1}$ and $k_{2}$ or by taking $\psi(t)$, or $\psi_{1}(t)$ and $\psi_{2}(t)$, equal to $t$.

### 2.1 Description of the Monte Carlo experiment

The performance of three estimators, one for each of the classes defined in the previous section are compared using Monte Carlo simulations. Five populations are used in the study, the first four contain 500 units; they were constructed using the procedure suggested by Robinson (1987). Population 1 is constructed according to model (M). The regression of Population 2 has a quadratic term. Population 3 is similar to Population I with $7 \%$ of random outliers. In Population 4, the slope of the regression is, for most of the data point, equal to 2; it is 1 for $7 \%$ of the units. Population 5 contains 235 units. It was obtained from the block population of Kish (1965) by discarding all the units with $x \leq 2$ and by perturbing the $y$-values of 13 units to create outliers. Graphs of the five populations are given in Appendix 2.

Table 1: Estimators in the Monte Carlo study

| Type of estimator | Label | $\psi$ function | Robustness constant |
| :---: | :---: | :---: | :---: |
| Ratio estimator | RE | $\psi(t)=1$ | not appllcable |
| M estimator | ME | $\psi(t)=\psi_{\text {m }}(t)$ | $k=1.2107$ 。 |
| GM estimator | GMM | $\psi_{1}(t)=\psi_{H 1}(t)$ | $k_{1}=1.46\left(\operatorname{med}\left(\sqrt{X_{t}}\right)\right.$ |
| Mallows type |  | $\psi_{2}(t)=\psi_{H_{m}}(t)$ | $k_{2}=1.2107 \sigma$ |
| GM estimator <br> Hampel Krasker type | GMH | $\psi(t)=\psi_{f}(t)$ | $k=\frac{\operatorname{med}\left(\sqrt{X_{i}}\right)}{.6745} 2.560$ |

The general form of the three estimators of $\bar{Y}$ under investigation is given by formula (1); the definitions of the estimators $\hat{\beta}_{g}$ considered are given in Table 1. Their small sample properties were investigated on samples of sizes 20 and 30 for each of the five populations of Appendix 2. Each simulation used 2000 Monte Carlo repetitions.

Some robustness constants depend on $\sigma$, the residual standard deviation of model ( $M$ ). To use the resistant estimator in practice, one nees an estimate of o which can possibly be obtained from data of previous surveys. Rey (1983, Section 6.4.5) notes an interesting property of estimators based on $\Psi_{f}(t)$ : they are not sensitive to changes in the robustness constant $k$. Thus, for estimators using function $\psi_{f}(t)$, a rough estimator of $\psi$ is all
what is needed to get meaningful results. The constants appearing in most formulae for $k$ were calculated in such a way that, at the normal model, the efficiency of the robust estimator with respect to the least squares estimator is $95 \%$ (see Rey, 1983 p. 105, and Hampel, Ronchetti, Roussecuw and Stahel, 1986, p.333).

For calculation purposes, it is of interest to consider resistant estimators $\hat{\boldsymbol{\beta}}_{g}$ as weighted ratios:

$$
\begin{equation*}
\beta_{g}=\frac{\sum w_{i} y_{i}}{\sum w_{i} x_{i}} \tag{2}
\end{equation*}
$$

where the weights depend on the type of estimator. For instance for estimator GMH, one has

$$
\begin{equation*}
w_{i}=\left(1+\frac{\left|y_{i}-\hat{\beta}_{k} x_{i}\right|}{k}\right)^{-1}, \text { for } i=1, \ldots, n . \tag{3}
\end{equation*}
$$

Formula (2) underlines that, with resistant estimators, the detection and the treatment of outliers occur simultancously. The weights identify outliers: an outlier has a large residual and receives a small weight. Resistant estimators treat outliers by incorporating the weights in the calculation of $\hat{\boldsymbol{\beta}}_{\boldsymbol{8}}$.

### 2.2 Simulation results

Ten situations are considered in this section. A situation is made of a sample size ( $\mathrm{n}=20$ or 30 ) and of a population. Simulation results for the 10 situations are presented in Table 2. Two statistics are given: the relative bias (that is the bias divided by $\bar{Y}$, the parameter of interest) and the coefficient of variation (which is equal to the square root of the mean squared error divided by $\bar{Y}$ ).

For the five populations the resistant estimators perform at least as well as the ratio estimator. In Populations 1 and 2, which are outlier free, the resistant estimators are as precise as the ratio estimator. When, as in Populations 3,4, and 5, there are outliers, resistant estimators can yield interesting gains of precision. These findings are similar to those of Lee (1990) who showed that resistant estimators increased the accuracy of estimated totals for some business surveys. In Populations 3 and 4, the relative biases of ME and GMM are almost equal to their coefficient of variation. This shows that most of their mean squared errors is related to bias. Among all the resistant estimators, the best results, in terms of bias and mean squared error, are obtained with GMH, the Hampel Krasker estimator based on function fair. This estimator is retained for further analysis in Sections 3 and 4.

Table 2: Coefficient of variation (CV) and relative biases (Rbias) of the four estimators of Table 1 for 5 populations and 2 sample sizes

| Est. | Population 1 |  | Population 2 |  | Population 3 |  | Population 4 |  | Population 5 |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\mathrm{n}=20$ | CV | Rbias | CV | Rbias | CV | Rbias | CV | Rbias | CV | Rbias |
| RE | 0.027 | 0.000 | 0.031 | -0.002 | 0.087 | -0.004 | 0.075 | -0.001 | 0.119 | 0.002 |
| ME | 0.027 | 0.000 | 0.031 | -0.009 | 0.054 | -0.044 | 0.072 | 0.058 | 0.108 | 0.028 |
| GMM | 0.027 | 0.000 | 0.030 | -0.011 | 0.054 | -0.045 | 0.072 | 0.058 | 0.098 | 0.019 |
| GMH | 0.027 | 0.000 | 0.030 | -0.011 | 0.052 | -0.034 | 0.066 | 0.039 | 0.100 | -0.005 |
| Est. | Population 1 |  | Population 2 |  | Population 3 |  | Population 4 |  | Population 5 |  |
| $\mathrm{n}=30$ | CV | Rbias | CV | Rbias | CV | Rbias | CV | Rbias | CV | Rbias |
| RE | 0.022 | 0.000 | 0.024 | -0.002 | 0.070 | 0.000 | 0.062 | 0.002 | 0.099 | 0.000 |
| ME | 0.023 | -0.001 | 0.025 | -0.009 | 0.049 | -0.042 | 0.069 | 0.060 | 0.090 | 0.031 |
| GMM | 0.023 | $-0.001$ | 0.025 | $-0.011$ | 0.049 | $-0.043$ | 0.069 | 0.061 | 0.080 | 0.022 |
| GMH | 0.022 | -0.001 | 0.025 | -0.011 | 0.044 | -0.032 | 0.059 | 0.042 | 0.082 | 0.008 |

A graph of the efficiencies of GMH as compared with the ratio estimator versus the coefficient of variation of the ratio estimator is presented in Figure 1. It suggests that interesting gain in efficiency can be obtained with resistant estimators when the coefficient of variation of the ratio estimator is larger than $5 \%$.

Figure 1: Graph of the efficiency of estimator GMH, defined as the MSE of the ratio estimator divided by the MSE for the resistant estimator, versus the coefficient of variation of the ratio estimator


## 3. ESTIMATION OF THE MEAN SQUARED ERROR OF ESTIMATOR GMH

The properties, with respect to the sampling plan, of resistant alternatives to the ratio estimator are investigated by Gwet and Rivest (1990). Their findings are used in this section to calculate the asymptotic mean squared error or resistant estimator GMH. Several estimators of this mean squared error are compared using Monte Carlo simulations. In this section $\hat{\boldsymbol{\beta}}_{g}$ denotes $G M H$, the Hampel Kraker GM-estimator based on function $\boldsymbol{\psi}_{f}$. Also function fair, $\psi_{f}$, is from now on denoted $\psi$.

Resistant estimators are not, in general, consistent for $\bar{Y}$. Their asymptotic biases can be expressed as of $\bar{Y}-\bar{X} \beta_{g}$, where $\beta_{8}$ is the population parameter estimated by $\hat{\boldsymbol{\beta}}_{g}$. Parameter $\beta_{g}$ is defined as the solution of the following equation,

$$
\frac{1}{n} \sum_{U} \psi\left(\frac{y_{i}-\beta x_{i}}{k}\right)=0
$$

Another expression for the asymptotic bias is $\bar{X}\left(\beta_{r}-\beta_{g}\right)$ where $\beta_{f}$, is the mean ratio $\bar{Y} / \bar{X}$. The asymptotic variances of resistant are derived in Gwet and Rivest (1990) by linearization. For estimator GMH, an asymptotically unbiased variance estimator is given by

$$
v\left(\bar{y}_{f}\right)=\frac{1-f}{n}\left(\frac{n \bar{X}}{\sum_{s} w_{i}^{2} x_{i}}\right)^{2} \frac{1}{n-1} \sum_{s} w_{i}^{2} e_{i}^{2}
$$

where $e_{i}$ is the residual, $y_{i}-\hat{\beta}_{g} x_{i}$, and the $w_{i}$ 's are the weights defined by formula (3). Note that when all the $w_{i}$ 's are equal to $1, v\left(\bar{y}_{g}\right)$ is equal to variance estimator $\boldsymbol{v}_{2}$ for the ratio estimator. A mean squared error estimator is given by $v\left(\bar{y}_{\delta}\right)+b^{2}\left(\bar{y}_{g}\right)$, where $b\left(\bar{y}_{g}\right)$ is an estimator of the bias $\bar{Y}-\bar{X} \beta_{g}$.

The basic bias estimator is $\hat{b}_{0}\left(\bar{y}_{g}\right)=\bar{y}_{p}-\bar{y}_{g}$, or $\hat{b}_{0}\left(\bar{y}_{g}\right)=\bar{X}\left(\hat{\beta}_{p}-\hat{\beta}_{g}\right.$ ) where $\hat{\beta},=\bar{y} / \bar{x}$. In mean squared error estimators, the squared bias is estimated. One problem with $\delta_{0}\left(\bar{y}_{g}\right)^{2}$, as an estimator of the squared bias, is that it overestimates the true squared bias. One has

$$
\begin{equation*}
\bar{X}^{2} E\left[\left(\hat{\beta}_{p}-\hat{\beta}_{r}\right)^{2}\right]=\bar{X}^{2}\left[\left(\beta_{g}-\beta_{r}\right)^{2}+c_{1}+c_{2}\right] \tag{4}
\end{equation*}
$$

where $c_{1}=2\left(\beta_{g}-\beta_{p}\right)\left(b\left(\hat{\beta}_{g}\right)-b\left(\hat{\beta}_{p}\right)\right), c_{2}=V\left(\hat{\beta}_{g}-\hat{\beta}_{p}\right)$, and $b(\hat{\theta})$ is the bias of $\hat{\theta}$ as an estimator of the population parameter $\theta$. In many instances, $c_{g}+c_{2}$ is positive and $\bar{b}_{0}\left(\bar{y}_{g}\right)^{2}$ has a positive bias. To obtain nearly unbiased estimators of the squared bias, two approaches have been considered.

First $b_{0}\left(\bar{y}_{g}\right)^{2}$ was corrected by substracting an asymptotic estimator of the bias. The quantities $c_{1}$ and $c_{2}$, appearing in formula (4), are easily estimated. Using the asymptotic linearization for resistant estimators presented in Gwet and Rivest (1990), one gets the following estimator for $c_{3}$ :

$$
\hat{c}_{2}=\frac{1-f}{n(n-1)} \sum_{s}\left[\frac{y_{i}-\hat{\beta}_{r} x_{i}}{\bar{x}}-\frac{k \psi\left(\left(y_{i}-\hat{\beta}_{g} x_{i}\right) / k\right)}{\Sigma_{s} w_{1}^{2} x_{i} / n}\right]^{2}
$$

To estimate $c_{p}$, estimators of the asymptotic biased of $\hat{\beta}_{f}$ and $\hat{\beta}_{g}$ are needed. The asymptotic bias of $\hat{\beta}_{r}$ is given by (Cochran, 1977 p. 161)

$$
\hat{b}\left(\hat{\beta}_{r}\right)=-\frac{1-f}{n} \frac{\left(s_{x y}-\hat{\beta}_{p}^{2} s_{x}^{2}\right)}{\bar{x}^{2}}
$$

where $s_{x y}$ is the sample covariance between $x$ and $y$ and $s_{x}^{2}$ is the sample variance of $x$. An expression for the asymptotic bias of $\hat{\boldsymbol{\beta}}_{g}$ as an estimator of $\boldsymbol{\beta}_{g}$ is derived in Appendix 1. It is given by

$$
\hat{G}\left(\hat{\beta}_{g}\right)=-\frac{1-f}{n}\left(\frac{\sum_{s} x_{i} e_{i} w_{i}^{3} / n}{\left(\sum_{s} x_{i} w_{i}^{2} / n\right)^{2}}-\frac{1}{k} \frac{\sum_{g} x_{i}^{2} \operatorname{sgn}\left(e_{i}\right) w_{i}^{3}}{\sum_{s} x_{i} w_{i}^{2}} v\left(\bar{y}_{g}\right)\right) .
$$

In constructing estimators of the squared bias, it seemed appropriate to prevent them from becoming negative. Thus, formula (4) suggests the following corrected estimator of the squared bias,

$$
\hat{b}_{1}\left(\bar{y}_{g}\right)^{2}=\max \left(0, \hat{b}_{0}\left(\bar{y}_{g}\right)^{2}-\bar{x}^{2}\left(\hat{c}_{1}+\hat{c}_{2}\right)\right)
$$

The variance term, $c_{3}$, was suspected to be the dominant term in the bias, thus a second corrected estimator of the squared bias was considered,

$$
\hat{b}_{2}\left(\bar{y}_{g}\right)^{2}=\max \left(0, \hat{b}_{0}\left(\bar{y}_{g}\right)^{2}-\bar{x}^{2} \hat{c}_{2}\right)
$$

The second approach was to use a jackknife estimator of the square bias. It is given by

$$
\hat{b}_{j}\left(\bar{y}_{s}\right)^{2}=n \hat{b}_{0}\left(\bar{y}_{g}\right)^{2}-\frac{n-1}{n} \sum_{s} \hat{b}_{(i)}\left(\bar{y}_{s}\right)^{2},
$$

where $\bar{b}_{(1)}\left(\bar{y}_{s}\right)$ denotes the bias estimator $b_{0}\left(\bar{y}_{z}\right)$ calculated without observation i . This estimator requires extensive computations; $n$ implicit equations have to be solved in order to calculate the $n$ pseudo values needed to evaluate one value of $\hat{b}_{j}\left(\bar{y}_{g}\right)^{2}$.

## 32 Simulation results

Monte Carlo comparisons of four mean squared error estimators for the resistant estimator, obtained by adding to the variance estimator $\mathrm{v}\left(\bar{y}_{g}\right)$ one of $\hat{b}_{0}\left(\bar{y}_{g}\right)^{2}, \hat{b}_{1}\left(\bar{y}_{g}\right)^{2} b_{2}\left(\bar{y}_{g}\right)^{2}$, or $\bar{b}_{j}\left(\bar{y}_{g}\right)^{2}$, and of the variance estimator $v_{2}$ for the ratio estimator are presented in Table 3. For the 5 MSE estimators, the relative bias, the coefficient of variation and the coverage of the $95 \%$ confidence interval are reported.

Table 3: Relative biases (RB expressed in \% of the target population parameter; for $v_{2}$, it is the MSE of the ratio estimator, and for the others, the MSE of the resistant estimator), coefficients of variation (CV) and coverages of the $95 \%$ confidence interval (for $\mathbf{v}_{2}$, the confidence interval is defined by $\bar{y}_{f} \pm t_{025}\left(v_{2}\right)^{1 / 2}$, for the other $\left.\bar{y}_{z} \pm t_{.025}\left(v\left(\bar{y}_{z}\right)\right)+b^{2}\right)^{1 / 2}$

| MSE Esi. | Population 1 |  |  | Population 2 |  |  | Population \$ |  |  | Population 4 |  |  | Population 5 |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\mathrm{n}=20$ | RB | CV | CO | RB | CV | CO | RB | CV | CO | RB | CV | CO | RB |  | CO |
| $\mathrm{V}_{2}$ | -1 | 0.4 | 93 | -8 | 0.46 | 93 | -12 | 0.83 | 82 | -8 | 0.59 | 83 | -12 | 0.57 | 89 |
| $v\left(\bar{y}_{g}\right)+G_{1}^{2}$ |  | 0.47 | 95 | 2 | 0.57 | 93 | 106 | 2.93 | 84 | 29 | 1.18 | 83 | 24 | 0.78 | 94 |
| $v\left(\bar{y}_{g}\right)+b_{2}^{2}$ |  | 0.46 | 95 | 1 | 0.69 | 93 | 62 | 2.86 | 82 | 29 | 1.33 | 81 | 23 | 0.8 | 84 |
| $v\left(\bar{y}_{g}\right)+\delta_{0}^{2}$ |  | 0.46 | 95 | 3 | 0.61 | 93 | 96 | 2.28 | 82 | 12 | 1.07 | 80 | 15 | 0.69 | 94 |
| $v\left(\bar{y}_{g}\right)+b_{j}^{2}$ |  | 0.45 | 94 | 5 | 0.62 | 93 | 12 | 2.72 | 69 | 16 | 1.22 | 79 | 10 | 0.71 | 94 |
| MSE Est. |  | pulati |  |  | ulatio |  | Pop | ulatio |  | Pop | pulatio |  |  | pulation | 5 |
| $\mathrm{n}=30$ | RB | CV | CO | RB | CV | CO | RB | CV | CO | RB | CV | CO | RB | CV | CO |
| $V_{2}$ | -4 | 0.31 | 95 | -4 | 0.4 | 93 |  | 0.71 | 85 |  | 0.51 | 86 | $-13$ | 0.5 | 88 |
| $v\left(\bar{y}_{g}\right)+b_{0}^{2}$ |  | 0.32 | 95 | S | 1.04 | 93 |  | 2.92 | 84 | 22 | 1.07 | 82 | 21 | 0.67 | 94 |
| $v\left(\bar{y}_{\varepsilon}\right)+b_{1}^{2}$ | -1 | 0.32 | 95 |  | 1.04 | 93 |  | 2.64 | 80 |  | 1.15 | 78 | 16 | 0.67 | 94 |
| $v\left(\bar{y}_{p}\right)+b_{2}^{2}$ | -1 | 0.28 | 95 | 0 | 0.94 | 92 |  | 2.28 | 79 | 5 | 0.97 | 78 | 10 | 0.58 | 93 |
| $v\left(\bar{y}_{g}\right)+b_{j}^{2}$ | -5 | 0.28 | 95 |  | 0.87 | 92 |  | 2.53 | 59 | 2 | 1.02 | 75 | -1 | 0.6 | 92 |

Table 3 reveals that the MSE estimators for $\bar{y}_{g}$ have positive biases. The three modifications to the basic squared bias estimator $\hat{b}_{0}^{2}$ reduce this bias. The coefficients of variation of resistant mean squared error estimators are quite large. In comparing them to the coefficient of variation of $\mathbf{v}_{2}$, one should bear in mind that, in Populations 3, 4, and 5, $v_{2}$ is estimating a parameter which is much larger than the resistant mean squared
error. In Population 3, the variances of $\mathbf{v}_{2}$ and of $\mathbf{v}\left(\bar{y}_{\boldsymbol{z}}\right)+\hat{b}_{2}^{2}$ are comparable; in Population 5, $\mathbf{v}_{2}$ has a larger variance than $v\left(\bar{y}_{g}\right)+\hat{b}_{2}^{2}$. Among the three modifications to the basic squared bias estimator, $\hat{b}_{2}^{2}$ provides the best results. It yields a mean squared error estimator which, among the mean squared error estimators for $\bar{y}_{g}$, has a low coefficient of variation, a low bias and which leads to confidence interval swith coverage rates relatively close to $95 \%$. For Populations 3 and 4 the coverages of $95 \%$ resistant intervals are poor. For these populations, the MSE of $\bar{y}_{z}$ is smaller than that of $\bar{y}_{s}$; however, the actual coverages of the confidence intervals obtained with $\bar{y}_{r}$ are closer to $95 \%$ than those obtained with $\bar{y}_{g}$, especially when $n=30$. Figure 2 sheds some light on this phenomenon. It shows a strong relationship between the coverages of resistant confidence intervals and the bias components of the MSE. Resistant confidence intervals have good coverage rates when the squared bias of $\bar{y}_{g}$ is small, say less than $15 \%$, with respect to its mean squared error.

Figure 2: Coverage of the resistant $95 \%$ confidence interval obtained with bias estimator $\hat{b}_{2}$ versus the bias component of the resistant estimator


## 4. CONDITIONAL PROPERTIES OF RESISTANT ESTIMATORS

The actual number of outliers in the sample is an important feature which influences the behaviour of estimators. For instance, one suspects that the ratio estimator performs well when outliers are not numerous. This had led researchers to consider the conditional behaviour, for a fixed number of outliers in the sample, of estimators (Hidiroglou and Srinath, 1981; Rao, 1985). When populations, such as Populations 4 and 5, can be divided in two strata, one containing the outliers $\left(U_{t}\right)$ and one outlier free $\left(U_{2}\right)$, it is a of interest to compare estimators when a fixed number of outlier is sampled. This can be done by using a stratified sampling plan of $n$, units in $U_{1}$ and $n_{2}$ units in $U_{2}\left(n_{1}+n_{2}=n\right)$. This section presents conditional analyses for the populations of Appendix 2 having an easily identifiable outlier stratum.

### 4.1 Conditional biases and coverages for samples of size 20 drawn of Population 4

In Population 4, the outlier stratum is made of the 35 units generated with a slope of 1 . This stratum is easily identifiable on the graph of Appendix 2. In samples of size 20, drawn out of Population 4, the expected number of outliers is $20 * 7 \%=1.4$. Such samples should contain between 0 and 4 outliers; the unconditional probabilities, calculated using a Poisson approximation to the distribution of the number of outliers, are given in Table 4. For each value of $n_{1}$ between 0 and $4,2,000$ Monte Carlo repetitions of a stratified random sampling plan of $n_{1}$ units in $U_{1}$ and of $20-n_{k}$ units in $U_{3}$ were used to calculate the conditional biases of $\bar{y}_{g}$ and $\bar{y}_{r}$ and the conditional coverages of confidence intervals for $\bar{Y}$ obtained with $\bar{y}_{g}$ and $\bar{y}_{r}$. The confidence interval constructed with $\bar{y}_{g}$ used mean squared error estimator $v\left(\bar{y}_{g}\right)+\hat{b}_{2}^{2}$; the one constructed with $\bar{y}$, used $v_{2}$ as a variance estimator.

Table 4: Conditional relative biases (RB) and conditional coverages (CO) of $95 \%$ confidence intervals constructed with estimators $\bar{y}_{g}$ and $\bar{y}_{r}$ for samples sizes 20 drawn from Population 4

| * of outliers | 0 | 1 | 2 | 3 | 4 |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Probability | . 247 | . 352 | . 242 | . 113 | . 039 |
| RB ( $\bar{y}_{8}$ ) | . 071 | . 046 | . 019 | -. 012 | -. 047 |
| $\mathrm{RB}\left(\bar{y}_{p}\right)$ | . 071 | . 019 | -. 029 | -. 075 | -. 118 |
| $\operatorname{CO}\left(\bar{y}_{g}\right)$ | . 311 | 871 | . 999 | 1.000 | 1.000 |
| $\operatorname{CO}\left(\bar{y}_{r}\right)$ | . 280 | . 994 | . 999 | . 995 | . 925 |

As given in Table 2 for $n=20$ in Population 4, the unconditional bias of $\bar{y}_{g}$ is $4 \%$. This is much larger than that of $\bar{y}_{f}$ which is essentially 0 . The conditional analysis of Table 4 tells a different story: for 0 outliers $\bar{y}_{g}$ and $\bar{y}_{g}$ have the same bias, the bias of $\bar{y}_{g}$ is large with 1 outlier ( $p=352$ ), and $\bar{y}_{r}$ is more biased than $\bar{y}_{g}$ when there is more than one outlier ( $p=.394$ ). Thus, from a conditional point of view, one can say that $\bar{y}$, is more biased than $\bar{y}_{g}$. The conditional coverages of the $95 \%$ confidence intervals show that the unconditional low coverage is caused by samples with no outliers where $\bar{Y}$ is overestimated and where the variability of the estimators is underestimated.

### 4.2 Conditional biases and coverages for samples of size 30 drawn of Population 5

In population $5, U_{1}$ is defined as the set of units with the 13 largest absolute residuals. These units are easily identified on the graph of Appendix 2 as units with $y$-values smaller than what was expected for their $x$-values. In a sample of size 30 , the average number of outlier is 1.1 ; there should be between 0 and 5 outliers. The corresponding unconditional probabilities, calculated under a Poisson approximation, are given in Table 5. For each value of $n_{l}$ between 0 and $5,2,000$ Monte Carlo repetitions were used to calculate the conditional biases and the conditional coverages. The results are presented in Table 5.

Table 5 reveals that the conditional bias of $\bar{y}_{g}$ is systematically smaller than that of $\bar{y}_{p}$, also the coverage of the resistant confidence interval is systematically closer to the nominal level of $95 \%$ than the coverage of the confidence interval for the ratio estimator. In this example the resistant procedure provides results which are more accurate than those provided by the classical approach.

Table 5: Conditional relatives biases ( RB ) and conditional coverages (CO) of $95 \%$ confidence intervals constructed with estimators $\bar{y}_{g}$ and $\bar{y}_{p}$ for samples of size 30 drawn from Population 5

| \# of outliers | 0 | 1 | 2 | 3 | 4 | 5 |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: |
| Probability | .190 | .316 | .262 | .145 | .060 | .020 |
| $\operatorname{RB}\left(\bar{y}_{g}\right)$ | .067 | .031 | -.005 | -.044 | -.091 | -.131 |
| $\operatorname{RB}\left(\bar{y}_{r}\right)$ | .093 | .031 | -.024 | -.075 | -.129 | -.172 |
| $\operatorname{CO}\left(\bar{y}_{g}\right)$ | .824 | .949 | .979 | .978 | .968 | .912 |
| $\operatorname{CO}\left(\bar{y}_{r}\right)$ | .640 | .939 | .932 | .964 | .907 | .782 |

## 5. CONCLUSIONS

Outlier resistant procedures are interesting alternatives to the ratio estimator when a regression model through the origin provides a good description of most population units. In the development of such procedures the choice of the robustness constant $k$ seems to be a crucial point. In this work it has been selected to satisfy efficiency criteria. In investigating resistant estimators for a particular type of data, it is suggested to try several values of $\boldsymbol{k}$. Taking too small a value can yield an estimator with a small variance as compared to its bias, the results of Section 3 suggest that this leads to poor confidence intervals. Taking too large a value of $\mathbf{k}$, one gets the usual ratio estimator, with almost no bias and a possibly large variance.

## REFERENCES

Chambers, R.L. (1986). Outlier robust finite population estimation, Joumal of the American Statistical Association, 81, 1063-1069.

Cochran, W.G. (1977). Sampling Techniques. Third edition. New York: John Wiley \& Sons.
Ernest, L.R. (1980). Comparison of estimators of the mean which adjust for large observations, Sankhya C, 42, 1-16.

Fuller, WA. (1970). Simple estimators for the mean of skewed populations. Technical Report. Iowa State University. Department of Statistics.

Gwet, J.-P., and Rivest, L.-P. (1990). Outlier resistant alternatives to the ratio estimator. Prépublication, Département de mathématiques et de statistique, Université Laval.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.E. (1986). Robust Statistics: The Approach Based on Influence Functions. New York: John Wiley \& Sons.

Hidiroglou, M.H., and Srinath, K.P. (1981). Some estimators of the population total from simple random samples containing large units. Joumal of the American Statistical Association, 76, 690-695.

Kish, L. (1965). Survey Sampling. New York: John Wiley \& Sons.
Lee, H. (1990). Outlier resistant regression estimators. Presentation given at the 1990 Annual Meeting of the Statistical Society of Canada in St-John's, Newfoundland.

Rao, C.R. (1971). Some aspects of statistical inference in problems of sampling from finite populations in Foundations of Statistical Inference, Godambe and Sprott (Eds.). Toronto: Holt, Rinehart \& Winston.

Rao, J.N.K. (1985). Conditional inferences in surveys sampling. Survey Methodology, 11, 15-31.
Rey, W.J.J. (1983). Introduction to Robust and Quasi-Robust Statistical Methods. New York: Springer Verlag.
Robinson, J. (1987). Conditioning ratio estimates under simple random sampling, Joumal of the American Statistical Association, 82, 826-831.

Searl, D.T. (1966). An estimator which reduces large true observations, Joumal of the American Statistical Association, 61, 1200-1204.

Tambay, J.L. (1988). An integrated approach for the treatment of outliers in sub-annual economic surveys, in Proceedings of the Section on Survey Research Methods, American Statistical Association.

## APPENDIX 1: DERIVATION OF THE ASYMPTOTIC BLAS OF $\hat{\beta}_{\boldsymbol{g}}$

Define $\hat{8}(\beta)$ as follows:

$$
\hat{\mathcal{E}}(\beta)=\frac{1}{n} \sum_{z} \psi\left(\frac{y_{i}-\beta x_{i}}{k}\right) .
$$

and let $g(\beta)$ denote the expectation of $\hat{\mathscr{g}}(\beta)$. By using a second order Taylor series expansion, one obtains the following

$$
\begin{equation*}
\hat{g}(\hat{\beta} g)-\hat{g}\left(\beta_{g}\right)=\left(\hat{\beta}_{g}-\beta_{g}\right) \hat{g}^{\prime}\left(\beta_{\varepsilon}\right)+\frac{1}{2}\left(\hat{\beta}_{g}-\beta_{\varepsilon}\right)^{\prime \prime} \hat{g}^{\prime \prime}\left(\beta_{\varepsilon}\right)+o_{p}(1 / n), \tag{A.I}
\end{equation*}
$$

where $\hat{g}^{\prime}(\beta)$ and $\xi^{\prime \prime}(\beta)$ are the first and second derivatives with respect to $\beta$ of $\hat{g}(\beta)$. For instance,

$$
g^{\prime}(\beta)=\frac{-1}{n} \sum_{s} \frac{x_{i}}{k} \psi^{\prime},\left(\frac{y_{i}-\beta x_{i}}{k}\right) .
$$

Solving (A.1) for $\hat{\beta}_{g}-\beta_{g}$ and picking the solution converging towards 0 yields

$$
\hat{\beta}_{g}-\beta_{g}=\frac{-\hat{g}\left(\beta_{\varepsilon}\right)-\sqrt{\left[\delta^{\prime}\left(\beta_{\varepsilon}\right)\right]^{2}-2 g^{\prime \prime}\left(\beta_{g}\right) \delta\left(\beta_{\varepsilon}\right)}}{\xi^{\prime \prime}\left(\beta_{g}\right)}+o_{p}(1 / n) .
$$

Since $\hat{g}\left(\beta_{q}\right)$ is $o_{p}(1 / \sqrt{n})$ and all the other terms are $o_{p}(1)$, one can expand the square root of the right hand side of the previous expression. This yields

$$
\hat{\beta}_{g}-\beta_{g}=-\frac{\hat{\varepsilon}\left(\beta_{g}\right)}{\hat{g}^{\prime}\left(\beta_{g}\right)}-\frac{\hat{g}^{\prime \prime}\left(\beta_{g}\right)\left[\hat{\delta}\left(\beta_{\varepsilon}\right)\right]^{2}}{2\left[\hat{g}^{\prime}\left(\beta_{g}\right)\right]^{3}}+o_{p}(1 / n) .
$$

Consider the right hand side of this expression. In the second term $\left[\delta\left(\beta_{q}\right)\right]^{2}$ is $o_{p}(1 / n)$, thus, up to $o_{p}(1 / n)$, one can replace $\hat{\boldsymbol{g}}^{\prime \prime}\left(\beta_{\varepsilon}\right)$ and $\hat{g}^{\prime}\left(\beta_{\varepsilon}\right)$ by their expectation $\boldsymbol{g}^{\prime \prime}\left(\beta_{g}\right)$ and $\boldsymbol{g}^{\prime}\left(\beta_{g}\right)$. The first term of the left hand side of the equality is equal to

$$
\begin{equation*}
-\frac{g\left(\beta_{g}\right)}{g^{\prime}\left(\beta_{g}\right)} \frac{1}{1+\left(g^{\prime}\left(\beta_{g}\right)-g^{\prime}\left(\beta_{g}\right)\right) / g^{\prime}\left(\beta_{g}\right)} . \tag{A.2}
\end{equation*}
$$

Since $\left.\left(\hat{g}^{\prime}\left(\beta_{g}\right)-g^{\prime}\left(\beta_{g}\right)\right) / g^{\prime}\left(\beta_{g}\right)\right)$ is $o_{p}(1 / \sqrt{n})$, by taking a one term Taylor series expansion, one can get an $o_{p}(1 / n)$ approximation of (A.2). Putting this back in the expansion for $\hat{\beta}_{g}-\beta_{g}$ yields

$$
\hat{\beta}_{g}-\beta_{g}=-\frac{\hat{g}\left(\beta_{g}\right)}{g^{\prime}\left(\beta_{\varepsilon}\right)}+\frac{\hat{g}\left(\beta_{\varepsilon}\right)\left(\delta^{\prime}\left(\beta_{g}\right)-g^{\prime}\left(\beta_{\varepsilon}\right)\right)}{\left[g^{\prime}\left(\beta_{g}\right)\right]^{2}}-\frac{g^{\prime \prime}\left(\beta_{\varepsilon}\right)\left[\delta\left(\beta_{\varepsilon}\right)\right]^{2}}{2\left[g^{\prime}\left(\beta_{g}\right)\right]^{3}}+0_{p}(1 / n) .
$$

Thus, up to $o(1 / n)$, the expectation of $\hat{\beta}_{g}$ is equal to

$$
\beta_{g}+\frac{\operatorname{cov}\left(\delta\left(\beta_{\varepsilon}\right) \delta^{\prime}\left(\beta_{\varepsilon}\right)\right)}{\left[g^{\prime}\left(\beta_{g}\right)\right]^{2}}-\frac{\delta^{\prime \prime}\left(\beta_{g}\right)\left(\delta\left(\beta_{g}\right)\right)^{2}}{2 g^{\prime}\left(\beta_{g}\right)\left[\delta^{\prime}\left(\beta_{g}\right)\right]^{2}} .
$$

For estimator GMH, the first two derivatives of g can be estimated by

$$
g^{\prime}\left(\beta_{8}\right)=\frac{-1}{n} \sum_{s} \frac{x_{i}}{k} w_{i}^{2} \text { and } \xi^{\prime \prime}\left(\beta_{g}\right)=\frac{-1}{n} \sum_{s}\left(\frac{x_{i}}{k}\right)^{2} \operatorname{sgn}(e i) w_{i}^{3} .
$$

This leadsto equation (5) as an estimator of the asymptotic bias of $\hat{\beta}_{g}$.

Appendix 2. Populations for the Monte Carlo study


## SESSION 9

## Are Our Statistical Products Fit for Use?

# MOVING FROM DESCRIPTIVE STATISTICS TO INFERENCE 

K. O'Conor, B. K. Atrostic, and R. Gillette ${ }^{1}$


#### Abstract

Government statistical organizations primarily provide descriptive statistics. However, as the needs of their users have changed, the outputs have evolved. Organizations have begun to change their sample designs, public use files, publications.

This paper argues that one of the goals should be to produce data sets that allow users to bring the full power of modern statistics to the analysis of government data, not just to produce more descriptive statistics. Wider access to existing descriptive statistics is clearly needed too. For example, these should be distributed in a variety of ways such as floppy disks, CD-ROM, or bulletin boards. A balanced customer-driven and supplier-taught blend of high quality descriptive and inferential measures is needed.


KEY WORDS: Microdata; Sample design; Public use files.

## 1. INTRODUCTION

Over the past few years, the quality revolution -- led by Deming (1986) and Juran (1988) -- has had an enormous impact on the production of Federal Statistics. In the U. S., as in Canada, major improvements are being seen in all areas of statistical methodology. This paper describes four major U. S. Federal surveys and the efforts they are making to improve the quality and usability of their public-use files, particularly for tax policy research.

Each program discussed here has addressed quality concerns in a different way. For example, the Internal Revenue Service's Statistics of Income Division (SOI) recently established an interactive process of customers and data suppliers for redesigning its SOI Individual File of sampled income tax returns. Along another vein, the Current Population Survey (CPS), produced by the U. S. Census Bureau, has calculated and released replicate weights, which can be used to produce sampling errors for virtually any type of estimate. In another Census survey, the Survey of Income and Program Participation (SIPP) has created an on-line relational database to be more responsive to users' multiple needs. Finally, the Survey of Consumer Finances (SCF), produced by the U. S. Board of Governors of the Federal Reserve, introduced stochastic relaxation to adjust for item nonresponse.

These surveys are just a few examples of work underway in the U. S. Federal sector. They were chosen because they are used by the authors in developing one of the microsimulation models maintained by the U.S. Treasury Department for tax policy research, (the U. S. Treasury Tax Model). (See Figure 1.) The following description of that model provides a brief overview of an intense use of public data files and classic insight into the needs of modelers. Later sections describe both the four component surveys' samples and their efforts to improve the quality and usability of their public-use files, so as to achieve better tax policy models.

[^37]Figure 1.-- U.S. Department of the
Treasury's Tax Model


### 1.1 Brief Description of the U. S. Treasury Tax Model

In Canada and the United States, tax policy is developed with the aid of microsimulation models that combine the data available from tax administration sources with survey data and other information. The combined database is then adjusted to fit economic and demographic projections for future years, in order to calculate the revenue effects and distributional consequences of proposed tax law changes. The precision of these estimates varies according to the accuracy and availability of the supplemental data. The challenge is to raise the breadth and accuracy of data provided to policymakers, so that they can make more informed decisions. This requires better data, better use of those data, and better understanding by data producers of how users actually make decisions.

The development of the U.S. Treasury Tax Model has two phases: first, a match and imputation phase and, second, an extrapolation phase. The former is actually the matching or linking of statistically "similar" individuals (i.e., statistical matching) -- rather than exact record linkage of identical individuals -- using age and income as determining variables. The match is constrained so that the marginal distributions of the two populations are maintained. (Barr et al., 1987) Then, the CPS, the Survey of Income and Program Participation (SIPP), and the Survey of Consumer Finances (SCF) are used in the imputation phase. After matching and imputation, the extrapolation process forecasts to achieve the Administration's economic forecast. An overview of the U. S. Treasury Tax Model is given in Cilke and Wyscarver 1987, and detailed documentation, in Cilke and Wyscarver 1990.

In Canada, models like the Wolfson model do more exact and statistical matching than the U. S. Treasury Tax Model, in part because the structure of the Canadian statistical system facilitates such approaches. For one thing, they are able to exchange data among Federal agencies more freely.

Comparability of income measures among the surveys and reliability of their income measures are important considerations in choosing the surveys to use in building the Treasury Tax Model. Because the SOI Individual File contains little information other than (selected) income items, income is one of the key variables in any statistical match or imputation between the SOI Individual File and another survey. While income definitions vary among the CPS, SIPP, and SCF, there is a high degree of overlap.

In fact, much of the new work we wish to focus on in this paper has led to improvements in comparability of income. The sections that follow provide a brief description of the component systems used in the Treasury Tax Model and discussion of some of the new and creative projects that each of the selected surveys is undertaking to improve the usability of its data.

## 2. STATISTICS OF INCOME INDIVIDUAL FILE

### 2.1 The SOI Individual File Sample

The U. S. Internal Revenue Service began the Statistics of Income (SOI) Individual File in 1918 and released its first public-use file in 1960. The current SOI Individual File sample size alternates between 80,000 and 120,000 returns. In the current design, data from Forms $1040,1040 \mathrm{~A}$, and 1040 EZ are stratified by the larger of total net income or total net loss and the size of business income plus farm receipts. In addition, the strata are based on the presence or absence of Foreign Earned Income (Form 2555), Foreign Tax Credit (Form 1116), Profit or Loss from Business or Profession (Schedule C) and Farm Income and Expenses (Schedule F). Beginning in June of 1991, data will be collected using a new sample design, briefly described in section 2.2.

In both the current and the new designs, returns are selected in each strata using two methods. The first approach uses certain ending digits of the social security number; the second method uses ending digits of random numbers generated from transformations of the social security numbers. This two-stage process was instituted in order to build in an overlap with the Social Security Administration's Continuous Work History Sample, one of the longest running longitudinal panels in the world, which is used to collect earnings data for social security covered employees. The overlap accounts for approximately 10,000 returns when the sample size is 80,000 (or 20,000 for the 120,000 sample size). It will account for about 20,000 returns in the new design (Smith et al., 1989).

### 2.2 Redesign of the Cross Sectional Sample

### 2.2.1 Background

As mentioned, the Statistics of Income (SOI) Division of the Internal Revenue Service is redesigning its sample of individual income tax returns to provide better data for modelers to estimate the effects of proposed changes in tax policy. The redesign has three major components:

1. a longitudinal panel of about 83,000 returns, with 1987 as the base year;
2. inclusion of the returns of the dependents of selected families, so that a "tax family" can be constructed; and
3. development of a new design for the annual cross-sectional sample.

All of these components will be selected and processed annually. The first component, the longitudinal panel, will allow the measurement of actual changes in the components of income reported on individual taxpayers tax returns over time, rather depending on the repeated cross-section comparisons now available. This will provide a base for quickly drawing specialized panels. The second component, the definition of tax families, will permit the measurement of the incomes of actual families, instead of the synthetic family units previously formed by statistical matches of the SOI Individual File. This will, also, make it easier to reconcile the SOI Individual File with other data sources. The third component, the new design for the annual cross-sectional sample, will address the twin goals of strengthening the sample of income components which are the subject of tax policy and of obtaining better coverage for certain demographic groups (Hostetter et al., 1990; Czajka et al., 1990).

This redesign has been a very long task. Planning for the effort began two years ago and full implementation is expected in 1991. The longitudinal panel already has data for the years: 1987, 1988, and 1989. Definitions are being constructed and revised so that preliminary family data will be produced for our primary cases in late 1991. The system for selecting the new cross-sectional sample is being programmed and will begin selecting the sample in June 1991 (Bates, 1991). The rest of this section will focus on the interactive process which made the redesign effort unique.

### 2.2.2 Participants and Roles

In 1987 a committee of the data users and producers met to define the objectives of the overall redesign, to lay out the components and needs of the effort, and to build a framework of understanding. The committee consisted of representatives from the Treasury's Office of Tax Analysis (the principal data users), the IRS computer systems area, the IRS service centers (responsible for editing and coding SOI data), independent experts, and the SOI Division (which is charged with managing the survey). By far the most important accomplishment of this committee was the development of a framework of understanding and the agreement on goals. As with many data users and producers, neither had taken the time to understand each others needs. (For more information on how this process came about, see the upcoming paper by Susan Hostetter and Karen $O^{\prime}$ Conor, to be presented at the 1991 Joint Statistical Meetings in Atlanta.)

The next stage of planning, which began in the fall of 1989 , was spent listening to and quantifying ideas for the new sample design of the annual cross-sectional sample and defining the processing of the panel and family data. The research on the former required the most interactions and, to make that happen, a new team was formed. This team was much more focused than the 1987 committee and consisted of three or four data users from the Office of Tax Analysis (OTA), two sampling experts from Mathematical Policy Research (MPR) and a mix of economists and mathematical statisticians from the SOI Division.

### 2.2.3 Interactive Redesign Process

The Redesign Committee had many issues to decide in order to meet several competing goals. Perhaps the most difficult issue was reaching a definition of returns with complex income structures - returns most likely to be the focus of proposed tax policy. Complex returns are relatively rare and, therefore, likely to be undersampled. Certain relatively rare filing characteristics, however, are of policy interest and, hence, needed to be identified, so that sufficient numbers were sampled.

After many discussion, an initial consensus was reached. As shown below, the presence of any of 10 income components was taken as a possible indicator of complexity, and 3 filing status indicators that might be undersampled were identified:

## Complex income indicators;

- capital gain or loss
- partnership or small corporation income or loss
- itemized deductions (Schedule A)
- deduction for bome mortgage interest
- social security income
- pension or annuity income
- child care credit
- unemployment compensation
- alimony income
- alternative minimum tax


## Filing status indicators

- aged exemption
- unmarried head of household
- exemption for dependent child living at home or dependent parents

Next, in order to examine the frequency of occurrence of all possible combinations of related items, MPR constructed indices. The lower positive income strata were the focus of this investigation because a high proportion of returns with income greater than $\$ 250,000$ were already included in the sample design. A series of tabulations showed that complex returns with income less than $\$ 250,000$ were dispersed relatively evenly among the remaining indicators.

Two other objectives conflicted with the "complex return" goal. The first was to retain sufficient numbers of noncomplex returns to provide coverage for modelling and descriptive statistics. The second was to maintain as much simplicity as possible, with just two levels of stratification: income and form type. This consideration is important because the series of statistical matches, imputations, and extrapolations brings a high degree of complexity to the construction of the Treasury Tax Model -- the modelers also wanted the sample design to be as simple as possible, with just two levels of stratification: income and form type. Test sample designs balanced these conflicting objectives.

The team also determined that it was important to identify noncomplex returns on the basis of income and other characteristics that were unlikely to be the focus of proposed tax law changes. Several iterations were needed to classify returns in a way that both satisfied tax policy needs and maintained the reliability of published SOI estimates.

An initial proposal identified returns that could be undersampled as those where a substantial proportion of the total positive income ( 75 percent for returns with positive income of $\$ 60,000$ to $\$ 250,000$ and 90 percent for returns with positive income of $\$ 0$ to $\$ 60,000$ ) came from wage and salary or retirement income. (See Figure 2.) This definition turned out to be too narrow, retaining too few returns in the 13 primary categories. An expanded definition added three types of noncomplex returns: returns with alternative minimum tax preference items but zero alternative minimum tax; returns identified as complex only because of substantial tax exempt interest income; and returns which had predominately Schedule C (sole proprietorship) income, and interest and dividend income.

The expanded definition solved one problem only to raise another. The total number of sole proprietors in the sample would have been reduced by approximately 2,300 returns, concentrated in the $\$ 0-\$ 30,000$ and $\$ 30,000$ $\$ 60,000$ strata. The resulting sample would not have provided sufficient stability in the estimates of total sole proprietors' net or gross income. Since the majority of the total net or gross income in the population comes from the low income positive strata that would have been undersampled, such a change would have been unacceptable to our users. The Bureau of Economic Analysis (BEA) uses these data to produce estimates of the National Income and Product Accounts.

As a result, a third condition was added to the first two definitions: a noncomplex return was reclassified as complex if total negative income exceeded 40 percent of its total positive income. By so doing, an estimated 520 sole proprietors returned to the sample in the $\$ 0-\$ 60,000$ income range (Czajka, 1988; Hostetter, 1990).

### 2.3 Benefits

Test tabulations were run to observe the impact on the sample statistics of each of the income and form type classification constraints imposed. These provided insight for subsequent decisions. Furthermore, since several different sets of interests had to be met by the final sample, the simulation studies helped provide a more factual basis for compromise. Clearly, the give and take process which followed led to greater understanding of all of the participants' needs. It helped the team explore facts rather than criticize opinions. As the committee worked to achieve an operational definition of complexity, there were differences of opinion on the value of various tabulations because some were quite costly. The high cost of such calculations usually prevents large numbers of them from being calculated. But, all the discussed tabulations were completed; and they were worth every penny because the resulting data kept the decisionmaking on a factual level. The final result is a sample design which satisfies the needs of both the tax modelers (OTA) and the users of descriptive statistics (BEA and IRS).

## 3. CURRENT POPULATION SURVEY

### 3.1 The Current Population Sample

Treasury uses CPS and other sources to fill in the demographic and some financial information unavailable from the SOI Individual File. Begun in 1942 when the Survey of Unemployment was transferred to the Bureau of the Census, the CPS is perhaps the most widely known of the component files discussed here when the Survey of

Unemployment was transferred to the Bureau of the Census. The first CPS public-use file was distributed in 1968.

The CPS, a monthly household survey conducted by the U. S. Census Bureau for the Bureau of Labor Statistics, collects data on a wide range of topics. While initially designed to produce labor force and demographic data, CPS now also collects (primarily through a supplement to the annual March survey) information on topics such as hours worked, occupation, industry, periodic, personal and family income, migration, educational attainment, etc. It is the March CPS that is linked to the Tax Model at Treasury.

The Current Population Survey covers the civilian noninstitutional population of the United States. The sample size is 60,000 households or about 113,000 persons 16 or older. The CPS is a cluster sample of housing units stratified geographically. The sample is rotated, such that each household is interviewed for four months, not interviewed for eight months, interviewed for another four months, and then retired from the CPS sample. (Bureau of the Census, 1978) Of particular interest, here, are the replicate weights recently developed for the CPS, which allow researchers to calculate variances for a wide range of estimates.

### 3.2 Generalized Variance Functions

In the past, the CPS has used generalized variance functions to produce sampling errors. This had several drawbacks. For example, the Bureau of Labor Statistics (BLS), which is particularly interested in using the CPS to estimate the unemployment rate for each state, had to use this approximation to compute a measure of reliability for each state estimate. In so doing, they encountered two serious problems:

1) the national within-primary-sampling-unit design effect is assumed constant across states, with an adjustment for sample units determined to be out-of-scope of the survey; and
2) the ratio of between-primary-sampling-unit variance to total variance is assumed constant across time (Lent, 1991).

The BLS is now working with a file of replicate weights developed by the Census Bureau to compute variance and correlation estimates by state.

Another problem with the old procedure was that, because of confidentiality, most CPS users could not get the stratification (geographic) information needed to calculate their own variance estimates. They were forced to rely on the generalized variance tables which are only useful for percentages and totals. For very complex estimators, it was often impossible to calculate a variance.

The new generalized replication technique attempts to address both these concerns. This advance has made it possible to provide BLS with a file of replicate weights to compute variance and correlation estimates by state, while maintaining flexibility for meeting smaller users' needs -- all without sacrificing confidentiality. The Census Bureau will continue to calculate the generalized variance function and provide the corresponding tables for those who prefer to use them.

## 33 Overview of Generalized Replication Theory

For the new approach, the CPS is using the generalized replication theory developed by Robert Fay to produce the replicate weights. (Fay, 1984 and 1989; Wolter, 1985) This technique is very similar to Balanced Repeated Replication, a procedure for repeatedly constructing estimates from half the sample, such that each stratum has half of its observations included in each estimate. Any observation is in half of the samples.

In generalized replication theory, all of the observations are used in each replicate. Replicate factors are used to weight the contribution of each half sample in the replicate weight. For example, with classical balanced repeated replication, the replicate factors for a half sample in a stratum are either 0 or 2 . Using the generalized replicate theory, the replicate factors could be one half or one and a half. Since CPS has only one primary sampling unit per stratum, pseudo strata are then created by collapsing sampling strata. Some of the pseudo
strata in the CPS replication system are portions of self representing strata rather than groups of collapsed strata. The generalized replication theory uses a method to assign factors which differ by the pscudo strata.

Simulation tests have been run to evaluate Fay's generalized variance estimator. They conclude that the technique is useful when variance estimates are needed for both smooth and nonsmooth statistics or when there are very few degrees of freedom available for variance estimation (Judkins, 1990). These are the types of estimates many users of CPS data are making.

### 3.4 Calculation of the Uncertainty Using Replicate Weights

The resulting replicate weights have made it easy to calculate a measure of uncertainty. Each record has 48 replicate weights. Estimates are constructed by using the replicate weight in the calculation of the estimate, instead of the household weight, which previously had been the only source available. An estimate is calculated for each replicate. The variance is calculated using the standard sum of the differences squared variance formula multiplied by a factor of 4 to account for the use of the whole sample in each estimate.

## 35 Benefits

This method permits the researcher to calculate variances for virtually any estimate and subgroup of the population. Although there are problems with the variance estimates for small geographic subsets (the coefficient of variation of the variance for small states is as high as 50 percent), this is a tremendous improvement over generalized variances.

## 4. SURVEY OF INCOME AND PROGRAM PARTICIPATION (SIPP)

### 4.1 The Survey of Income and Program Participation Sample

The Survey of Income and Program Participation was first collected in 1983. SIPP is a multi-pancl longitudinal survey of persons aged 15 and over which measures their economic and demographic characteristics over a period of two and a half years. Core questions cover demographic characteristics, labor force participation, program participation, amounts and types of earned and unearned income received, asset ownership, private health insurance, and pension receipt. The sample size for SIPP has varied between about 11,500 and 37,000 households, due to budget constraints and panel overlap. From February to July there are three panels in the survey; the rest of the year there are two. SIPP data are collected by the Census Bureau (Census, SIPP User's Guide). The new relational database, which Census is implementing for SIPP users, will vastly improve quality and usability.

### 4.2 Early SIPP Public Use Files

The Survey of Income and Program Participation represented a major departure from the Census Bureau's approach to current surveys. SIPP was much more complex and ambitious, and resulting data offered great promise to demographers and researchers in the area of tax policy and transfer issues. Unfortunately, the producers of the SIPP initially received very negative comments from file users. The record layout was confusing. There were redundant fields (e.g., age original, age edited, and age imputed). The location of variables shifted among files. "Zero" had different meanings, such as missing, not applicable, and zero. The documentation of the editing and imputation procedures was not available and the release of clean files was not timely (ADPU, 1989).

Drawing on the experience of the user community, the Census Bureau has redesigned the SIPP public-use files. Data from the 1990 panel will be released in a person-month format--one record for each month a person is in sample. This provides a one-to-one relationship between the person interviewed and the timing of the data reported. The redesign also solves one of the most troubling problems of the public-use files where a single record recorded data representing several different time spans. In addition, the 1990 data have been reorganized. Redundant data have been eliminated; a single record layout has been developed for all waves of core data; and,
new recodes have been added to aid the user in working with some of the more complicated concepts in the survey.

With many user concerns ameliorated, Census proceeded to explore other approaches to improve the quality and usefulness of SIPP. A large scale research project at the University of Wisconsin, funded by the National Science Foundation, explored ways of making SIPP data more accessible to the academic community (David, 1989). That project developed a relational database using the 1984 SIPP panel. The database is now supported by and available through the Census Bureau. The Census Bureau is also in the process of adding data from the 1985 panel to the database.

### 4.3 SIPP Relational Database

A relational database permits the user to develop his or her own aggregate tabulations to meet individual needs. The current system developed at Census consists of the 1984 SIPP panel data, which is loaded into INGRES, a relational database software package. The relational database has 210 tables. These tables can be broken down into three types: wave, stacked and other. The wave tables have long records with many variables for one wave of data. In order to reduce the amount of storage required for the tabie, the number of rows or records in the tables is restricted. There are separate tables for persons, households, and families. There are also separate tables for different kinds of people (e.g., those earning wages and salary). The rows consist of the records for persons with valid data for the table. The columns are variables related to the subject matter of the table.

The stacked tables have short records with a few variables for all responding waves. These tables were specifically designed for use when the number of members in the sample is smaller -- e.g., persons with Medicare or persons with self-employment income.

The remaining tables provide additional data, such as:

- link tables, which link people, families and households;
- miscellaneous tables of constants, couples, marital status, retention, and recipiency change; and
- utility tables, with survey and reference dates.

These other tables are used in links to the stack and wave tables to find out information, such as how many people from a household get Medicare. The relational nature of the database permits users to link several files to get the data they want.

### 4.4 Benefits

The increased flexibility and ease of access created by the relational database has many benefits for SIPP users. It should be noted the SIPP has not been used in constructing previous Treasury Tax Models because access to the microdata was too complex before the relational database was developed. Instead, SIPP had been used to supplement analyses of tax provisions when the SIPP contained information unavailable on the CPS or other sources. Even with the relational database, SIPP is still more likely to be used for imputations than for statistical matching, because of the difference in sample size between SIPP and SOI Individual File.

## 5. SURVEY OF CONSUMER FINANCES (SCF)

### 5.1 The Survey of Consumer Finances Sample

The last component of the Treasury Tax Model that has undergone some important changes in order to better meet user needs is the Survey of Consumer Finances. The Federal Reserve Board has conducted the Survey of Consumer Finances on a triennial basis since 1983. The primary purpose of the SCF is to gather comprehensive data on assets, liabilities, pensions and income, together with descriptive data on employment, marital history, family structure, health, and other demographic variables. The 1989 SCF (like the 1983 SCF, but not the 1986) actually consists of two related surveys: a household survey and a survey of pension providers.

It is the household survey what will be discussed here because of the work of Arthur Kennickell, of the Board of Governors of the Federal Reserve, is doing in multiple imputation.

The household sample was selected using a dual frame, composed of a list and an area frame. The stratification of the list frame was a wealth proxy index and that of the area frame was geographic. For the 1989 SCF, the number of respondents from the list frame was approximately 870 households; from the area frame, 1,130 households. This household sample is important because of the high number of wealthy households responding -- 100 of the respondents each have an estimated wealth of $\$ 10-\$ 250$ million.

## 52 Household Nonresponse

The SCF is the only source of survey data on wealth that has a sufficiently large sample in the upper income strata to permit separate analyses of them. (The SIPP also surveys wealth, periodically, but its design does not provide enough observations for separate analyses of the highest strata of the wealth distribution.) High income taxpayers are very important for modeling changes in tax policy, as well as being of considerable interest to other researchers. In fact, taxpayers in the top 10 percent of the income distribution pay 50 percent of the taxes; those in the top 1 percent pay 26 percent.

Despite the importance of these data, it is generally felt that response rates decline as wealth rises. As Figure 2 illustrates, the SCF response rates corroborate this and point out the difficult problem of collecting data from such "large" taxpayers.

Figure 2: Response Rates From List Frame (Wealthy)

## WEALTH PROXY INDEX

$0<100,000$
$100,000<500,000$
$500,000<1$ million
1 million < 2.5 million
2.5 million < 10 million

10 million < 250 million
TOTAL

RESPONSE RATE
48.4\%

## \#RESPONDENTS

45
116
$39.6 \% \quad 158$
39.4\% 232
30.6\% 215
$20.1 \% \quad 100$
$34.1 \%$
866

The high income portion of the SCF sample is a population subgroup whose income and wealth are very hard to measure, due, in part, to the houschold and item nonresponse. A study focusing exclusively on wealthy houscholds as identified through a wealth index would not meet the expected 70 percent response rate required for approval by the U. S. Office of Management and Budget. (Federal surveys must be approved by the Office of Management and Budget before data can be collected.) However, these households are crucial to tax modeling. For the purposes of comparison and to make national estimates, an area frame of primarily lower income households supplemented the list sample. It raised the overall response rate to about 70 percent. Still, all surveys have subgroups with poor response rates. One of the strengths of the SCF is that there are strong variables from the sampling frame to help in modeling the missing data. Hence, it was this area which was the focus of new efforts to improve data quality for the users.

### 5.3 Item Nonresponse Adjustment

Item nonresponse is also a serious issue in the SCF. Although not nearly as large as the unit nonresponse problem, the adjusted gross income variable has an item nonresponse rate of 28.6 percent. (See Figure 3.) To deal with this problem, the SCF uses sophisticated regression techniques combined with multiple imputation to adjust for item nonresponse. The method, called stochastic relaxation, is also known as Gibbs sampling and expectations maximization algorithm (Rubin, 1987, 1990 and 1990; Geman, 1984).

# Figure 3. Nonresponse Rates For The Survey of Consumer Finances 



The theory for this approach requires all the variables in the imputation to be continuous and that all the variables can be transformed to normal. These assumptions are approximately satisfied. By using an iterative procedure, a randomized regression model is employed to estimate all the missing values for a particular item -- for example, adjusted gross income (AGI). Once all the records with missing adjusted gross income have had a value imputed, the process moves on to the next variable. The process moves through all the variables in turn, then the cycle starts again.

The actual imputation is done using a randomized regression model. There are four major steps in the imputation of each variable, as follows:

1. A set of conditioning variables are determined. For AGI there are 300 conditioning variables.
2. The variance covariance matrix is generated using cases with at least 75 percent of the conditioning variables responding.
3. For each record with nonresponse in the variable being imputed, the following substeps a and $b$ are carried out:
a. A covariance matrix is constructed from the matrix in Step 2, based on the responding conditioning variables for that record; and
b. Five randomized, independent replicates of the missing variable are imputed, using the regression defined by the covariance matrix for that record.
4. The imputed value is used as a conditioning variable in the variance covariance matrix of the next variable.

Once all the variables have been imputed, Step 1 is repeated. In this way, the procedure iterates toward the maximum likelihood estimate of the variance covariance matrix (Kennickell,1991).

### 5.4 Benefits

The benefits of stochastic relaxation are that inferences made on the imputed database (using a multiple imputation technique) have the same validity, because the distribution of the imputed data is the same as the underlying distribution. Also, the imputation takes advantage of the relationship between the variables as apposed to independently imputing each variable.

## 6. CONCLUSION

Data quality has traditionally been measured by looking at issues which concern data producers. A more complete measure of data quality would look at both sides of the question: conformance to standards and fitness for use.

The vast majority of the resources allotted to data collection are spent for the production of descriptive statistics. Increasingly, many users are also interested in the microdata (i.e., public-use files). As the sophistication of data users and their technological capabilities has increased in recent decades, the production and improvement of more flexible products such as public-use files has not kept pace. That is changing now. More importantly these improvements are coming about because data producers are listening to their customers.

As a result of growing quality consciousness in the U. S. Federal statistical agencies, the community of data producers recognized the importance of the user perspective. These four surveys exemplify a wide range of efforts to improve public access to data and the resultant usefulness of that information. The SOI Individual File redesign paid increased attention to user meeds, through greater user involvement in the redesign process. The redesigned SOI Individual File should provide more reliable basic data than the previous design. It also should provide sufficient numbers of sample cases for complex returns and other classes of returns that are likely to require separate analysis for tax policy.

Attention to user needs has made the CPS and SCF important data sources for modeling the components of economic income, and for estimating the effects of tax policy. Similar attention in the SIPP program increases the likelihood that its rich array of data can now be used, as well. All of these non-tax sources of data are needed because the basic source of tax information, the SOI Individual File, contains only the variables available on income tax returns. Many policy questions, on the other hand, require additional information; for example, calculating the effects of a tax change on a family -- rather than on a return -- basis, or modeling the effect of including an income source currently not reported on tax returns. Increased ease of use, enhanced design, and improved public use data sets expand the range of information available for tax policy analysis.

## REFERENCES

Association of Public Data Users (1989). SIPP Supplement to the APDU Newsletter, 1 supplements $1-6$ and 2 supplements $1-2$, NJ: Princeton University Computing Center.

Barr, R., and Turner, S.J. (1978). A New Linear Programming Approach To Micro Data File Merging, 1978 Compendium of Tax Research, Department of the Treasury, 131-149.

Bates, J. (1991). Creating a Database for Longitudinal Analysis of Families, Proceedings of the Section on Survey Research Methods, American Statistical Association.

Cilke, J.M., and Wyscarver, R.A. (1987). The Treasury Individual Income Tax Simulation Model, Compendium of Tax Research 1987, Washington, D. C.: Department of the Treasury, Office of Tax Analysis.

Cilke, J.M., and Wyscarver, R.A. (1990). The Treasury Individual Income Tax Simulation Model, Washington, D.C.: Department of the Treasury, Office of Tax Analysis.

Czajka, J., and Schirm, A. (1990). Overlapping Membership in Annual Samples of Individual Tax Returns, Proceedings of the Section on Survey Research Methods, American Statistical Association.

Czajka, J. (1988). Development of a New Income Classifier for a Sample of Individual Tax Returns," Proceedings of the Section on Sunvey Research Methods, American Statistical Association.

David, M., and Robbin, A. (1989). Database Design for Large Scale, Complex Data, U.S. Bureau of the Census, SIPP Working Paper 8923.

Deming, W. E. (1986). Out of the Crisis, Massachusetts Institute of Technology, Center for Advanced Engineering Study: Cambridge, Ma.

Fay, R.E. (1984). Some Properties of Estimates of Variance Based on Replication Methods, Proceedings of the Section on Survey Research Methods., American Statistical Association.

Fay, R.E. (1989). Theory and Application of Replicate Weighting for Variance Calculations, Proceedings of the Section on Survey Research Methods, American Statistical Association.

Geman, S., and Geman, G. (1984). Stochastic Relaxation, Gibbs Distribution, and the Bayesian Restoration of Images, IEEE Transactions on Pattern Analysis and Machine Intelligence.

Hostetter, S. et al. (1990). Choosing The Appropriate Income Classifier For Economic Tax Modeling, Proceedings of the Section on Survey Research Methods, American Statistical Association.

Hostetter, S., and O'Conor, K. (1991). Satisfying the Needs of Income Tax Policy Modelers While Preserving The Reliability of Descriptive Statistics, Proceedings of the Section on Survey Research Methods, American Statistical Association.

Judkins, D.R. (1990). Fay's Method For Variance Estimation, Joumal of Official Statistics, 223-239.
Juran, J.M. (1988). Juran on Planning Quality, The Free Press: New York, NY.
Lent, J. (1991). Variance Estimation for Current Population Survey State Labor Force Estimates, Proceedings of the Section on Survey Research Methods American Statistical Association.

Schirm, A., and Czajka, J. (1990). Intertemporal Stability in Total Income and the Overlap in Annual Samples of Tax Returns, Proceedings of the Section on Survey Research Methods, American Statistical Association.

Smith, C. (1989). Social Security Administration Continuous Work History Sample, Social Security Bulletin. (The Continuous Work History Sample was begun by the Social Security Administration in the 1930's. Prior to 1980 the CWHS released public-use files; however, due to concerns about the possibility of disclosure, the Social Security Administration no longer releases the data.)

Rubin, D. (1987). Multiple Imputation for Nonresponse in Surveys, John Wiley \& Sons: New York, NY.
Rubin, D. (1987). EM and Beyond, European Meeting of Biometric Association.
Rubin, D. (1990). Imputation Procedures and Inferential Versus Evaluative Statistical Statements, Proc. Census Annual Research Conference VT, U.S. Bureau of the Census.

United States Bureau of the Census (1978). The Current Population Survey: Design and Methodology, Technical Paper 40, 2.

United States Bureau of the Census (1987). Survey of Income and Program Participation: Users' Guide, 1.
Wolter, K.M. (1985). Introduction to Variance Estimation, Springer Series in Statistics, New York: SpringerVerlag Inc.

## AN EXAMINATION OF STATISTICS CANADA'S DATA QUALITY RELEASE CRITERIA

R.D. Burgess ${ }^{\prime}$


#### Abstract

When publishing data Statistics Canada authors are to advise users of any major potential limitation imposed by the quality of that data. Users are sometimes advised to use certain data "with caution". In some cases specific criteria are used to ensure lower quality data are not released. Release may be precluded through the specification of tabulations or through "data suppression" within tabulations. A user can only make direct use of these lower quality data through a more deliberate process of acquiring custom tabulations, and may there too face restrictions.


Some statisticians and users believe that data quality, by itself, should not be used as a release criteria; that only the individual user is able to decide on the appropriateness of data and that all data potentially have some information value.

The purpose of this paper is to describe data quality release practices currently in use in Statistics Canada, to examine and compare the criteria being used by various statistical programs, and to assess the merits and shortcomings of data quality release criteria.

KEY WORDS: Data Quality; Release Criteria; Data Suppression.

## 1. INTRODUCTION

### 1.1 Purpose and Scope

When publishing or otherwise disseminating data, Statistics Canada authors are to make information on concepts and methodology, and quality indicators available to users (Statistics Canada, 1986). These will afford a user the opportunity to assess whether the data describe the characteristics he or she wishes to measure, and whether the data are sufficiently accurate for the intended purpose. Users are sometimes advised to use certain data "with caution". However, most data of lower quality are not included in the specification of tabulations for publication or pre-planned dissemination. This is primarily for reasons of the survey objectives, and of tabulation and publication cost or space constraints.

Some statistical programs do have specific data quality standards or guidelines in place which place restrictions on the release of low quality data. Release of such data may be precluded through the tabulation specifications for publications. In some cases low quality data encompassed by a specification are individually deleted from or "suppressed" within a released distribution. The user can only make use of these lower quality data through the more deliberate process of acquiring unpublished or custom tabulations, and may there too face data quality release restrictions.

Some users believe that data quality, by itself, should not be used as a release criteria; that only the individual user is able to decide on the appropriateness of data and that all data potentially have some information value. Some statisticians support this point of view. Others would argue that a statistical agency has the responsibility to withhold "data" which are not fit for use and which otherwise might be taken as "official statistics".

[^38]The purpose of this paper is to describe data quality release practices currently in use in Statistics Canada, to examine and compare the criteria being used by various statistical programs, and to assess the merits and shortcomings of data quality release criteria.

Issues of data quality release criteria are separate and largely supplementary to those of confidentiality. Confidentiality issues are not discussed in this paper. However, it is understood, without effect on the discussion, that disclosure avoidance is a routine practice or consideration for all statistical outputs of the Agency.

## 12 Organization

In Section 2 there is a basic description of the nature of release criteria and how they may be applied. In Section 3 some background on the examination and discussion of release criteria within Statistics Canada is given. The general issues also are presented in Section 3. The use of release criteria at Statistics Canada, and at a few other statistical agencies, is described in Section 4. The effect, or potential effect, of certain release criteria is described in Section 5. Some arguments for and against the use of release criteria are presented and discussed in Section 6. Concluding remarks are given in Section 7.

## 2. DEFINITIONS AND TERMINOLOGY

### 2.1 Release Criteria and Suppression

In the context of this paper a data quality release criterion is first, simply a rule or set of rules used to make an subjective or objective assessment of the potential or actual quality of a set of data. It is, second, a rule or set of rules used to decide, based on the assessed quality, whether to withhold the data from publication or dissemination, or to place special restrictions on their use. Since such criteria would be applied in order to determine which data are sufficiently reliable to be made available without restriction to users, they constitute quality standards for statistical output. Data which do not meet such standards and which are not disseminated are said to be "suppressed".

For the purposes of subsequent discussion suppression will be described as "cell suppression", as suppression through table design or as suppression through "secondary restictions". For cell suppression, the rules or criteria used are applied to individual data cells or to totals within a table: resulting in the replacement of the number in a cell, or in each cell in a column or row of a statistical table, with a symbol.

For table design suppression the rules could apply to an individual variable or characteristic, or to a publication, tabulation set or a single statistical table. Table design suppression may result in the exclusion of a row or column from a table specificalion, collapsing of rows or columns, removal of a variable from a table specification or the withholding of a particular array of data from publication or release. Such suppression may apply only for certain geographic levels or areas, or for specific cross-tabulations.

Release criteria in the form of secondary restrictions on the use of data may preclude further dissemination of the data by the user; such as in quoting unreliable estimates in a report.

Suppression may apply to one or more media of dissemination used for a statistical program.

## 22 Quality and Utility

Juran and Gryna (1980) define quality as "fitness for use". Following their line, "data fit for use" are considered here as data which are of adequate accuracy and which conform to user requirements. In this paper, however, quality is equated to accuracy. Data which conform to user requirements, are simply termed "useful" or as having utility. They are those which are based on concepts suitable to the users' requirements, and which give the geographic and detail of cross-tabulated characteristics needed by them. (The data must also be timely and of reasonable cost, but these are not directly at issue here.)

Among other things, Juran and Gryna characterized fitness for use by the quality of design and the quality of conformance to design. This will be paralleled to some extent here by sometimes distinguishing between variance and bias as the two aspects of quality.

## 3. BACKGROUND AND ISSUES

### 3.1 Background

Data quality release criteria are one potential aspect of the presentation of statistical data and of quality standards. The development of statistical standards for Statistics Canada is the mandate of the Methods and Standards Committee. The Committee has an ongoing process of identifying what needs the Agency has for written standards. Release criteria are currently on the Committee's agenda. It has already had some standards and guidelines on table design prepared. These include a requirement for authors to balance utility with quality of data in the design of statistical tables.

The Agency's Advisory Committee on Statistical Methods, members of which are outside statistical experts, has discussed release criteria in principle, and current practices, and has given its views. In follow-up a Task Force is to review practices in more detail and, along with broader issues, to consider the implication of changes in these on individual statistical programs.

This paper can be considered a preliminary examination on the subject, but separate from the activities of the Task Force and the Committees.

### 3.2 General Issues

Release criteria can be viewed as one aspect of the broader subject of the quality and utility of statistical products; starting with consideration of the objectives and design of the statistical program. However, it might be argued also that the merits of release criteria have to be assessed conditioned on the fact that potentially suppressed data have already been collected and processed, and some such data might be released for no or relatively little additional cost. One might also speculate that there would be no argument with the conclusion that what is not fit for use should not be disseminated. It seems less likely to be possible to obtain a consensus on what is not fit for use, and specifically on what level of quality will render the data unfit. Indeed, much of the background discussion on the subject of release criteria to date has centred on the acceptability of cell suppression. There has been also consideration of the issues for pre-planned versus custom output, and for the medium of dissemination.

Within the broader subject of quality and utility the issues of the discussion here are:
1/ the net benefits that are or could be provided by the application of release criteria by Statistics Canada, in controlling what quality of data it releases and publishes; and
2 / the nature or form of release criteria that might or should be applied by the Agency.
These can be worded more simply as the following questions:
1/ Should there be minimum quality standards set for all statistical output? For variance? For bias? Will these increase the utility of data products, absolutely or relative to cost? Is this a responsibility of the Agency or of benefit to it?
2/ What quality criteria should be used? For table design? For data cells?
3/ To what products and types of data should these quality criteria be applied? Economic indicators? Catalogued publications versus other release vehicles? Pre-planned versus custom products?

To place these issues in context it should be noted that as currently applied by Statistics Canada, there does not seem to be a wide spread view that suppression per se is a serious impediment to the use of data. Although there are exceptions, many of the users consulted will accept that the quality and not suppression is the limitation
(e.g., Program Evaluation Division, 1988). On the other hand this view is dependent on the extent to which suppression is currently practiced by the Agency.

## 4. USE OF RELEASE CRITERIA

### 4.1 Statistics Canada Release Criteria

Statistics Canada has no explicit standards by which it can be decided whether data are of a quality adequate for release. Equally there is no policy that prohibits placing explicit data quality restrictions on the release of data.

The Agency has standards and guidelines on the design of tables for its products (Methods and Standards Committee, 1990). These include consideration of the quality of the data. In part these read:

The design of a table or a set of tables should minimize the proportion of data cells which present no information. Similarly, the level of detail presented should not implicitly overstate the accuracy of the data."
"If all cells in two or more rows or columns within the same hierarchy of characteristics within a table are, or are likely to be:
a) blank/zero (other than through random rounding);
b) suppressed due to confidentiality; or
c) of marginal or unacceptable quality;
then, subject to a need for consistency between survey cycles, these rows or columns should be collapsed, deleted or the table redesigned."

A similar guideline is given in the event there is or is likely to be a large proportion of cells of "marginal or unacceptable quality" in a table.

The definition of what is low quality data and a decision on whether such data should be released are at the discretion of program managers. The guidelines recognize the need to maximize the utility of the data, given the objectives of the individual statistical program, the known users, and cost and physical design constraints. Thus, they do not preclude the release of low quality data, which is usually a by-product of presenting more reliable data. However, they are intended to limit the relative amount of unreliable data in the Agency's statistical products. These guidelines do not restrict release for custom products. They do not imply or preclude the use of cell suppression, although the proportion of suppressed data, by implication, is intended to be limited.

Perhaps the most common manifestation of the principles underlying these guidelines is variation in the level of detail of cross-tabulations by size of geographic areas or of various sub-populations. Decisions on the level of detail, however, are not always based on explicit release criteria.

A review of several of Statistics Canada's statistical programs has found the systematic application of release criteria most common for sample surveys. It is much less common to find a census, or a survey which combines census methods with sampling methods, for which systematic release criteria are applied. Six basic forms of release criteria are in place.

1/ Coefficient of variation limit: Not surprisingly the most common release criteria are based on a coefficient of variation limit for estimated or predicted sampling error. Such a criterion is used, for example, for the Labour Force Survey, the Survey of Consumer Finance, the Family Expenditure Survey, and most of the other of the Agency's household surveys. The Labour Force Survey practice seems to be the common model or precedence for the others. (It is also the source of the sample for most household surveys.) For the Labour Force Survey a coefficient of variation of $331 / 3 \%$ is used as the release criterion. Cell suppression is used for data items which exceed this limit. The practice applies for all tabular output, including custom tabulations. If an estimate of a count is suppressed, a corresponding rate is also suppressed.

The Health and Activity Limitation Survey data were also subject to suppression based on estimated relative sampling error. For this 1986 post censal survey, data for Canada, provinces and Census Metropolitan Areas were suppressed in a table if the coefficient of variation exceeded $25 \%$. For tables containing data for sub-provincial areas, other than Census Metropolitan Areas, a criterion of $33 \%$ was used so as to not suppress too great an amount of the data.

A $16.5 \%$ upper limit on coefficient of variation is used for the International Travel Survey annual statistics. For catalogue publications this limit is largely imposed through table design.

For some of the surveys with coefficient of variation criteria for publication data, these criteria are not applied for custom tabulations or for data disseminated on micro-data tapes.

2/ Rounding error limit: The error introduced by rounding is infrequently used in release criteria. It is used for the Labour Force Survey. For this survey estimates of counts are usually rounded to the nearest 1,000 or 100 . The expected error due to this rounding is assessed as a proportion of the expected sampling error for a given cell size and a limit placed on it. Cells and corresponding rates are subject to suppression accordingly.

3/ Comparison with benchmark data: There are those statistical programs for which some of the data collected monthly from a sample of respondents can be compared with a more reliable, but much less frequently available data source. These more reliable data are sometimes used as benchmark values. If the monthly value deviates by more than a pre-specified amount from the benchmark, then the monthly cell value is suppressed. The Monthly Survey of Manufacturing uses such cell suppression criteria based on a comparison with the Annual Census of Manufactures. The tolerance limits vary with the time difference between the Monthly and the Census being used as the benchmark.

4/ Imputation rate limit: Imputation is a potential source of error for most statistical programs. Limits are sometimes, but it seems infrequently, set on the rate of imputation for released data. Such a limit is used for the Annual Retail Trade Survey. Table rows (Major Trade Group) will be collapsed if $30 \%$ or more of the data for a row in a provincial table have been imputed.

5/ Restriction on secondary publication: For some surveys unpublished data are released without suppression, but as part of the purchase agreement restrictions are placed on subsequent publication. Either the data not meeting the standard release criteria are not to be presented in reports or disseminated further, or the data quality indicator or a caveat is to be presented with them. These types of rules are used, for example, with micro-data files from the Labour Force Survey and with sub-provincial Consumer Price Index data.

6/ General rating of data and data sources: It is not uncommon in the case of sample surveys to fix limits on the geographic detail of output. This is inherently implied or explicit at the sample design stage. No consideration is given to publishing such data whether there are or are not other release criteria for pre-planned or custom output.

Less common are release criteria for analytical data. For much of these data the input data are highly aggregated and are probably of more than adequate quality so that release criteria would have little effect. A major exception is the System of National Accounts (SNA). For some sectors, those producing industry estimates, a quality rating system is in place and is sometimes used as the basis for release. Under this rating system there is a clearly articulated procedure for assessing the quality of various data and aggregations. The highest quality rating, a " 1 ", is used for data based on a census, or for data from administrative records which need minimum adjustment. The lowest "acceptable" quality rating is "3". The lowest rating, "unacceptable", is " 4 ". Until recently data with a rating of 4 were not published. This is still the case for some sectors of the SNA. In one sector the methodology of the rating system was revised extensively in 1990 (Taillon, 1990). What is published for this sector is currently unchanged from that of the older rating system and release criteria. However, some of the published data now have a rating of 4 .

The review also found statistical programs which do not have explicit or systematically applied release criteria, but which restrict dissemination on an ad hoc basis. Such restrictions through table design are not uncommon and potentially might be applied for any statistical program.

Two important examples are the Census of Population and the Census of Agriculture. For these there are no explicit standards or guidelines by which it can be decided, based on quality, whether data can or should be released to users. There are, however, formal certification processes by which the quality of the data is to be assessed. If quality problems are found, a judgement is made on whether a cautionary note is to be given with the data or the data suppressed.

This judgement has, but infrequently, resulted in collapsing of the planned published breakdown of a variable into a smaller number of values; e.g, data from the structural type of dwelling question for the 1981 Census. It has, but infrequently, resulted in the removal of data for a variable from planned output; e.g., data from the Aboriginal Peoples question for the 1986 Census. It has resulted in the suppression of characteristic data for specific geographic areas; e.g., for the incompletely enumerated Indian reserves and Indian settlements in the 1986 Census. (In this case suppression was in large part dictated by the absence of data on characteristics of the population in the given geographic areas.)

Such suppression in the Census of Population has not extended to custom output. The quality of data is routinely discussed with custom data users and quality documentation provided.

Similar practices are used for the Census of Agriculture.

### 4.2 Comparison With Release Criteria In Use At Other Agencies

Statistics Canada, of course, is not the only statistical agency where release or suppression criteria are used. A review of the statistical standards and guidelines of a few of these other agencies found the following items on the topic of release criteria:

1/ Office of Management and Budget - United States (Office of Management and Budget, 1988)
" ... no data collected in a statistical survey should be suppressed or withheld from release unless ... the data fail to meet the quality standards for publication that the sponsoring agency has specified in its request for ... data collection ... ". This standard (in addition to confidentiality restrictions) is to apply to all U.S. federal statistical agencies.

2/ Bureau of the Census, U.S. Department of Commerce (Bureau of the Census, 1975)

- Individual data cells should not be suppressed for reason of large sampling error or non-sampling variances, although good judgement has to be exercised in the definition of the tables to be published.
- Tables not published because most of the estimates are unreliable, in terms of sampling error, can be released to users with appropriate indications of the sampling error.
- Serious bias is a valid reason for suppression of data in a published table.
- If unreliable data are released (e.g., in a custom retrieval), the recipient should be told, in writing, to include in any publication of the data the sampling error estimates and to outline any serious biases known to exist.

A description of the Bureau of the Census practices was presented by Shapiro et al (1985).
"... where the samples are so small as to render the estimates of the sampling error themselves highly unreliable, estimates are often suppressed. The definition of 'very large errors' will vary from 15 percent to 100 percent or more, usually depending on the relative error of the major statistics released from a given survey. For example, if a U.S. total has a relative error of 1 percent, area detail with more than 15 percent
may be suppressed, whereas in a survey where the U.S. total has a 10 percent relative error, estimates for area detail may not be suppressed until the relative error reaches 75 percent or more."
"The suppression of data with high non-response rates is a widespread practice throughout the economic areas but it is not done in the demographic areas where non-response rates are rarely very high. Since higher non-response rates increase the potential for substantial biases in published survey estimates, publication of data is sometimes suppressed. However, the non-response level requiring suppression will vary from division to division and even from survey to survey within the same division. Generally, when non-response exceeds ... 25 percent ... "

The practices of the Bureau of the Census and of Statistics Canada are similar in their diversity. The specifics are different. For example, Statistics Canada's suppression seems directed more to cell totals or counts. A review of a few Bureau of the Census publications suggests that Agency places more emphasis on suppressing averages, medians and percentages for quality reasons, without suppressing the corresponding cell totals. One reason given for this is to permit users to form their own aggregations (Bureau of the Census, 1989). Indeed concealment of data is not the objective of cell suppression for either of these Agencies. For Statistics Canada tables, however, the user would have to do much more work to derive or approximate the suppressed data. Where enough cells have been suppressed these data are in effect concealed.

Like Statistics Canada there seems not to be universal application of release criteria within other statistical agencies; at least not for cell suppression. There are differences among Agencies on how a given type of survey is treated with respect to release criteria. Labour force surveys serve as an example. Like Statistics Canada, New Zealand's Department of Statistics and the United States' Bureau of Labor Statistics use cell suppression for high relative sampling error. For the New Zealand survey, however, corresponding percentages (e.g., unemployment rates) are not usually suppressed when the estimated count is (Department of Statistics, 1990). The suppression used by the Bureau of Labor is primarly for percentages, but estimated counts for states and defined geographic areas will also be suppressed if much stricter design criteria for these areas are not met (Bureau of Labor, 1988 and 1990). The Australian Bureau of Statistics does not use cell suppression for its Labour Force Survey. Instead data with a coefficient of variation exceeding $25 \%$ are flagged (Australian Bureau of Statistics, 1990).

The pattern of cell suppression by Agency will be found to be very different if one examines, for example, household expenditure and income surveys, and family expenditure surveys.

The table design practices of these other Agencies were not assessed for this paper.

### 4.3 Data Quality Measurement and Release Criteria

Issues of release criteria, of course, cannot be separated from those of data quality evaluation and of informing users of the quality of data. Suppression can be considered only if there is some form of quality evaluation or certification of data. Then, is it not necessary and sufficient to inform users of the quality? The implied responses of statistical programs in this regard can be very different.

In terms of the presentation of data quality information corresponding to criteria for suppression, the Statistics Canada statistical programs reviewed fall into four categories. There are those that:

1/ publish corresponding data quality measures within the data tables, for all data cells;
2/ publish corresponding data quality measures separate from the data tables, for all data cells;
3/ flag marginally unreliable data and give the corresponding quality criteria for these and suppressed data, but do not publish specific indicators of data quality for data not flagged and not suppressed;
4/ do not flag data based on quality or publish data quality measures.
The implication for data published under categories 3 and 4 is that data not flagged and not suppressed can be used without concern for their quality. That is, they are fit for any use, while flagged data are fit for some uses and suppressed data are unfit for any use. For some surveys this is more or less explicitly stated in the publications.

## 5. EFFECT OF RELEASE CRITERIA

### 5.1 Extent of Cell Suppression

The importance or potential importance of release criteria is more than a matter of philosophical discussion. Where they are based on major components of error, the effect of release criteria can give a clear and easily understood indication of the scope of data, from a statistical program, which are fit for use from the perspective of the survey designer. For some surveys the release criteria have little visible effect on data published in catalogued publications. For other surveys, for individual tables a majority of data are suppressed.

The effect of cell suppression for selected publications of 4 surveys has been quantified and presented in Table 1. The four are sample surveys and coefficient of variation criteria are used. In examining these results one can assume that the tables as defined represent what users want and what is consistent with the survey design standards. The publications reviewed for this assessment show some differences in presentation strategies for data and data quality indicators or measures. These are described in the last column of Table 1.

Table 1: Proportion of Cells Suppressed for Reason of High Relative Sampling Error - Selected Surveys and Publications

| Survey Publication | Geographic Detail | Data Cells <br> Total Suppressed |  | Data Presentation Practices |
| :---: | :---: | :---: | :---: | :---: |
|  |  | \# | \% |  |
| Labour Force Survey ${ }^{1}$ | Canada \& Provinces | 18,200 | 9\% | - table detail varies according to province size data quality indicators of sampling error included in most data tables |
| Health and Activity ${ }^{2}$ Limitation Survey | Canada, Provinces \& Territories | 39,700 | 18\% | - table detail does not vary by province or by sup-provincial area cells judged to be of marginal quality are flagged no other indicators of sampling error for published data |
| General Social Survey ${ }^{3}$ | Canada \& some Regional data | 11,900 | $36 \%$ | indicator of sampling error for published data are cell size ranges for "between $16.5 \%$ and $33 \%$ " and "less than $16.5 \%$ " |
| Household Facilities ${ }^{4}$ and Equipment Survey | Canada \& Provinces | 3,300 | 13\% | - table detail does not vary by province <br> - data quality indicators of sampling error included in most tables |

${ }^{1}$ The Labour Force, Catalogue No. 71-001, March 1990. The percentage of suppressed cells includes those suppressed for rounding error. Release criteria are as described in text.
2 Highlights: Disabled Persons in Canada, Catalogue No. 82-602, March 1990 and Sub-provincial Data for Nova Scotia, Ontario and British Columbia, Catalogue Nos. 82-605, 82-608 \& 82-612, March 1989. Release criteria are as described in text. Unlike the other surveys illustrated here, zero value cells were not suppressed.
${ }^{3}$ Patterns of Criminal Victimization in Canada, Catalogue No. 11-612 \#2, March 1990. Coefficient of variation release criteria limit is $33 \%$.
4 Houschold Facilities and Equipment 1990, Catalogue 64-202, October 1990. The coefficient of variation release criteria limit $25 \%$. A rounding error release criteria similar to that for the Labour Force Survey is also used. The percentage of suppressed cells includes those suppressed for rounding error and zero value.

The proportion of cells suppressed within these publications varies from table to table, but primarily for shorter (number of pages) tables such as are used for the Labour Force Survey publication. For that survey cell suppression varies from $0 \%$, for the primary data tables, for which the survey was designed (Singh et al, 1984), to $45 \%$ for more detailed tables. The proportion of cells suppressed at the most detailed level of cross classification, of course, is much higher. For the Labour Force Survey publication $16 \%$ of the 8,365 detailed cells are suppressed. The range by table is $0 \%$ to $71 \%$.

The Labour Force Survey (LFS) is a particularly useful example of the effect of suppression rules; of the proportion of data which have been judged to be unfit for use even for surveys with a relatively large sample size. The major users consulted on the matter agree that more detailed output would not be fit for use because of inadequate quality. Although, certain users have objected to cell suppression for specific tables, even they have only suggested table design changes (Program Evaluation Division, 1988). What could not be determined is the fitness for use of the suppressed cells, from the perspective of all users, or the value of this suppression.

## 52 Potential Effect of Cell Suppression

If cell suppression were to be applied to surveys without release criteria, what would the effect be? The potential effect of cell suppression for a selection of published tables from the 1986 Census of Population was quantified. For this exercise sample data publications and a coefficient of variation criteria of approximately $33 \%$ (almost the same as the LFS citeria), as the release criteria limit, were used. The Census of Population was chosen for illustrative purposes since the sample is large (approximately $20 \%$ of Canadian households), and it is used to provide a broad range of detailed cross-tabulations and small area data. Suppression rates were calculated with and without blank cells ("."). These cells represent either zero value or a non-zero value of less than 5 set to zero through random rounding (a confidentiality protection procedure). Not suprisingly the amount of data that would be suppressed varies greatly between tables and publications.

The publication Labour Force Activity (Cat. No. 93-111) provides data similar to that of the LFS. Of the 80,000 cells of data in this publication $16 \%$ would be suppressed under the criteria limit of $33 \%$ coefficient of variation. Considering only cells not already published as "-", $12 \%$ of cells would be suppressed. In a selected table of Census Tract data for the Census Metropolitan Area of Victoria, British Columbia, 20\% of the 15,600 data cells (or $17 \%$ of those not already "-") would be suppressed under the $33 \%$ limit (Profile Series: Census Tracts, Victoria: Part 2, Table 1, Cat. No. 95-170). For some social characteristics with skewed distributions the rate would be much higher. For example, for the selected table of 41,100 data cells dealing with mother tongue and language used in the home (Language Retention and Transfer, Table 2, Cat. No. 93-153) $64 \%$ of data cells would be suppressed under the $33 \%$ limit. A much lower $36 \%$ of data cells not already " - " would be suppressed in this last case.

Again, in examining these results one must assume that the tables as defined represent what users want and what is consistent with the survey design standards. Thus, it must be concluded that either:

1/ many data items of low quality meet users' criteria of fitness for use; or
2/ the typical rectangular hierarchical arrays of tables are a sometimes inefficient method of disseminating the data which are fit for use.

The first case would demonstrate a risk of suppression or a risk to users. The second would demonstrate a disadvantage for the particular tables of using the traditional dissemination method for these tables.

## 6. SOME ARGUMENTS FOR AND AGAINST RELEASE CRITERIA

The general issues stated in section 3 can be summarized as questions of:

[^39]when considered in the context of the particular statistical program, data, forms of error and medium of dissemination.

Some of the arguments for and against the use of release criteria can be summarized as follows:

## FOR

* Demonstration that quality standards exist and are applied
* Agency's reputation and responsibility
* Statistical agency has the expertise to arbitrate on fitness for use
* User expect the role of the Agency includes ensuring all data released are fit for use
* Sensitive data may be abused
*Will lead to improved table design
* Not all possible cross-tabulations should be considered within the scope of the survey design
* Information value of cells is not lost through cells suppression based on variance


## AGAINST

* Information control
* Data are from publicly funded statistical programs
* Criteria are arbitrary
* All numbers have information value
* Inconsistent with survey design
* Adds to difficulty of analysis
* Many suppressed cells are derivable
* Data quality measures are sufficient
* False sense of reliability of data not suppressed
* Users do not want to pay for suppressed data

Individual arguments do not necessarily apply to all aspects of the issues.

### 6.1 The value and appropriateness of release criteria

Release criteria represent a quality standard. One view is that the producers of statistical products are also users and are competent to judge what is or is not worth publishing. They also have more information at a more detailed level on what the limitations of the data are. Further, the external user should be able to assume a basic level of quality exists in the product. The professional judgement necessary to ensure this plays a key role in the design of statistical programs. Decisions on such things as sample size of statistical programs are based in part on quality standards. It should follow that it is appropriate and necessary that the Agency apply such design standards at all stages, including that of dissemination. By this argument fitness for use should not and cannot be equated to availability.

However, it can also be argued that fitness for use cannot be equated to quality either; certainly not to quality expressed by the coefficient of variation. It is recognized that for many surveys a wide range of reliable data are wanted by users. It may be possible to provide many of these through appropriate design (Cowan, 1988), however, not all potential users and uses will be known at the time of design, and not all data requests could be expected to meet the constraints and quality limitations of the statistical program. Despite these limitations the particular statistical program might be the only source of the required data available to the user. Unreliable data may be better than having no information or having to guess.

Further, data users who have any understanding of the survey process will not expect all data to be of high quality. What is "good" data presumably will be based on the users experience with other of the Agency's data for an, in some way, comparable use. To understand the quality of the data the user need have only this experience, the data and the data quality measures. By this argument suppression plays no positive part in this process.

### 6.2 Form of Release Criteria

In principle the method of suppression should minimize the loss of data for the user, and should not detract from the proper use of other data. As long as data are presented in the typical rectangular hierarchical arrays, without cell suppression either some unreliable data must be released or some reliable data must be suppressed. Cell suppression may be viewed by some as a reasonable compromise. However, a clear distinction is made between
the appropriateness of restrictions through table design and that of cell suppression. It can be argued that the former is as a matter of economy and efficiency, and unavoidable for most forms of output.

Cell suppression, on the other hand can be viewed as paternalistic and serving neither economy nor efficiency; particularly since suppressed data frequently can be deduced. Cell suppression may imply that all other data are fit for any use and some data fit for a particular use may be suppressed. Cell suppression is inconsistent with a view that data and their quality measures are inseparable. Given the definition of fit for use, it would be argued that it is better to just let users have the data with quality measures and determine on their own the fitness for each particular use.

There are those who suggest that many users do not concern themselves about data quality (Keane, 1987). It can be argued that such users must be prevented from misusing of the data, thus justifying cell suppression. If if it can be deduced, "suasion" prevents suppressed data from being used. Such suppression, however, does not prevent the misuse of data of marginal or higher quality.

### 6.3 Variance Versus Bias

The type of error is a key factor in arguments for and against suppression; especially in regard to cell suppression. Beyond what is done through table design, it is argued by some that no data should be suppressed for reason of high variance. To do so amounts to deciding, without consultation with users, what data are so unreliable as to be of no benefit to any user for any purpose. Most unreliable cells are relatively small. This is information in itself, the exact cell value is unimportant (Bureau of the Census, 1975). Implies the user cannot make an informed judgement of the quality and usefulness of the data, while at the same time using arbitrary criteria.

A case can be made that suppression should be permitted in all output, and may be the only responsible course of action, for biased or potentially biased cells. This may be valid, for example, because of systematic response error or high non-response for a question. Some data which are bias may have no information value even as a general indicator of level. As such there may be harm in giving credence to such data by publishing them.

On the other hand it can be argued that estimating bias might be difficult and establishing a criteria for suppression is also arbitrary. The case of $1,000 \%$ bias for a large cell may be clear, but that of $50 \%$ for a small cell may not be. However, measurable biased is not the real problem. If the size of the bias can be estimated, then informing the user rather than suppressing the data may be adequate and preferable. It is the bias for which not even a reasonable lower or upper bound has been approximated that is the problem. In such cases it may not be possible to provide users with quality indicators which will permit them to assess the data's fitness for use. In such a situation suppression may be the responsible course of action for some media of dissemination.

### 6.4 Nature of Data

It can be argued there may be some data for which the consequence of misuse of unreliable data is potentially serious to the Agency's reputation and for the user. In such a case it will be prudent to suppress the data. Even for less sensitive data, suppression in publications may prevent the occasional or less sophisticated user from making important decisions based on very unreliable data.

Some data have special status as what might be termed "official statistics". These might include those referenced in federal government legislation and those given special status by historical use: for example, Census population counts, unemployment rates, the Consumer Price Index and merchandise trade statistics. It can be argued that the status of these data must be protected by ensuring only meaningful data are disseminated.

There is also a difference between using Statistics Canada data for, what might be loosely termed, academic purposes and for policy purposes. In the latter case the person making the decisions is not necessarily the person who sees the data as published and does not know what the quality limitations or exact concepts are. The person who puts together the background information or report upon which a decision is based may have felt pressured
into being concerned only with having a "number" produced by Statistics Canada and not with its quality. It can be argued that private and public policy making should not be exposed to unreliable data.

On the other hand it can be argued that the data used for important policy purposes are generally not those which would be subject to suppression. This is arguable given some user requirements (e.g., D'Costa et al, 1989). Further, the opportunity for abuse potentially does exist now. The release criteria used do permit data to be published, which for policy use, might be considered very unreliable. Further, there is no extensive history of serious abuse specific to unreliable "official" statistics. However, it becomes a question of degree and of which user wants what data for what purpose. Most unreliable data are not now made available to be misused, or are not available in a timely fashion or in the right form for misuse to be a possibility. In terms of utility alone, there is no lack of demand for some of these data.

### 6.5 Medium of Dissemination

Release criteria seem to have greater support when dissemination is by catalogued publication. There is much less support and certainly less restrictive release criteria for other media of dissemination. One view is that table design techniques should be used for all pre-planned output to minimize the amount of low quality data disseminated; that cell suppression should not be used for any medium except for potentially serious bias; and that there be no data quality release criteria for custom tabulations beyond advising the user. Although, for the full range of media of dissemination some form of restrictions on secondary publication and reporting would remain a viable and appropriate exception.

Except for restrictions on secondary publication and reporting, it is not clear that even the most liberal of the data quality release criteria will be suitable for what will be the dissemination options of the near future. Electronic dissemination of data is increasing. Demand for small area data, and for data from a variety of statistical programs which can be integrated, seems to be increasing. The limits to these demands may only be confidentiality protection, and the technology needed to process and analyse the data inexpensively. The technology is unlikely to be a problem and will more likely drive the demand for more data. These users of small area data disseminated electronically will not want release criteria to be a limitation.

## 7. CONCLUSIONS

Within the bounds of legislation and government regulations it is at the discretion of the individual statistical program what, if any, explicit data quality release criteria to use. No specific arguments have been presented that this, in the most general sense, should change. On the other hand, arguments have been given against specific applications and the criteria being used, and against the absence of release criteria.

The use of release criteria for table design for pre-planned output seems an unavoidable requirement. Indeed, more rigorous criteria for some statistical programs might be in order. There is not a clear case for cell suppression for variance. Cell suppression seems acceptable for bias if the error is relatively large, or potentially so. Restrictions on secondary publication and reporting should be considered for all output to the extent that at least the quality of the data should be reported, if it is a potential issue. Other than this, there seems to be little reason for restrictions on custom output, with the possible exception of "official statistics". However, if there are to be data quality release criteria for "official statistics" these data and the output design standards might more appropriately be defined by policy.

The challenge for the Agency may be to provide more of the data quality measures necessary for a user to properly assess the fitness for use of data. This may be both difficult and critical for small area and integrated data.

## REFERENCES

Australian Bureau of Statistics (1990). May 1990: The Labour Fonce, Australia, Catalogue No. 6203.0, Australian Bureau of Statistics.

Bureau of the Census (1975). Standards for Discussion and Presentation of Errors in Survey and Census Data, Bureau of the Census, U.S. Department of Commerce.

Bureau of Labor Statistics (1988). Geographic Profile of Employment and Unemployment: 1987, Bulletin 2305, Bureau of Labor Statistics, U.S. Department of Labor.

Bureau of Labor Statistics (1990). Employment and Eamings, Volume 37 Number 7, Bureau of Labor Statistics, U.S. Department of Labor.

Cowan, C.D., and Malec, DJ. (1988). Sample Allocation for A Multistage, Multilevel, Multivariate Survey, Proceedings of the Fourth Annual Research Conference, Bureau of the Census, U.S. Department of Commerce.

Department of Statistics (1990). Household Labour Force Survey: The New Zealand Labour Force, December 1989 Quarter, Cat. No. 05.002, Department of Statistics, Wellington, New Zealand.

D'Costa, R., and Kennedy, L.W. (1989). Evaluation Report: Demography, Program Evaluation Division, Statistics Canada.

Juran, J.M., and Gryna, F.M. Jr. (1980). Quality Planning and Analysis, McGraw-Hill Book Company, New York.

Keane, J.G. (1987). Surveying Survey Development: Some Stage-Setting Perspectives, Proceedings of the Third Annual Research Conference, Bureau of the Census, U.S. Department of Commerce.

Methods and Standards Committee (1990). Standards and Guidelines on the Presentation of Data in Tables of Statistical Publications, Statistics Canada.

Office of Management and Budget (1988). Guidelines for Federal Statistical Activities, Federal Register, Volume 53 Number 12, Office of the Federal Register, National Archives and Records Administration, Washington, D.C.

Program Evaluation Division (1988). Statistics Canada Labour Data: Identifying the Issues, Statistics Canada.
Shapiro, G., Pollock, J., Harley, D., and Beach, M.E. (1985). Statistical Standards at the Census Bureau: Past, Present, and Future, Proceedings of the First Annual Research Conference, Bureau of the Census, U.S. Department of Commerce.

Singh, M.P., Drew, J.D., and Choudhry, G.H. (1984). Post '81 Censal Redesign of the Canadian Labour Force Survey, Survey Methodology Joumal, 10, 2, 127-140, Statistics Canada.

Taillon, J. (1990). Monthly Gross Domestic Product by Industry: Quality Assessment, Gross Domestic Product by Industry, Catalogue 15-001, Statistics Canada.

# STATISTICS: THE BALANCE BETWEEN DESIRED PRECISION AND INTENDED USE 

B.L. Khuong ${ }^{\text { }}$


#### Abstract

This presentation attempts to describe the present situation at the BSQ which has developed as the result of continued resource cutbacks, and identifies a few possible avenues which BSQ statisticians should probably explore in order to meet more effectively the increasingly varied and demanding needs of their clients.

In the traditional approach, statisticians, with the collaboration of other specialists, attempt to anticipate information needs before they are actually expressed; in this context, it is usually the statistics themselves that give rise to the need for information.

However, over the past few years, particularly at the BSQ, resource cutbacks have brought in their wake a curtailment in new activities, in order to make it possible to concentrate available resources on the refinement of knowledge in fields that have already been covered. Paradoxically, generalized resource cutbacks generate a greater need for information. However, this time, the needs are created by the users, who are ready to pay, as long as the information generated meets their specific needs. In view of this, should statisticians change their approach in order to meet this new situation? Should they approach certain requests by directing their analysis toward the problems themselves and to the particular nature of the needs, rather than concentrate upon the more theoretical aspects of the techniques to be used?


KEY WORDS: Statistical products; Precision; "Optimal quality" statistics.

## 1. THE PROBLEM

Like other scientific disciplines, statistics has its own precision criteria; these require that it produce proof, based on real numbers, of its scientific approach and more specifically, of the precision of the numbers it produces. These notions of rigour and precision are imbued in the minds of statisticians during their university training.

When young graduates embark on their careers with the Bureau de la statistique du Québec, their anxicty about producing extremely precise statistics is even greater, due to the expectations of a large proportion of the public and the belief, founded or unfounded, that the figures produced by the statistics bureaus are almost infallible. Thus, their main concern is the development of tools or techniques that will allow them to constantly improve the precision of the statistics produced, regardless of the use to which they will be put.

Figure 1 is a diagram of the so-called "traditional approach," in which the main concern of statisticians is directed towards the theory and technique necessary for the production of figures that are as precise as possible.

[^40]Figure 1: TRADITIONAL APPROACH

## GROUP OF POTENTLAL USERS

- Government and public organizations
- Private organizations
- Researchers
- BSO specialists

ANTICIPATED NEEDS

DESIRED
PRECISION $\quad\left\{\begin{array}{l}\text { - budget allocated } \\ - \text { turn-around time } \\ - \text { physical capacity } \\ - \text { technical capacity } \\ \text { LEVEL } 2 \\ - \text { use of the results }\end{array}\right.$

This obsession with the techniques used to improve precision is not a bad thing in itself, because often university researchers and the major users in the public sector blame the central statistical organizations (BSQ, Statistics Canada, the BLS in the United States) for the questionable quality of some of their statistics.

In the present context of generalized budget reductions, can the same approach always be used to meet specific needs in a society that is constantly becoming more and more complex? My own personal observations indicate that we are faced with two types of needs, requiring two different approaches.

The first type of need concerns statistics for current use, which meet a certain demand, but are not precise enough for some of the specialized users. In such a context, the Bureau de la statistique du Québec, as the central statistical organization, is entirely justified in adopting the traditional approach; that is, one where precision is dictated by physical and technical capacity and the time and budget allocated, without regard for the use of these statistics. Under such circumstances, the main problem to be solved is to continually improve quality.

These concerns are already imbued in the minds of established statisticians; furthermore, they are provided for in BSQ programming. In addition, no nation can provide perfect statistics, since the use of these general-interest statistics varies from one individual to another, and changes over time automatically bring about a change in the level of precision required.

I have been at the BSO for about two years, and have often faced another type of need, the need to supply specific information for a particular decision. In this case, general-interest statistics, as precise as they may be, are not entirely satisfactory for specific requirements. Since the new Act respecting the Bureau de la statistique was adopted in 1987, the Bureau has also been authorized to meet the specific needs of users, providing that they agree to pay the costs incurred. In order to carry out this new task, the BSQ plays the role of a consultant, in addition to its usual function as the central statistical organization. In carrying out this additional role, the "traditional approach" runs into some difficulty in its efforts to meet the demand effectively.

The needs in question are generally for specific information required for making a particular decision, and require an answer in a short period of time. The following are two concrete examples, one from the field of surveys and the other from the area of data analysis.

1) The Ministry of Energy and Resources wanted to measure and analyze the demand for vacation resorts on public lands.

In order to estimate accurately the decision-making variable that, is, the proportion of households interested in vacation resorts on public lands in Quebec, the classic systematic approach to provide the data (as the central statistical organization) to the applicant would have been to conduct a survey involving a sample of $1,504,000$ households.

Taking into account the budget required to implement such a project, the cost of obtaining the information, and the time necessary, the optimal decision would be not to carry out the survey. What then should be done?
2) The Ministry of Health and Social Services (MHSS) wanted to establish a criterion that would allow it to compare the performances of various hospitals.

This criterion should be based on the average number of nights spent in the hospital for different operations. However, from time to time, an extreme value can excessively inflate the average number of nights in hospital. In order to eliminate this extreme value, the MHSS used a normal distribution. Excessive values were considered to be extreme, and were eliminated from the calculation. When the problem was presented to the statisticians at the BSQ, they expressed doubts as to the validity of using a normal distribution. Instinctively, they thought that a Poisson distribution would be more appropriate, since the number of nights resembled a service process for a "waiting line" phenomenon. After several tests were conducted, no distribution (Poisson, Beta or $\mathrm{X}^{2}$ ) corresponded precisely to the distribution of the observations according to the threshold set. Therefore, nothing could be done, although these distributions came closer to the real situation than a normal distribution, unless we were to invest an amount of time and money that would not be justified in view of the useful value of the results.

## 2. DESIRED LEVEL OF PRECISION

In both of the two cases mentioned above, the decision-makers needed information or tools to help them make a decision. However, setting the desired level of precision too high would probably mean that they would make their decision without this specific information, because of the high cost and time required to obtain a precision beyond the level required for the intended use. Is it preferable to have no information at all, or to have less precise information? Are there any alternatives? This leads me to ask another question on the quality of data. Must the level of quality of the data be consistent and static, regardless of the situation? In my opinion, it should be neither.

There is an optimal level of quality that can meet a given requirement, and this given requirement is dictated by the intended use. This intended use is specific and known with certainty in some cases, whereas in others it is not known. Can we then separate the two statistical products based on these two categories of intended use, in order to determine a threshold of quality that would define the desired level of precision?

## A. Products meant for non-specific users

These are statistical products that are meant for a large number of users and are or will be part of the programming of the Bureau. For this group of products, the level of quality must be consistent, since the products are intended for a wide range of users. But this level must not be static, it must evolve, in order to attain the highest level of precision possible. Statisticians must concentrate their efforts on the theory or technique that will allow them to improve precision, while taking into account the four constraints mentioned in the figure.

## B- Products meant for specific and limited use

For this second group of products, statisticians must emphasize the analysis of user problems, rather than concentrate upon the technique that will allow them to obtain data with an level of precision that would be identical to that of the products in the first group. Here the desired level of precision must not be consistent,
and depends particularly upon the specific intended use. In this case, statisticians "do statistical research" rather than "produce statistics." I need not explain the difference in meaning between statistics in the singular and statistics in the plural, since Messrs Fellegi and Wilk have described it very clearly in their article entitled "Is Statistics Singular or Plural?"

When statisticians get involved in the analysis of a problem, beyond the simple production of reliable statistics, they can determine in an "optimal" way, the level of precision desired, taking into account the intended use. They actively participate in evaluating the impact of the desired level of precision on reliability, and the applicability of the decisions made on the basis of the statistics provided. Thus, they do more than issue warnings; they themselves see to it that these warnings are heeded in the use of the results.

In the case of the products in group B, there is need for information for a well-defined and limited use. If, for reasons of precision regarding the desired level of quality, statisticians allow themselves to approach the problems from both a theoretical and a technical point of view, the decision-makers will obtain their information elsewhere. In my opinion, it is preferable to supply less reliable information, accompanied by an evaluation of its shortcomings, rather than leave it up to the decision-makers to use the information without any evaluation of these shortcomings.

Figure 2 below illustrates the so-called "ALTERNATIVE" approach, according to which the main interest of statisticians would be the solution to the problem or the particular nature of the use of the statistics.

Figure 2: THE "ALTERNATIVE" APPROACH


## 3. CONCLUSIONS

In theory, the notion of the quality of data does not present any apparent difficulty. In practise, it leads to such interesting and difficult questions as reconciling scientific requirements with the need to produce data that are applicable to a specific problem, in a short time, and at a reasonable cost. Determining the desired level of quality is a matter of evaluating the balance between the desired precision and the intended use. In all situations, there are two errors that must be avoided when designing the process to follow:

- looking for a level of quality in the data that is more precise than is required for the intended use; and
- aiming for a level of quality that is not precise enough for the intended use.

Thus, in order to produce "OPTIMAL QUALITY" statistics, statisticians must at least take into account the following three elements:

- they must avoid being only "PRODUCERS OF STATISTICS;"
- they must try to meet the user's NEEDS, rather than his demands;
- they must remember that intended use is the factor that DETERMINES the desired precision.


## REFERENCES

Bierman, H., Bonini, C., and Hausman, W. (1977). Quantitative Analysis for Business Decision (4th Ed.), Illinois, Richard D. IRWIN.

Deming, W.E (1990). Sample Design in Business Research, New York, John Wiley \& Sons.
Fellegi, I.P., and Wilk, M.B. (1988). Is Statistics Singular or Plural?, The Canadian Joumal of Statistics, 16.

## SESSION 10

Quality Assurance

# APPLYING INDUSTRY QUALITY IMPROVEMENT TECHNIQUES TO THE STATISTICS CANADA ENVIRONMENT 

D.N. Williams ${ }^{\prime}$


#### Abstract

Industry development strategies and techniques for quality improvement are very applicable to and effective in work environments such as Statistics Canada. The steps required to make this application, e.g. the definition of quality, measurement of quality, improvement of quality, are reviewed and discussed. This discussion places strong emphasis on the recognized differences between industry and the Statistics Canada work processes, e.g. services vs. production, short $v s$. long life-cycle processes, labour intensive $v s$. capital intensive (e.g. production equipment), professional vs. blue collar.


KEY WORDS: Customer satisfaction; Understand the process; Process control.

## 1. INTRODUCTION

Quality improvement is an over-whelming priority in industry today. Leaders from all corners will testify to how important it is. You can go into almost any department in any organization, ask how the quality improvement effort is going...
and be told "It doesn't apply to us... that's the stuff production is doing."
One of the greatest barriers to implementing quality improvement in a given organization is the strong belief that "it doesn't apply to us, we are different, we have already improved, that's production stuff". I have heard it one time or another from management and employees in just about every organization I have worked with, from Research and Development to Marketing to service and support operations.

But the organizations that are serious about quality improvement have discovered that it does not only apply on the production floor. That is where it started but what we discovered is that there is such an interdependence between all aspects of developing, producing and selling a product or service that everyone has to be involved. There is no one work area where quality improvement applies and one where is does not. It is a futile effort for the production floor to decrease the number of defects it creates if the design is so bad it won't work anyway, or if no one is going to buy the product, or if product is shipped to the place.

There is no cookbook for making quality improvement work. The implementation of quality improvement in a wide variety of different work environments has required different strategies and different techniques to be developed. But there are a few high level rules that always hold true. I would like to share these rules with you today and to reflect a bit on how these rules apply to Statistics Canada.

As we go through these you may recognize some of the rules, or you may realize that you are already applying one or more of them to the areas in which you work. This is good, these rules are not unique, nor are they a major creative work. Perhaps what makes them different for industry is that they have been clearly defined and that companies, at least the ones that are serious about this, are attempting to put them in place as a whole

[^41]throughout the organization. Not just one or two here and there, but as a block. They are far more effective used as a block.

## Rule \#1: Quality means customer satisfaction

Quality is defined in terms of customer satisfaction, doing what it takes to satisfy the customer, e.g. understand what it is they want and need, set their expectations properly and adequately, and then deliver to their expectations. Of course at Statistics Canada quality also has to be defined as meeting the stringent requirements which must be set in order to ensure integrity of the data you present. These internal requirements play a big part in defining the expectations that you set for your customers.

Some examples of proper setting of expectations:

- Honda Motor Corporation sells eight basic cars with minor variations. If some one goes to Honda and says, "Well, what I would really like is a V-8 'muscle machine', that seats eight and does zero to 90 kph in six seconds", Honda will gladly direct them to the nearest GM dealership. People do not expect that of Honda. People expect a Honda to be free of problems. If a Honda does have a problem, Honda has a dissatisfied customer. Setting an expectation and then not meeting it causes customers to get upset and, maybe, not come back to buy again.

Mitel had a very serious problem some years ago with properly setting customer expectations. There was a tendency to promise customers things that were not there, functions on its switching systems that did not exist. Customers were quite upset about that.

Quality does not necessarily mean telling the customer what they want to hear or trying to deliver to them everything that they could possibly think they want. Quality means setting expectations and then delivering to or exceeding those expectations.

Think about it, what are your customers' expectations of your area in Statistics Canada? Are their expectations well set, e.g. specific data presented in a few publications on a reasonable schedule, or do they expect all the data on earth published when they want it, usually very quickly? And then ask yourself, "how did their expectations get set?"

Setting expectations is not an easy thing to do. Industry spends millions of dollars a year trying to set expectations. Changing expectations is even harder, lowering expectations is even harder yet. Lee lacoca of Chrysler was very good at this and people responded. Others are not so good.

If you do not actively set a customer's expectations, they will set them for you. People have set their own expectations about the government for years. Now budget constraints have presented a situation where the government is having to seriously consider doing less. Resetting expectations is critical. It is not an easy nor popular thing to do.

But one fact holds true, if customers expectations are set and not met, they will be dissatisfied; they will be angry and they probably will not come back. With government they do not have a choice about coming back so they just get angry...and they say mean and nasty things about you.

## Rule \#2: Focus on the process

So how do you improve quality? That brings us to the second rule: The only way to effectively make permanent, on-going gains in quality is by improving the work processes, shift your focus to the process, not the people.

When I am educating work groups I like to ask them two very difficult questions: first to imagine the perfect work week and then to contrast it with the actual work week. Then I have them tell me what the differences are. The real work week is full of hassles, fire fighting, conflicting priorities, angry customers, missed deadlines,
and so on. On the average, people spend between $45 \%$ and $65 \%$ of their work time dealing with these things, call them waste. Not because they want to waste time. Waste is defined as the frivolous expenditure of time or materials. People are not responsible for the waste, they can not avoid these problems, it is the process that wastes this time.


As a rule of thumb, the causes for $85 \%$ of these problems reside in the process, $15 \%$ reside with the individual. By the process I am talking about:

- procedures,
- training,
- policies,
- equipment,
- facilities,
- incoming materials and information.

By the individual I mean individual differences, such as problems they bring from home, physical differences, health, intelligence, etc.

In the past companies have focused primarily on the individual to improve quality, through motivational programs, threats, money, etc. Slogan campaigns were initiated once a year cajoling employees to "Do it right the first time" or to "Go for Zero Defects." The main result was that morale went up when they took down the posters. Other companies tried bribing employees to do a better job, "if you'll work harder, better, etc., we'll give you MORE MONEY." The main result was that the employees tried harder for a bit but trying harder can not last for long. Increased attention spans can only be bought for so long. What they discovered is that if the process is not changed, improved, these actions do not make any long-term or cumulative gains.

This is not to say the people are not important in making things work, or in improving the process; they are essential. Which leads us to the next guideline:

## Rule \#3: People want to do a good job

This is the corollary to the above. If we remove the blocks that get in the way, people will do a good job. Companies have spent millions of dollars and hours of time trying to figure out how to raise motivation and productivity by focusing on the employee, what is being suggested here is that perhaps companies need to remove the demotivators that are in the process, the things that cause people to not care.

Certainly there are people that do not care, that will do a poor job. But can you blame them after years of the process getting in their way? And that nobody did much about it. By believing that people want to do a good job it will free us up to focus on the process.

A second part of this is that employees are the experts on the processes that they work with. To effectively improve quality, e.g. the process, you have to get the employees involved in the effort. Industries are realizing that they have to not only believe that their employees want to do a good job, but the employees should be trained to be an integral part of the improvement effort; so that they can actively work to get rid of the things that are getting in their way. Take advantage of your best resource, your employees.

## Rule \#4: Understand your process the very best you can

O.K., so we understand that we have to focus on improving the process. But how do we go about doing that?

I remember quite sometime ago having a three a.m. conversation with a rather unhappy manager that had just arrived from Japan to do an audit on an operation. He was saying "I don't know why you Americans can't figure this quality improvement stuff out by yourselves. You try quality circles and they don't work, you try suggestion plans and they don't work, you try robots and they don't work; when are you going to understand that quality improvement means understanding your processes the very best you can. If you truly understand your process, improvements will come."

I thought about that for awhile and my reaction was "No way, its more complex than that." I had been in this field for two or three years and it just was not that simple. But when I started looking at the various things I was teaching people, the tools and techniques, 1 realized that they were all aimed at giving them more information about the process.

All work is done through processes. If you do not have a process, you have anarchy. So in response to the challenge of "It doesn't apply to us" I would ask "is there any work process out there that could not be better understood?" By "understood", I mean understanding all aspects of the process, not just the process flow but the employees, the customers, the suppliers and the managers roles as well.

By customers and suppliers I mean internal as well as external customers. The process of designing and conducting even a relatively simple survey can involve a variety of divisions, all of them relying on each other for input and output in order for them to do their jobs. Without an understanding by all involved, the suppliers and the customers alike, of their roles, their input, and how the process works can employees really be expected to produce high quality work on schedule?


The two initial tools for understanding a process are process definition and monitoring. Flow charting is one of the most common tools for defining processes, but you have to go beyond just documenting the high lavel steps in the process, you have to add information about the inputs, the outputs and the requirements on tham as well as the overall requirements on the process. This information will begin to give you an idea of how processes interact, how much they rely on each other.

Getting agreement on all of this is difficult, processes are complex and involve many different areas. And once you do have a picture of a process you are faced with an ever more difficult question "Does this thing really work the way it is supposed to? Is it capable of meeting all of the requirements and expectations placed on it?" If the answer is "No" then you know where your improvement effort has to begin.

Monitoring is very important in two ways:

- if you do not monitor process quality, you can not improve it. This is to say that you will not know if something needs to be improved, and if changes are made you will not know if it has improved.
priority setting and maintenance. In industry they measure and carefully report productivity and financial data; and there are objectives against these measures. It is very hard to take quality improvement very seriously if it is not measured and carefully reported as well.


## Rule \#5: You can't improve a process that is out of control

This is a rule from statistical process control, what it means is that if a process is not predictablc and is inconsistent it can not be substantially improved. Now to me making a process consistent and predictable is improving it, but there are people that would differ with me.

The first step in any improvement effort has to be to achieve a reasonable degree of consistency and predictability in the way the process works, in the schedules kept, in the output that is delivered. This is not a simple step, research and development groups that 1 have worked with have spent a lot of time trying to accomplish that. It is important to note that meeting schedules consistently is only a piece of the challengc, making the production process consistent is perhaps a far more major task.

A good example of this occurred in a research and development group I was working with recently. The process seemed simple, R\&D was supposed to be given product specifications by marketing and then they designed a product to meet the specifications. But when we started looking closcly at the process we discovered that the information R\&D needed was delivered inconsistently, parts were usually missing, the format differed every time, and the accuracy of the data was poor. This resulted in a lot of rework, scrapped plans and upset customers. One of our first steps was to work with the marketing department to set mutually agrceable standards on this input. Now they can start figuring out how to meet these standards.

Do examples at Statistics Canada come to mind? Consider how hard it is to get consistent specifications on uscr requirements for questionnaire designs, or how frustrating it must be for computer programmer/analysts trying to write code for analysts that have not communicated their needs clearly.

## Rule \#6: The only way to permanently improve quality is to remove the causes of poor quality

Understanding the process has to include understanding the problems as well, understanding problems right down to understanding the root causes in the process that make the problems occur. Process improvement means permanently removing these causes, one after the other.

In the past, and for many in the present, the only way companies have tried to ensure and improve quality was to inspect the final product. If an error was found it was rejected, reworked and maybe someone got a black mark against them. This is an expensive habit to get into. You pay people to produce the work, people to find the mistakes, people to rework the error and put the product right.

In the service field we do not have that luxury because you can not rework an angry customer.

Rule \#7: Over 65\% of all solutions are lost $\qquad$


No matter how good they are, solutions can and do disappear. There are many reasons, too many to expand on today, but this rule has to be stated to underscore that just understanding the causes of a problem and developing a solution are not enough, the work is not done yet.

Proper implementation of solutions means doing what it takes to make them permanent.
Protect process improvements carefully. They can slip away without it ever being noticed.

## Rule \#8: Do it over again

To be effective, quality improvement has to be continuous. It should be a part of each person's job, part of their objectives. No one can be expected to everything right the first time, but they can be expected to always be working to improve the processes they work with.


When an improvement objective is met it means that we go looking for another one; it does not mean that we are done with quality improvement for the year or until we are ready to redesign the survey again in five years. It means that you constantly look at what the processes are that you are working with and their monitoring measures and try to make them work better.

## Rule \#9: If the managers do not lead, do not expect any one to follow

Management leadership and ownership of a quality improvement effort is essential. It can not be delegated.
So these are the rules that we try to apply to all areas of an organization.
As I have mentioned, there are whole set of tools and techniques for understanding and improving processes that I did not go into today. Many of them are statistically based, such as histograms, Pareto charts and control charts. Quality improvement is often confused with these tools, and it is true that the tools do not all work in all areas. But this does not relieve any one from the responsibility of going out and finding ones that do.

I would like to close not with a rule but with an attempt to answer the most often asked question about quality improvement, "What's in it for me?" This is a question that is easy to answer in industry. Often the answers go along the lines of "survival of the company" and "a job."

But in the government these responses do not necessarily follow. What is in it for you? I would say having more control over the processes you work with, increased pride in being able to do a job better and better, decreased stress from not having to deal with all of the hassles, fire fighting, and problems that we were talking about earlier, being able to produce the same amount of, or more, products and services, not from working harder but because you are not fighting the process every step of the way. And above all increased pride in delivering a better product, your own individual product or service as well as the critical information Statistics Canada supplies.


# WEIGHTED CUSUM SCHEMES AND THEIR APPLICATIONS 

E. Yashchin ${ }^{1}$


#### Abstract

We consider a class of weighted control charts that can be viewed as a generalization of the basic Cusum charting technique. The statistical power of this type of scheme is related to their efficient use of scores based on likelihood ratios. The weighted Cusum technique of type 1 proves to be especially useful when handling charts with varying sample sizes. The weighted Cusum of type 2 is used to enhance the statistical performance of a Cusum chart with respect to drifts in the process mean. In the present work we describe the graphical and decision making aspect of the technique as well as some applications related to monitoring of data quality. We also discuss some more general types of weighted schemes.


KEY WORDS: Average run length; Control charts; Exponentially weighted moving average; Likelihood ratio; Quality control.

## 1. INTRODUCTION

Let $\mathbf{X}_{1}, \mathbf{X}_{2}, \ldots$ be a sequence of observations related to a certain process, which may represent, for example, single measurements, sample means, sample standard deviations or sample proportions of errors found in successive data sets. In practical situations, such sequences are usually monitored by using control charts. The decision part of a control chart represents a set of criteria to judge, at any given moment of time, whether the process generating the observations is within acceptable variation. Such a set is called a control scheme. The criteria of performance of a control scheme are usually related to the behavior of some selected characteristics of its Run Length (RL), such as the Average Run Length (ARL) or some quantile of the Run Length. In most cases the scheme needs to be designed to ensure both a good sensitivity (i.e. short RL) with respect to undesirable patterns of incoming observations and a reasonable degree of protection against false alarms.

Superior decision making capabilities are necessary for any control charting technique to be recognized as suitable. However, these capabilities alone are frequently not sufficient. Of special value are techniques that also offer a meaningful and convenient graphical representation, which enables one not only to detect the presence of change in the behavior of the monitored data, but also to establish its point of origination and evaluate the current status of the controlled processes.

The charts used in industrial applications include the classical Shewhart chart, the Moving Average chart (simple or weighted, see Lai (1974), Nelson (1983)), Exponentially Weighted Moving Average chart (EWMA, see Roberts (1959, 1966), Bather (1963), Hunter (1987), Lucas and Saccucci (1990)), Cusum and Cusum - Shewhart chart (Page (1954), Barnard (1959), Lucas (1982), Yashchin (1985)), Girshik - Rubin chart (Girshik and Rubin (1952), Roberts (1966)), Quangles (North (1980)), Polyplots (Blazck, Novic and Scott (1987)) and some others. A summary of methods that usually lead to control schemes with good resolution can be found in Yashchin (1987). One of the most popular is Page's approach which is derived under the assumption that the on - target density of the observations is $f_{0}(x)$ and at some unknown moment of time it may switch to a "bad" density $f_{1}(x)$. This approach is based on the likelihood ratio considerations and calls for an out-of-control signal at time $T$ if

[^42]for some $l \geq 1$ the last $l$ observations $\left(X_{T-l+1}, T_{T-l+2}, \ldots, X_{T}\right.$ ) are "significant" in the sense of Sequential Probability Ratio Test (SPRT) i.e. if for some specified signal level $h$
\[

$$
\begin{equation*}
\sum_{i=T-1+1}^{T} \phi\left(X_{i}\right)>h \tag{1.1}
\end{equation*}
$$

\]

where $\phi\left(X_{i}\right)=\log \left[f_{1}\left(X_{i}\right) / f_{0}\left(X_{i}\right)\right]$ represents a score contributed by the $i$-th observation. In many situations, specifically, those involving exponential distribution families, the score function is linear in $\boldsymbol{X}_{i}$. For example, in the case, where $f_{0}(x)$ and $f_{1}(x)$ are both normal with a common standard deviation $\sigma$ and means $\mu_{0}$ and $\mu_{1}>\mu_{0}$, the score is

$$
\begin{equation*}
\log \frac{f_{1}\left(X_{i}\right)}{f_{0}\left(X_{i}\right)}=\frac{1}{\Delta}\left(X_{i}-k\right) ; k=\left(\mu_{0}+\mu_{1}\right) / 2, \Delta=\frac{\sigma^{2}}{\mu_{1}-\mu_{0}} . \tag{1.2}
\end{equation*}
$$

Clearly, in nonlinear situations one can simply use the transformation $Z=\phi(X)$. Therefore, we can work within the linear framework without loss of generality. The approach (1.1) is associated with a Cusum graphical representation which enables one to perform a "post-mortem" investigation which is likely to provide additional insight into the nature and magnitude of change as well as its origin. The Page's procedure was proven to possess certain optimality properties (eg. see Lorden (1971), Moustakides (1986), Banzal and Papantoni-Kazakos (1986), Ritov (1990)) and there are many reasons to believe that it is likely to play a dominant role in the area of process control.

In the present work we consider a chart which suggests associating weights with our sequence of observations and so generalizes the Cusum technique in a way analogous to that in which the weighted Moving Average technique generalizes the simple MA technique. In weighted schemes of type 1 (e.g. see Davies and Goldsmith (1972), Bissell (1973), Yashchin (1989)) we assume that the weights are either given in advance or become available together with the observations. In weighted schemes of type 2 the weights are essentially the parameters of a control scheme that enhance its capability of detecting drifts in the values of controlled parameters (see Yashchin (1989)). We shall also discuss schemes involving simultaneous use of both types of weighting.

In Section 2 we introduce the graphical formats associated with the weighted Cusum technique. In Section 3 we consider an example related to monitoring data integrity. In Sections 4 and 5 we discuss the weighted technique of type 2 and its connections to the generalized EWMA. Finally, in Section 6 we consider an example involving simultaneous control of a large number of parameters by using the Tunnel chart involving both types of weights.

## 2. DATA, PAGE'S AND CUSUM FORMATS FOR REPRESENTING WEIGHTED SERIAL DATA

Let us suppose that the controlled process parameter is the mean of the observations $\boldsymbol{X}_{1}, \boldsymbol{X}_{3}, \ldots$, and that the corresponding sequence of non-negative weights is $v_{1}, v_{2}, \ldots$. A typical example of such situation would correspond to the case where $X_{i}$ represents the proportion of defective items found in the $i$-th sample and $v_{i}$ is the corresponding sample size.

Let us suppose that we are primarily concerned about the possibility that the process might shift up towards an unacceptable level. To detect the presence of such conditions one could use a weighted upper Page's scheme defined in terms of three parameters: $h^{+} \geq 0$ (signal level), $k^{+}$(reference value) and $0 \leq s_{0}^{+} \leq h^{+}$(headstart). It is applied by starting from $s_{0}^{*}$ and then computing the sequence of cumulative sums

$$
\begin{equation*}
s_{i}^{*}=\max \left\{s_{i-1}^{*}+v_{i}\left(X_{i}-k^{*}\right), 0\right\}, i=1,2, \ldots \tag{2.1}
\end{equation*}
$$

If $T$ is the first index $i$ for which $s_{i}^{+}>h^{+}$, an out-of-control signal is triggered at time $T$.
If an additional signal criterion is introduced, which calls for a signal at the moment $i$ if a single observation $\boldsymbol{x}_{\boldsymbol{i}}$ exceeds $c^{+}\left(v_{i}\right)$, the procedure will be called an upper Page's scheme supplemented by a Shewhart's rule. In most applications the function $c^{+}\left(v_{i}\right)$ is decreasing in $v_{i}$; in the present work we shall not address the questions related to choice of this function in detail.

It is not difficult to see that when all the weights are the same the scheme (2.1) becomes a usual Cusum Shewhart scheme as described, for example, in Yashchio (1985). Moreover, in light of (1.2), the scheme (2.1) follows directly from (1.1) in the normal case with variable sample sizes. As usually, the reference value $\boldsymbol{k}^{+}$is chosen to be close to the midpoint between the acceptable and unacceptable levels of the process, while the signal level $h^{+}$characterizes the degree of accumulation of information allowed in the control scheme, i.e. represents the main instrument for achieving the desired trade-off between the degree of protection against false alarms and sensitivity requirements.

In a similar way one can define a lower weighted Page's scheme ( $s_{i}$ ) in order to detect changes in the process level downwards. As in the case of the usual Page's schemes, this can be achieved by applying an upper weighted Page's scheme with parameters $h^{-} \geq 0, k^{-}, 0 \leq s_{0} \leq h^{-}$(and, possibly, $c^{-}\left(v_{i}\right)$ ) to the sequence of "reflected" observations, $-\boldsymbol{X}_{1},-\boldsymbol{X}_{2}, \ldots$, i.e.

$$
\begin{equation*}
s_{i}=\max \left\{s_{i-1}+v_{i}\left(-X_{i}-k^{-}\right), 0\right\}, \quad i=1,2, \ldots, \tag{2.2}
\end{equation*}
$$

with a signal if $s_{i}^{-}>h^{-}$. In this case the reflected reference value, $\left(-k^{-}\right)$is typically chosen close to the "midway" between the acceptable and (lower) unacceptable process level. Finally, two-sided control can be achieved by combining an upper and lower schemes into a two-sided weighted Page's scheme.

Analogously to the case of usual Cusum-Shewhart schemes, weighted schemes can be applied in one of two graphical formats. To represent this data in a V - mask (Cusum) format one can choose some convenient centering constant $t_{0}$ (which may correspond, for example, to the target level of the process or to the average of all observations) and then plot the points $(0,0)$ and

$$
\left\{\sum_{j=1}^{i} v_{j}, \sum_{j=1}^{i} v_{j}\left(X_{j}-t_{0}\right)\right\}, \quad i=1,2, \ldots
$$

as shown in Fig. 2.1, bottom. Since the actual scale of the $y$ - axis is irrelevant in most applications we replace it by the scale of slopes corresponding to the protractor and computed under convention that the slope corresponding the horizontal line on the graph is equal to $t_{\theta}$; thus, if $t_{0}$ happens to be zero, our "generalized" slope becomes merely slope in the usual sense of this word. In our example the centering constant is chosen $10 t_{\theta}=0.1$ therefore the "flat" path of a Cusum trajectory corresponds to the level of observations being about 0.1 , as indicated on the vertical scale. If the level of a sequence $\left\{X_{i}\right\}$ is 0.02 above the target, the weighted Cusum trajectory will show a slope corresponding to the ray of the protractor marked 0.12 , and so on. It is clear that the "implicit" actual scale of the vertical axis can be easily reconstructed by using the known length of the horizontal ray of the protractor and its slopes. Further, by connecting two points of our trajectory and relating the (generalized) slope of the resulting segment to the protractor, one can obtain the value of a weighted average of observations corresponding to this segment, analogously to how it is done for classical Cusum paths. In the example where observations correspond to successive sample proportions of defectives and weights - to corresponding sample sizes, this slope is equal to the proportion of defectives within the segment.

As in the case of a classical Cusum, the weighted Cusum control scheme can be implemented via a V.mask, defined in terms of the signal level and reference value of the upper and/or lower schemes as shown on Fig. 2.1. Note that the reference values correspond to (generalized) slopes of the arms of the mask, expressed in terms
of the protractor, i.e. the slope of the horizontal ray is set to the centering constant $\ell_{0}$. For example, the mask shown on Fig. 2.1. corresponds to the two - sided control scheme ( $h^{+}=6, k^{+}=0.15, h^{-}=4, k^{-}=-0.06$ ) and an out-of-control signal is not triggered as long as the Cusum trajectory stays within the arms of the mask. As usual, one-sided control is achieved by applying an appropriate half-mask only. Finally, a supplemental Shewhart rule can be implemented by local "parabolization" of the mask by means of rays of generalized slopes $c^{+}\left(v_{i}\right)$ and $-c^{-}\left(v_{i}\right)$ exiting from the origin of the mask.

In the Page's format (Fig. 2.1, middle) the sequence $\left\{s_{i}{ }^{*}\right\}$ (and/or $\left\{s_{i}{ }^{\top}\right)$ ) is plotted against $i$ and assessed against the horizontal line corresponding to the signal level $h^{+}\left(h^{-}\right)$. This format of displaying serial data is relatively compact, since only information relevant to the control process is retained; on the other hand, it is closely associated with a particular control scheme which limits its value as a graphical tool. In Page's representation it is not necessary to use the non - homogeneous horizontal axis with respect to sequential numbers of observations. However, it still pays to do so, as it makes it easier to interpret the slopes and estimate the current level of the process and thus makes the chart more meaningful. Similarly, we scale the x - axis in accordance with the cumulative sum of weights to produce the data plot.

We refer to the above formats as W-Cusum, W-Page's and W-Data formats of displaying serial data. By examining the plots on Fig. 2.1 the reader will notice that, as in the case of classical Cusum procedures, using the weighted Cusum format enables one to assess the process level and detect the points of change much more easily than using the Shewhart (Data) format. One of the reasons for that is related to the fact that points based on a relatively small sample size (eg. points \#4 and \#13 on Fig. 2.1) introduce a substantial degree of confusion into the data plot and can in some cases prevent a change in the level from being noticed. At the same time, these points play no substantial role in the W-Cusum plot where they are assigned relatively small weights.

Figure 2.1: Data, Page's and Cusum formats of displaying weighted serial data. The upper horizontal scale represents the cumulative sum of weights. The lower scale corresponds to observation numbers.


## 3. WEIGHTED CUSUM OF TYPE 1: AN EXAMPLE

In this section we show how one can apply the weighted Cusum technique for monitoring the level of data integrity in a database. Let us suppose that the stream of hardcopy questionnaires related to buying patterns of "representative" consumers is arriving to the central office of a data processing agency, where they are entered into a computerized data base by a group of clerks. The data base management is interested in monitoring the proportion of erroneous entries related to various clerks. To do so, the quality control unit randomly extracts $5 \%$ of the processed records at the end of each day. The selected records are compared with the corresponding questionnaires and the number of detected errors is recorded. As a result, one obtains a data set (see Fig. 3.1) containing the date ( 1 -st column), the corresponding number of detected errors ( 2 -nd column) and the number of inspected records ( 3 -rd column). Such a set is maintained for each clerk. The sample sizes in each record are different, due to random sampling and differences in productivity among the clerks.

The automated statistical process control system employed by the agency is activated at the end of each day. Among other reports, the system produces plots of type shown in Fig. 3.1 for every clerk. The situation shown in Fig. 3.1 corresponds to the end of the 40 -th day of 1990 . It reflects the performance of a given clerk whom we shall name Alex. The nominal (historic) proportion of defectives is $p_{0}=0.01$ and the typical sample size is around $n=100$. By default, the defect rates are plotted in the $W$-data and $W$-Cusum formats, as shown in the right half of Fig. 3.1. In addition, the set of scores $z_{i}$ computed by means of the standard formula,

$$
\begin{equation*}
z_{i}=\frac{\hat{p}_{i}-p_{0}}{\sqrt{p_{0}\left(1-p_{0}\right) / n_{i}}}-\frac{\hat{p}_{i}-0.01}{0.99 / \sqrt{n_{i}}} \tag{3.1}
\end{equation*}
$$

is plotted in the upper left corner, in the W-data format.

Figure 3.1: The computerized report used for monitoring error rates of the data entry operation.


Dates. Efrors and Sample Sizes

| Dates. Errors and Somple Sizes |  |  |  |  |  |  |  |  |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 90001 | 1 | 99 | 90016 | 4 | 167 | 90031 | 2 | 145 |
| 90002 | 1 | 98 | 90017 | 0 | 91 | 90032 | 3 | 161 |
| 90003 | 0 | 104 | 90018 | 1 | 102 | 90033 | 2 | 161 |
| 90004 | 0 | 88 | 90019 | 0 | 107 | 90009 | 0 | 90 |
| 90005 | 3 | 145 | 90020 | 1 | 124 | 90035 | 2 | 144 |
| 90006 | 2 | 125 | 90021 | 4 | 145 | 90036 | 1 | 143 |
| 90007 | 0 | 117 | 90022 | 4 | 168 | 90037 | 4 | 138 |
| 90008 | 0 | 93 | 90023 | 1 | 116 | 90038 | 2 | 157 |
| 90034 | 5 | 148 | 90024 | 1 | 134 | 90039 | 2 | 151 |
| 90010 | 0 | 101 | 90025 | 2 | 157 | 90040 | 4 | 137 |
| 90011 | 1 | 104 | 90026 | 5 | 164 |  |  |  |
| 90012 | 1 | 102 | 90027 | 0 | 75 |  |  |  |
| 90013 | 1 | 118 | 90028 | 4 | 154 |  |  |  |
| 90014 | 1 | 113 | 90029 | 4 | 137 |  |  |  |



The error rates above $p_{i}=0.02$ are considered unacceptable and thus should be detected as soon as possible. To detect the presence of out-of-control conditions, the agency uses an upper weighted control scheme with the signal level $h=4.2$ and the reference value $k=0.015$. For the typical sample size, $n=100$, the rate of false alarms of this scheme is about 1 in 190 days. On the other hand, the condition corresponding to $p_{i}=0.02$ will be detected, on the average, after 8.5 days. In light of the likelihood-ratio origin of the weighted Cusum of type 1 , the performance of this scheme with respect to variable sample sizes can be expected to be roughly the same (e.g. see Yashchin (1989, Table 1)).

Shortly after the onset of an alarm (the 28 -th day of the year) the data base manager had a discussion with Alex about the importance of data integrity. At this point, it looked as if the Alex's error rate, which was in line with expectations for most of the time, has worsened substantially around the 20 -th day of the year, without any visible reason. At the point shown in Fig. 3.1, the Cusum chart clearly indicates that the Alex's rate of error is still around 0.02 , even though no new out-of-control signals were triggered by his chart. Note that both W-data charts corresponding to the rate of errors and scores do not provide a good evidence about the presence of change and its point of onset.

Looking at the plots, the data base manager realized that the lower scale symbols corresponding to the X -axis of the plot are somewhat non-uniform, suggesting the load increase in the period after the 20 -th day of the year. Indeed, in this period he was chronically missing two or more employees, primarily because of the "sick" season of the year. In the absence of these employees, their share of the work was shared among the remaining ones, in some cases requiring overtime work. The manager thus suspected that the quality problem of Alex might be related to his inability to handle highly variable loads of work. Thus, he asked the Quality Control unit to chart the relevant sample sizes and produce a scatterplot of the sample sizes vs. error rates. The resulting plots are shown in Fig. 3.2. They clearly show the increase in load for Alex that originated 20 days ago. The scatterplot also suggests that Alex's rate of errors tends to be much higher during the high load days. This conclusion is supported by several standard statistical tests, which will not be discussed in the present work.

Figure 32: The Cusum plot of the daily sample sizes related to a given clerk (left) which shows the increase in the workload. Plotting these sample sizes vs, the daily error rates (right) reveals the association between the workload and the level of data entry errors. The plotting symbols represent the sequential numbers of the points.


## 4. WEIGHTED SCHEMES OF TYPE 2.

In weighted schemes of the second type the weights are not associated with the data but rather represent parameters of the control scheme. In this case, weighting is used to emphasize the relative importance of the most recent observations. Let us suppose that we are given a set of positive weights $1=w_{0} \geq w_{1} \geq \ldots \geq 0$. Then a natural extension of the scheme (1.1) calls for triggering a signal at time $T$ if for some $l \geq 1$ the last $l$ observations $\left(X_{T-1+1}, X_{T-1+2}, \ldots, X_{T}\right)$ satisfy

$$
\begin{equation*}
\sum_{i=T-i+1}^{T} w_{T-\phi} \phi\left(X_{i}\right)>h \tag{4.1}
\end{equation*}
$$

Without loss of generality, we shall assume that the controlled parameter is the process mean and that the score function is linear in terms of $\boldsymbol{X}_{i}$. Then the control scheme calls for a signal at time $T$ if for some $l \geq 1$

$$
\begin{equation*}
\sum_{i=T-1+1}^{r} w_{T-i} v_{i}\left(X_{i}-k\right)>h \tag{4.2}
\end{equation*}
$$

where, as usual, $k$ is about midway between the "good" and "bad" process levels. The weights of type $1,\{v$, usually appear in a natural way as part of the score function $\phi\left(X_{i}\right)$, eg. they frequently represent sample sizes associated with $\boldsymbol{X}_{i}$. As shown in Yashchin (1989), this type of a scheme enables one to improve substantially the sensitivity with respect to drifts in the process mean without a major sacrifice in sensitivity with respect to shifts. The mentioned paper contains a discussion on applications of the weighted Cusum of type 2 , including the important case of multivariate process control.

The graphical representation of (4.2) looks exactly like in Fig. 2.1, except that the lengths of intervals on the $x$-axis will tend to increase. In other words, at time $T$ we can just plot the points $(0,0)$ and

$$
\begin{equation*}
\left(\sum_{j=1}^{i} w_{T-j} v_{j}, \sum_{j=1}^{i} w_{T-j} v_{j}\left(X_{j}-t_{0}\right)\right\}, i=1,2, \ldots, T \tag{4.3}
\end{equation*}
$$

and apply the V-mask $(h, k)$ at the last point.
In the present work we shall discuss the problem of representing the scheme (4.2) in Page's format. Indeed, the fact that the Cusum scheme (1.1) can be represented in the form (2.1) is extremely important in applications, mainly because it is not necessary to analyze, at point $i$, the whole set of preceding data to determine whether an out-of-control signal is to be triggered. The only quantities needed to reach the conclusion are the previous value of the scheme, $s_{i-1}$, and the most recent point, $\boldsymbol{X}_{i}$. This property enables the Cusum method to be used efficiently for monitoring large and intense volumes of dat a in real time. Furthermore, the Markov character of the Page's scheme enables one to design Cusum charts and analyze their statistical properties in a relatively simple way. A natural question is whether a similar representation exists for the weighted Cusum of type 2.

To answer this question, let us define the constants $Y_{1}, \gamma_{2}, \ldots$ on the basis of our sequence of weights $\left\{w_{j}\right\}$ by means of the following process:

$$
\begin{equation*}
Y_{0}=1, Y_{i}=w_{i}-w_{i-1} Y_{1}-w_{i-2} Y_{2}-\ldots-w_{1} Y_{i-1}, i=1,2, \ldots ; \tag{4,4}
\end{equation*}
$$

the first several members of this sequence are $\gamma_{1}=w_{1}, \gamma_{2}=w_{2}-w_{1}^{2}, \gamma_{3}=w_{3}-2 w_{1} w_{2}+w_{1}^{3}$, etc. Further, define the weighted upper Page's scheme of the second type in the following way:

Weighted Page's scheme, 2-nd type: Start with $s_{0}=0$ and compute the values of the sequence

$$
\begin{equation*}
s_{i}=\gamma_{1} s_{i-1}+\gamma_{2} s_{i-2}+\ldots+\gamma_{i} s_{0}+v_{i}\left(X_{i}-k\right), \quad i=1,2, \ldots ; \tag{4.5}
\end{equation*}
$$

until this process either exceeds $h$ and signals or becomes less or equal to 0 . In the latter case, re-set $s_{i}$ to 0 and start the process (4.5) from scratch, ignoring all the previous history.

Now we can prove the following
Theorem 4.1. The upper weighted Page's scheme of the second type as defined above and the scheme (4.2) are equivalent if and only if $\gamma_{i} \geq 0$ for every $i$.

Proof: We shall prove the theorem for the case $v_{i} \equiv 1, i=1,2, \ldots$. The modification of the proof to cover the general case is quite straightorward. Suppose that the Page's scheme triggered a signal at time \&T. and that the last time the scheme was re-started from 0 was $r$ observations ago, i.e. $s_{T-r}=0$. Then it is not difficult to see that the last $r$ values of the Page's scheme (4.5) can be obtained by solving the system:

$$
\left[\begin{array}{c}
s_{T}  \tag{4.6}\\
s_{T-1} \\
s_{T-2} \\
. \\
\cdot \\
s_{T-r+1}
\end{array}\right]=\left[\begin{array}{ccccccc}
1 & w_{1} & w_{2} & w_{3} & w_{4} & \cdots & w_{r-1} \\
0 & 1 & w_{1} & w_{2} & w_{3} & \ldots & w_{r-2} \\
0 & 0 & 1 & w_{1} & w_{2} & \ldots & w_{r-3} \\
. & . & . & . & . & . & \cdot \\
. & \cdot & . & . & . & . & \cdot \\
. & . & . & . & . & . & . \\
0 & 0 & . & . & . & 0 & 1
\end{array}\right] \times\left[\begin{array}{c}
X_{T-k} \\
X_{T-1}-k \\
X_{T-2}-k \\
. \\
. \\
\cdot \\
X_{T-r+1}-k
\end{array}\right]
$$

i.e. indeed, the scheme (4.2) must also trigger a signal at time $\boldsymbol{T}$.

Conversely, assume that the scheme (4.2) has triggered a signal at time $T$ with the last $r$ observations being "significant" and consider the values of the weighted Pages scheme (4.5). If the value of this scheme $r$ observations ago was 0 , then the relationship (4.6) shows that the Page's scheme should also signal at time T. If all the coefficients $\gamma_{i}$ are non-negative, this scheme should signal at time $T$ also if $s_{T-r}>0$, since under these conditions $s_{T}$ is an increasing function in all the previous values of the scheme. If, however, one of these coefficients is negative - one can find a realization of observations for which the weighted Page's scheme will not signal at time $\boldsymbol{T}$. Indeed, let us suppose that the first negative coefficient in the sequence $\left\{\boldsymbol{\gamma}_{\boldsymbol{J}}\right\}$ is $\boldsymbol{\gamma}_{3}$ and consider the following realization of the observations:

$$
\begin{equation*}
X_{1}-k=h, \quad X_{2}-k=\epsilon-\gamma_{1} h, \quad X_{3}-k=-\gamma_{1} \epsilon-\gamma_{2} h, X_{4}-k=h, \tag{4.7}
\end{equation*}
$$

where $\epsilon$ is some positive number not exceeding $h$. For this realization the Page's scheme is re-started after the third observation and an out-of-control signal is not triggered. On the other hand, since the weighted sum of the last two scores is

$$
\begin{equation*}
h+\gamma_{1}\left(-\gamma_{1} \in-\gamma_{2} h\right)>h, \tag{4.8}
\end{equation*}
$$

provided $\epsilon$ is small enough, the procedure (4.2) can lead to a signal at the last observation. It is easy to construct a realization similar to (4.7) and use the above argument for the case when the index of the first negative cocfficient in the sequence $\left\{\gamma_{\}}\right\}$is arbitrary; the details will be omitted.

It is not difficult to generalize the Theorem 4.1 for lower schemes as well as the case when the weighted upper scheme is supplemented by a headstart or a Shewhart limit, as suggested in Lucas (1982) and Lucas and Crosier (1982). In the first case the scheme (4.5) is initiated from some point $0<s_{0} \leq h$ rather than 0 . To supplement the scheme with a Shewhart limit $c\left(v_{i}\right)$ (i.e. an out-of-control signal is triggered if a single observation exceeds $c\left(v_{i}\right)$ ) we "parabolize" the V - mask by using a ray of slope $c\left(v_{i}\right)$ exiting from the origin of the mask.

Of special interest is the case of geometrically decreasing weights, $\left\langle w_{i}=\gamma, i=0,1,2, \ldots\right\rangle$ where $\gamma$ is usually between 0.7 and 1. In this case the scheme (4.2) is called a Geometric Cusum with parameter $\gamma$ and denoted $G C(\gamma)$. By (4.4), we have $\gamma_{1}=\gamma, \gamma_{i}=0, i=2,3, \ldots$, and the above theorem essentially reduces to an equivalence result for Geometric Cusums proven in Yashchin (1989)). The Markov character of the Geometric Cusum ensures a relatively painless analysis and implementation of this technique. In the generalized form, the $G C(\gamma)$ is represented as follows:

Start with $0 \leq s_{0} \leq h$, compute the sequence

$$
\begin{equation*}
s_{i}=\max \left(0, \gamma s_{i-1}+v_{i}\left(X_{i}-k\right)\right), \quad i=1,2, \ldots, \tag{4.9}
\end{equation*}
$$

and trigger a signal at time $T$ if $s_{T}>h$.
To design the above $C C(\gamma)$ one could simply assume that all the weights of type 1 are equal to the same "typical" value, $v_{i} \approx v$, and then find the ARL of the usual GC scheme

$$
\begin{equation*}
s_{t}=\max \left(0, \gamma s_{i-1}+X_{i}-k\right) \tag{4.10}
\end{equation*}
$$

with signal level $h / v$ and headstart $s_{0} / v$ eg. by using the tables in Yashchin (1989).

## 5. WEIGHTED CUSUM OF TYPE 2 AND THE GENERALIZED EWMA

Let us define the generalized Exponentially Weighted Moving Average scheme (G-EWMA) as follows:
Compute the sequence $\left\{\bar{s}_{i}\right\}$

$$
\begin{equation*}
\bar{s}_{i}=\frac{v_{i} X_{i}+\gamma v_{i-1} X_{i-1}+\ldots+\gamma^{i-1} v_{1} X_{1}}{v_{i}+\gamma v_{i-1}+\ldots+\gamma^{i-1} v_{1}} i=1,2, \ldots \tag{5.1}
\end{equation*}
$$

trigger a signal at time $T$ if $\left|\bar{s}_{T}-\operatorname{Target}\right|>L o\left(\bar{s}_{i}\right)$, where Target is the desired process level and $\sigma\left(\bar{s}_{j}\right)$ is the standard deviation of $\bar{s}_{i}$. The value of $L$ is selected so as to assure a low rate of false alarms.

The scheme (5.1) cannot be represented as a Markov Chain and, therefore, it is in general impossible to compute its ARL exactly. However, its statistical flaws make it anyway unsuitable as a main control tool, primarily because of the high inertia, as discussed in Yashchin (1987). We believe that the primary tool should be the two-sided $G C(\gamma)$. However, representing the G-EWMA on the chart, together with its control limits, enables one to assess graphically the state of the process and, therefore, the G-EWMA could well be used as a secondary control scheme, provided its on-target ARL is much larger than that of the primary scheme. The recommended values of $L$ are given in Table 5.1. These values were obtained by assuming in (5.1) that (a) $v_{i} v,(b) X_{i}$ are normal independent and identically distributed random variables and (c) the number of variables on which the average (5.1) is based is infinite. Under these assumptions the scheme (5.1) is equivalent to the usual EWMA (eg. see Roberts $(1959,1966)$ ), which starts from $\tilde{s}_{0}=$ Target and is run in the interval Target $\pm L \sigma(\tilde{S})$ by means of the relation

$$
\begin{equation*}
\tilde{s}_{i}=\gamma \tilde{s}_{i-1}+(1-\gamma) X_{i}, i=1,2, \ldots \tag{5.2}
\end{equation*}
$$

It is not difficult to see that to obtain the RL distribution of the above two-sided scheme it is sufficient to obtain a transition matrix of the upper $G C(\gamma)$ scheme (4.10) with respect to the signal level $h=2 L \sigma(s) /(1-\gamma)$ and $k=$ Target $-L \sigma(s)$, turn 0 into an absorbing state and compute the RL distribution associated with the resulting transition matrix for the headstart $s_{0}=h / 2$.

In many practical situations, the variance of $\boldsymbol{X}_{i}$ is inversely proportional to $\boldsymbol{v}_{i}$, i.e.

$$
\begin{equation*}
\sigma^{2}\left(X_{i}\right)=\sigma^{2} / v_{i}, \quad i=1,2, \ldots \tag{5.3}
\end{equation*}
$$

In this case $\sigma\left(\bar{s}_{i}\right)$ is given by

$$
\begin{equation*}
\sigma\left(\bar{s}_{i}\right)=\sigma \times r_{i} \text { where } r_{i}=\frac{\sqrt{v_{i}+\gamma^{2} v_{i-1}+\cdots+\gamma^{2(i-1)} v_{1}}}{v_{i}+\gamma v_{i-1}+\cdots+\gamma^{1-1} v_{1}} \tag{5.4}
\end{equation*}
$$

The value of can be estimated from a stable set of $n$ data points $\left\{\tilde{X}_{i}\right\}$ based on the weights $\left\{\tilde{v}_{i}\right\}$, eg. by using the formula

$$
\begin{equation*}
\hat{o}^{2}=\frac{1}{n-1} \sum_{i=1}^{n-1} \frac{\tilde{v}_{i} \tilde{v}_{i}+1}{\tilde{v}_{i}+\tilde{v}_{i+1}}\left(\bar{X}_{i+1}-\bar{X}_{i}\right)^{2} \tag{5.5}
\end{equation*}
$$

The formula (5.5) has the advantage of being robust with respect to possible changes in the process level present in the data. However, it results in biased estimates in the presence of serial correlation. In cases involving serial correlation, one could use the formula

$$
\begin{equation*}
\hat{\sigma}^{2}=\sum_{i=1}^{n} \frac{\left(\tilde{X}_{i}-\bar{X}\right)^{2}}{\left(n / \tilde{v}_{i}\right)-(1 / \bar{v})} \text { where } \bar{v}=\frac{1}{n} \sum_{i=1}^{n} \tilde{v}_{i}, \bar{X}=\frac{1}{n \bar{v}} \sum_{i=1}^{n} \tilde{X}_{i} \tilde{v}_{i} \tag{5.6}
\end{equation*}
$$

Both (5.5) and (5.6) reduce to well known formulas in the case of equal weights. Further, it is not difficult to see that the values of $\sigma\left(\bar{s}_{i}\right)$ are invariant with respect to the scale $v_{i}$. It is also worth mentioning that the above argument can be easily generalized to include more general, i.e. non geometric, weights of type 2 .

Table 5.1: The values of $L$ for which the corresponding EWMA scheme (5.2) has a specified ARL. These values closely approximate $L$ for the G-EWMA (5.1). It is assumed that $\left\{X_{i}\right\}$ are Normal with $\mu=$ Target and standard deviation $\sigma$. The control limits of (5.2) are Target $\pm \operatorname{Lo}(\bar{s})$ where $\sigma(\bar{s})=\sigma \sqrt{(1-\gamma) /(1+\gamma)}$.

| $\gamma$ | ARL $=50$ | 100 | 200 | 500 | 1000 | 2000 | 5000 | 10000 | 20000 | 50000 | 100000 |
| :--- | ---: | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| 0.7 | 2.17 | 2.45 | 2.71 | 3.02 | 3.24 | 3.44 | 3.69 | 3.87 | 4.04 | 4.25 | 4.41 |
| 0.75 | 2.12 | 2.41 | 2.68 | 3.00 | 3.22 | 3.42 | 3.67 | 3.85 | 4.03 | 4.24 | 4.40 |
| 0.8 | 2.05 | 2.36 | 2.64 | 2.96 | 3.19 | 3.40 | 3.65 | 3.84 | 4.01 | 4.23 | 4.39 |
| 0.85 | 1.96 | 2.28 | 2.57 | 2.91 | 3.14 | 3.36 | 3.62 | 3.81 | 3.98 | 4.21 | 4.37 |
| 0.9 | 1.81 | 2.15 | 2.45 | 2.81 | 3.06 | 3.29 | 3.56 | 3.75 | 3.93 | 4.16 | 4.33 |

## 6. WEIGHTED CUSUM OF TYPE 2: AN EXAMPLE

Let us consider once again the situation described in Section 3 and suppose that the data base manager is interested in a weekly report on the current status of the basic quality characteristics corresponding to the four sampling regions, Central, South, East and West. The summary data base for monitoring quality related to the Western region is shown in Table 6.1. Now we shall provide the interpretation of the last data row of this table, corresponding to June 18, 1990. The "target" or "historic" levels of the corresponding parameters are given in the bottom row of the table.

On June 18, a total of 16000 questionnaires corresponding to the Western region were received and processed by the data entry clerks (the "target" level of input is 20000 per week). Of these, 500 , i.e, $0.5 \times 1000$, were randomly selected by the Quality Control department for assessing the level of errors introduced in the process of data entry. The "target" level of inspection is $5 \%$, i.e. normally the department should have inspected about 800 forms. The proportion of errors found in these forms is referred to as "Our Error Rate" and represented in the table by the column "Error". Thus, $1.1 \%$ of the inspected questionnaires contained errors introduced by the data entry clerks. Further, $10.4 \%$ of the 16000 questionnaires contained missing values, $4.9 \%$ contained invalid data and $3.5 \%$ failed tests for data consistency, which revealed that the responses, though technically valid, were in logical contradiction. The response rate was $29 \%$. Finally, all 16000 of the questionnaires are subject to screening to prevent sampling bias. The management is especially interested in maintaining the proportion of respondents making more than $\$ 50,000$ in the Western region at about $20 \%$, the proportion of those who are 30 years or older at about $40 \%$ and the proportion of males at $50 \%$.

The serial data corresponding to the four regions of interest are represented, in the GC format with $\gamma=0.8$, in Fig. 6.1. Since the GC naturally occupies a sector region, we produce a pie-like set of charts with a common origin. This form of display is called a Tunnel chart (see Yashchin (1989)). Consider, for example, the plot for monitoring missing values in the Western region. To obtain this plot, we start at the origin of the chart and then plot the points (4.3) with $w_{i}=\gamma^{i}, i=0,1,2 \ldots$ and $v_{i}$ proportional to the entries in the Input column of the data. It is clear that the coefficient of proportionality has no effect on the appearance of the plot. So, let us suppose, for the sake of clarity, that $v_{i}$ is chosen by dividing the Input column of the Table 6.1 by its "target" value, 20 . This choice suggests that under "good" conditions the typical value of $v_{i}$ is 1 .

Now let us estimate the standard deviation of the values in the column corresponding to missing values. Let us assume that the variance of these values is $\sigma^{2} / v_{i}$ where $\sigma^{2}$ is the coefficient of proportionality. The value of $\sigma$ can be estimated, for example, by using the first $n=20$ points of the data in (5.5) to be $\hat{\sigma}=0.128$. This value is, however, somewhat lower than the "historic" level of $\sigma$, which is known to be about 0.15 . Under the assumption that $\sigma=0.15$ and the process mean is on its target level, 10 , the ARL of the two-sided GC( 0.8 ) scheme with the signal level $h=0.5$ and $k^{+}=10.1, k-=-9.9$ is approximately 1000 and thus we shall use this scheme to detect changes in the process level of magnitudes 0.2 and up. The corresponding V-mask is shown in Fig. 6.1. In addition, we show the control limits of the EWMA (5.1). The formula (5.4) indicates that the constant $r_{25}$ to be used at the current point is 0.341 . Thus, under the assumption that $\sigma=\hat{\sigma}=0.128$, we obtain that the estimated standard deviation of the EWMA is $0.128 \times 0.341=0.044$. With $L=3.84$ suggested by Table 5.1, the control limits $10 \pm 0.168$ of the EWMA ensure an on-target ARL of about 10000 .

The corresponding plot clearly indicates a change in the level of missing data that occurred 4 days ago. The process level prior to the point of change was about 9.8 , which can be established by simply connecting this point to the origin and estimating the resulting slope by using the protractor. Similarly, the process level corresponding to the last 4 points was about 10.3. It is important to realize that the tunnel chart also enables one to "see" the individual observations by evaluating slopes of the individual segments connecting the corresponding adjacent points.

In a similar way, we can monitor the remaining parameters. In Fig. 6.1 we do not show the control tools associated with these parameters. We shall only note that the central values of the protractors correspond to the target values in Table 6.1 and the scales are chosen in such a way that no trajectory can "run away" from the designated sector under nominal conditions. In general, the target values differ from one region to another. Among other "interesting" trends shown in Fig. 6.1, one can notice the drift in the level of inconsistent data (Central and West regions) which leads to "curved" W - Cusum trajectories and the drift of the data entry error
rate upwards in the South region. At the same time, the rate of similar errors in the Central region is getting much better and at the current time stands at about $0.8 \%$ of the processed questionnaires. This may suggest, for example, problems with the workload distribution, in light of the conclusions reached in the example of Section 3. Another interesting observation is related to the fact that the levels of missing data in the East and West regions show opposing trends: for example, at the point where the level of the missing data in the West region shifted up, the corresponding parameter in the East region shifted down. Finally, the current proportion of respondents who are 30 years or older in the East region is about 41.9 , which is significantly above the target level for this area.

Table 6.1: Integrity of data for the Western region. The bottom row of the table corresponds to the target values.

| West Region |  |  |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Date | Input | Inspec | Miss | Inval | Lacons | Error | Resp | $>50 \mathrm{k}$ | $>30 \mathrm{yr}$ | Males |
| 05/25/90 | 25 | 1.2 | 9.9 | 5.1 | 2.6 | 1 | 30 | 22 | 39 | 50 |
| 05/26/90 | 26 | 1.2 | 9.6 | 5.1 | 2.7 | 0.8 | 31 | 19 | 40 | 51 |
| 05/27/90 | 21 | 1.0 | 9.7 | 5.1 | 3 | 1.1 | 30 | 19 | 41 | 49 |
| 05/28/90 | 22 | 1.1 | 9.8 | 5.3 | 3.5 | 0.9 | 31 | 19 | 39 | 50 |
| 05/29/90 | 22 | 1.1 | 9.9 | 5.1 | 2.8 | 0.8 | 30 | 20 | 40 | 50 |
| 05/30/90 | 16 | 0.8 | 10.1 | 5.2 | 3 | 1 | 29 | 21 | 40 | 51 |
| 05/31/90 | 19 | 1.0 | 10.4 | 4.8 | 3 | 1 | 30 | 20 | 40 | 50 |
| 06/01/90 | 20 | 1.0 | 10.2 | 4.6 | 3 | 1.1 | 30 | 19 | 40 | 50 |
| 06/02/90 | 19 | 0.9 | 10.4 | 5 | 3.2 | 0.9 | 30 | 21 | 40 | 50 |
| 06/03/90 | 15 | 0.8 | 10.2 | 4.9 | 3.4 | 0.9 | 30 | 20 | 39 | 51 |
| 06/04/90 | 15 | 0.8 | 10 | 5.1 | 2.7 | 1.1 | 30 | 20 | 40 | 51 |
| 06/05/90 | 24 | 1.2 | 9.7 | 4.8 | 2.9 | 1 | 30 | 18 | 40 | 50 |
| 06/06/90 | 28 | 1.3 | 9.7 | 4.9 | 2.8 | 1.2 | 31 | 22 | 40 | 49 |
| 06/07/90 | 22 | 1.0 | 9.8 | 4.8 | 2.7 | 0.9 | 30 | 20 | 40 | So |
| 06/08/90 | 28 | 1.3 | 9.7 | 4.9 | 2.8 | 1.2 | 31 | 21 | 39 | 51 |
| 06/09/90 | 18 | 0.9 | 9.9 | 4.7 | 2.8 | 0.9 | 30 | 20 | 41 | 49 |
| 06/10/90 | 20 | 1.0 | 9.7 | 4.7 | 2.8 | 0.8 | 29 | 20 | 40 | 49 |
| 06/11/90 | 21 | 1.0 | 9.8 | 5.2 | 2.9 | 0.9 | 29 | 21 | 41 | 49 |
| 06/12/90 | 22 | 1.1 | 9.9 | 5 | 2.9 | 1.1 | 30 | 21 | 40 | \$1 |
| 06/13/90 | 15 | 0.7 | 9.8 | 4.9 | 3 | 1.2 | 30 | 21 | 38 | 50 |
| 06/14/90 | 11 | 0.5 | 9.6 | 4.9 | 3.1 | 0.9 | 31 | 23 | 39 | S1 |
| 06/15/90 | 25 | 0.9 | 10.3 | 5.1 | 3.2 | 0.9 | 30 | 20 | 40 | 50 |
| 06/16/90 | 19 | 0.7 | 10.2 | 5.2 | 3.3 | 0.9 | 29 | 20 | 40 | 50 |
| 06/17/90 | 17 | 0.5 | 10.3 | 4.7 | 3.4 | 0.9 | 30 | 19 | 42 | 49 |
| 06/18/90 | 16 | 0.5 | 10.4 | 4.9 | 3.5 | 1.1 | 29 | 20 | 40 | 50 |
| "Target" | 20 | 5\% | 10.0 | 5.0 | 3.0 | 1.0 | 30 | 20 | 40 | 50 |

The horizontal axis in the Tunnel chart has the property that if all the weights of type 1 are the same, $v_{i} \equiv \boldsymbol{p}$, the total length of the axis is $v\left(I+\gamma+\gamma^{2}+\ldots\right)=v /(I-\gamma)$. Therefore, the length of the trajectory being too short to reach the protractor scale indicates a possible undersampling. For example, the last four sample sizes for estimating the error rate in the Western region are substantially smaller compared to what is expected under nominal conditions and thus the resulting trajectory corresponding to this error rate is too short. By comparing the lengths of the segments between the adjacent points of this trajectory to the corresponding lengths for other trajectories, one can get a rough idea on the amount of undersampling. Similarly, oversampling will lead to somewhat longer paths, as is the case for the Central region. Note that, as indicated in the header of this plot, the nominal four-week level of input is 60000 questionnaires, while the actual number received was 65000 .

The author believes that in many practical situations, the weighted technique of type 2 and the associates tunnel chart enables one to display the status of a large number of parameters in a compact way, which permits a meaningful interpretation of the most recent trends and individual data points. This technique is especially useful in conjunction with the modern computing technology. For example, in the workstation environment, one can
not only display charts of type shown in Fig. 6.1, but also "run" these forward and backward in time by using computer animation, call up and remove the V-masks and control limits of the EWMA's, explore individual charts, points, related charts and tables - in other words, use the Tunnel chart as a starting point for exploratory and diagnostic work.

Figure 6.1: The Tunnel charts used to monitor the data integrity parameters. All the plots are based on the Geometric Cusum with $\gamma=0.8$ used together with the weights of type 1. The symmetric V-mask $h=0.5, k^{+}=10.1, k^{-}=-9.9$ is shown for the Missing data chart of the Western region only. The dotted vertical line on this chart gives the zone for the EWMA which, under the normal conditions, is violated ahout once in 10000 days. The weights of type 1 for the Error Rate correspond to the column "Inspec" of Table 6.1. For the remaining parameters, these weights correspond to the column "Input".


June 18, 1990


## ACKNOWLEDGEMENT

I would like to thank Dr. Betty J. Flehinger (IBM Research) for useful consultations and discussions on this subject and related examples.

## REFERENCES

Banzal, R.K., and Papantoni-Kazakos (1986). An Algorithm for Detecting a Change in a Stochastic Process, IEEE Tran. Information Theory, IT-32, 2, 227-235.

Barnard, G. (1959). Control Charts and Stochastic Processes, Joumal of the Royal Statistical Society, (B), 21, 239-257.

Bather, J. (1963). Control Charts and Minimization of Costs, Joumal of the Royal Statistical Society, (B), 25, 49-80,

Bissell, A.F. (1973). Process Monitoring With Variable Element Sizes, Applied Statistics, 22, 226-238.
Blazek, L.W., Novic, B., and Scott, D.M. (1987). Displaying Multivariate Data Using Polyplots, Joumal of Quality Technology, 19, 2, 69-74.

Davies, O.L., and Goldsmith, P.L., (ed.) (1972). Statistical Methods in Research and Production, Oliver \& Boyd, Edinburgh.

Girshick, M., and Rubin, H. (1952). A Bayes Approach to a Quality Control Model, Annals of Mathematical Statistics, 23, 114-125.

Hunter, J.S. (1986). The Exponentially Weighted Moving Average, Joumal of Quality Technology, 18, 4, 203-210.
Lai, T.L. (1974). Control Charts Based on Weighted Sums, The Annals of Statistics, 2, 1, 134-147.
Lorden, G. (1971). Procedures for Reacting to a Change in Distribution, Annals of Mathematical Statistics, 42, 1897-1908.

Lucas, J.M. (1982). Combined Shewhart-Cusum Quality Control Schemes, Joumal of Quality Technology, 14, 2, 51-59.

Lucas, J.M., and Crosier, R.B. (1982). Fast Initial Response for Cusum Quality Control Schemes: Give your Cusum a Head Start, Technometrics, 24, 199-205.

Lucas, J.M., and Saccucci, M.S. (1990). Exponentially Weighted Moving Average Control Schemes: Properties and Enhancements, Technometrics 32, 1-29 (with discussion).

Moustakides, G.V. (1986). Optimal Stopping Times for Detecting Changes in Distributions, The Annals of Statistics, 14, 2, 1379-1387.

Nelson, L.S. (1983). The Deceptiveness of Moving Averages, Joumal of Quality Technology, 15, 2, 99-100.
North, W.R.S. (1980). The Quangle - A Modification of the Cusum Chart, Applied Statistics, 31, 155-158.
Page, E. (1954). Continuous Inspection Schemes, Biometrika, 41, 100-115.
Ritov, Y. (1990). Decision Theoretic Optimality of the Cusum Procedure, The Annals of Statistics, 18, 3, 1464-1469.

Roberts, S. (1959). Control Chart Tests Based on Geometric Moving Averages, Technomertics, 1, 239-250.
Roberts, S. (1966). A Comparison of Some Control Chart Procedures, Technomerrics, 8, 411-430.
Woodward, R., and Goldsmith, P.L. (1964). Cumulative Sum Techniques, ICI Monograph 3, Oliver and Boyd, London.

Yashchin, E. (1985b). On a Unified Approach to the Analysis Of Two-sided Cumulative Sum Schemes With Headstarts, Advances Applied Probability, 17, 562-593.

Yashchin, E. (1987). Some Aspects of the Theory of Statistical Control Schemes, IBM Joum. Res. Devel., 31, 199-205.

Yashchin, E. (1989). Weighted Cusum Technique, Technometrics, 31, 321-338.

# TECHNIQUES TO CONTROL AND IMPROVE QUALITY OF DATA IN LARGE DATABASES 

R.W. Pautke and T.C. Redman ${ }^{1}$


#### Abstract

As information and service become increasingly important to the economy, the quality of data also becomes increasingly important. In this paper we first review evidence that suggests that current quality levels are quite low. Then we describe some techniques which have been developed and implemented at AT\&T Bell Labs and AT\&T's Network Services Division to achieve statistical control and make sustainable improvements. In particular, we describe data tracking, a technique developed to quantitatively evaluate the processes by which data are introduced into a database and then modified or transformed. We show how data tracking may be used to control and improve data accuracy and consistency and to accelerate processes. Throughout, we use the so-called "access process" to illustrate the main points.


KEY WORDS: Data tracking; Access process; Accuracy; Consistency; Currency.

## 1. INTRODUCTION AND SUMMARY

This paper deals with the quality of data in databases, especially databases used by large corporations and government agencies. Over the last several years data in databases has become critical to the functioning of many such enterprises. In particular, Svanks (1988) notes that data are often a corporation's most valuable resource. And there seems to be general agreement that poor data can be extremely costly. (Ballou, D.P. et al., 1989) (Lipiens, G.E., 1989) Yet, despite its importance, data quality has received relatively little attention in the quality and management information services literatures. In particular, the number of references to data quality levcls is low. At the same time, they do provide clear evidence that error rates, in some databases at least, are quitc high. Liepens, Garfinkel, and Kunnathur, (1982) for example, cite earlier literature (Pritzker, L. ef al.) (Terry, M.E., 1963) and estimate that the error rates are typically between one and ten percent. And more recent references suggest that in at least some applications the rates may be much higher. For example, Johnson et al, (1981) Laudon, (1986) and Morey, (1982) for a variety of applications, cite error rates in the $10-50 \%$ range. ${ }^{2}$ But astounding as these error rates are, they understate the true extent of the data quality problem because they concern only the accuracy dimension of data quality. These figures do not reflect inconsistencies in supposedly identical data items in overlapping databases, incompleteness - data omitted for whole segments of the relevant population, or data that is out-of-date.

Over the last couple of years we have had success using a technique we call data tracking to make sustainable improvements in data accuracy (Huh, Y.U. et al., 1990) (Huh, Y.U. et al.). Our work has stemmed from the observation that identifying and correcting errors in databases is extremely expensive, time-consuming, and often yields unsatisfactory results. The alternative is to focus on ongoing processes, which more or less continuously add ncw records, update existing records, and so forth. Only when such processes have been improved and put into a state of statistical control, can sustained improvements in data quality be expected. The essence of data tracking is to provide information about a process performance so that the necessary improvements can be made and control achieved.

[^43]tracking is to provide information about a process performance so that the necessary improvements can be made and control achieved.

In this paper, we consider data tracking more generally, and investigate ways that it can be extended to address data consistency and data currency. For the purposes of this paper, we define accuracy, consistency, and currency as follows:

1. Accuracy. Accuracy is a measure of agreement with an identified source. ${ }^{3}$
2. Consistency. Two sets of data are consistent if the correctness of one set implies the correctness of the other (note: there are forms of data consistency other than that selected for this paper. See Caby for details).
3. Currency. Data is current if it is up-to-date.

There are other dimensions of data quality and data consistency can be much more generally defined; see Levitin and Redman (Levitin, A.V. et al.) for a general discussion.

In Section 2 we review the basic ideas behind data tracking and describe one way to implement it. In Sections 3 and 4, we show how it is used to address each of the three problem areas (accuracy, consistency, and currency) noted above. Section 5 presents conclusions. Throughout this paper, we will use the so-called access process, by which AT\&T orders and receives access service from telephone companies, to illustrate the main points. Different aspects of the access process are shown in Figures 1-3. These Figures are used respectively to highlight key features of the process with respect to data accuracy, data consistency, and data currency.

Figure 1: The Access Process USO - Inventory DB


Legend
USO - Universal Service Order
ASR - Access Service Request
FO - Firm Order
LEC - Telephone Company Activities
FOC - Firm Order Confirmation
INV - Inventory Database
The access process is initiated when a customer orders service and a "universal service order" is created. An "access service request" is then created from the USO. When the ASR is complete, it is translated into an industry standard format and a "firm order" is sent to the "local exchange company". The LEC confirms that it will provide the service

[^44]Figure 2: The Access Process USO - Expense, Revenue DBs


USO $\quad-\quad$ Universal Service Order
ASR $-\quad$ Access Service Request
FO $-\quad$ Firm Order
LEC $-\quad$ Telephone Company Activities
FOC $-\quad$ Firm Order Confirmation
EXP $-\quad$ Expense Database
REV $-\quad$ Revenue Database
The access process also leads to data populating so-called "Expense" and "Revenue" databases, as shown here.

Figure 3: The Access Process: Currency Considerations

USO - Universal Service Order
ASR - Access Service Request
FO - Firm Order
LEC - Telephone Company Activities
FOC - Firm Order Confirmation
INV - Inventory Database

In this figure, the same activities as described in Figure 1 are shown. The lengths of the boxes and arrows, respectively, represent the time spent in value-added activities and in queues. The date/times $t_{e} t_{j} \ldots, t_{j o}$ indicate actual entry and exit date/times for the various subprocesses. Date/times $d_{1} d_{3} d_{5} d_{7} d_{9}$ and $d_{10}$ are nominal date/times when value-added activities are to be completed and $d_{p \varphi}$ is the nominal date/time the record is due in the Inventory database.

## 2. DATA TRACKING

One of the most basic ideas in the quality literature is that prevention of defects and/or errors is superior to defect/error detection and correction (i.e., to inspection and rework). In this regard a database is like a lake -to ensure a clean lake, one should first ensure that clean water flows in from upstream. This involves stopping the sources of pollution along the stream and/or ensuring that the sources of lake water are clean. This approach is fundamentally different from "cleaning the lake." For if the lake is cleaned up but the sources of pollution are left unchecked, the lake will once again become polluted.

A similar concept applies to a database and its associated infornation processes. To improve the quality of a database, one must first ensure that such processes enter quality data into the database.

The lake analogy extends further. If only clean water is introduced into the lake, the overall impact will be to cleanse it. This cleansing is facilitated by natural (biological, chemical, etc.) processes that are removing lake water the stream is replenishing. For databases, "churn" in the entities represented, the data values, and so forth has the same impact.

This analogy and the concepts of process management (Process Control Management and Improvement Guidelines) motivate the following high-level approach to improving the quality of data in a database:

1. Put processes which introduce data into the database into a state of statistical control.
2. If the controlled quality level does not meet requirements, improve the process until it does. Ensure that the improved process is under statistical control.
3. The quality of data that will be in the database at any time in the future is now predictable. If this quality level is not adequate, focus database clean-up efforts on portions of the database with relatively lower churn.

It is important to note the fundamental roles played by statistical control. First, statistical control makes quality levels (of both the database and processes) predictable. Only then can quality levels be meaningfully compared with requirements. Second, statistical control provides the foundation for and targets the improvement activities. The primary goal of data tracking is to establish statistical control.

Key steps in applying data tracking are as follows:

1. Take a random sample of records entering the process.
2. Track sampled records as they progress through the process. To do this, write out the contents of each sampled record as it enters and exits each subprocess. Also note the time of entry and exit.
3. Identify defects and errors produced by the process and its constituent subprocesses.
4. At appropriate intervals, summarize the progress of sampled records through the process. Develop relevant plots and summaries to verify conformance to standards and analyze changes or errors to reestablish control and suggest process improvements.

The remainder of this paper concerns implementation of these steps. In this section, we focus on Steps 1 and 2 , which concern the gathering of data tracking data.

### 2.1 Step 1: Sampling

Sampling is employed because it has many advantages and over complete inspection (Cochran, W.G., 1977). The details of the sampling scheme will depend on many factors, including the structure of the population to be sampled, the nature of the process, and the precision required. For example, many information processes are "batch" in the sense that groups of records enter the process together. In this case, a random sample of each
batch is appropriate. If records arrive "one-at-a-time," it may be more appropriate to sample each record independently of the others, with a pre-determined selection probability.

### 2.2 Step 2: Tracking

Once sampled, the records are marked (or "tagged") in a special way, so that their progress can be tracked through the process. This marking can be done in one of the following ways:

- Unique identifying field: If there is a data field, or a set of data fields, which uniquely identifies each record and this data field cannot through the process, it may be used for marking sampled records. To use this method, the values of this data field for the sampled records are stored for subsequent referencing. This method requires no change to the data structure.
- Flagged field: A special data field may be added to each record as a flag which identifies the record as sampled or not sampled. This method allows for quick, easy identification of the sampled records through the process. However, it requires a change of the data structure to include the flag. It is important that this flag be bidden from people who work within the process.

As the sampled records move through the process, their contents are recorded as each subprocess is initiated and completed, along with the date and time. The sampling and tracking can be accomplished manually or in any number of increasingly automated ways. Herein we briefly describe one way, which we call the decentralized image-capturing (DCI) system, to do so. The DCI can be applied when:

Flagged field marking, as described above, is feasible and

- Each step of the overall process is at least partially computerized. ${ }^{*}$

We consider this case because many important processes already meet these criteria and because we believe that more processes will meet them in the future.

The essence of the DCI is to build the necessary capabilities into the computer systems of each subprocess. These capabilities include:
a "sampler." This feature is built into the first subprocess and simply determines whether each record is to be tracked or not. In addition, the sampler puts the correct value into the "flag field."
"filters" which recognize sampled records and capture their contents at appropriate times. Such filters are built into each subprocess and capture the contents of the records when the subprocess is completed. The filters should also note the date/times when the subprocess is started and completed.
"communications" software, which sends the captured records to a secondary location. This software is also needed at each subprocess.

The secondary location associates common records from each subprocess together and stores them for analysis. An example of a record, tracked throughout the access process, is presented in Figure 4. In the next sections we describe analyses so that sustainable improvements, in terms of accuracy and consistency (Section 3) and currency (Section 4) can be made.

[^45]
## 3. ACCURACY AND CONSISTENCY

Data collected throughout the entire process is used to quantify accuracy, consistency, and currency. For accuracy, the emphasis is on changes in data fields as records progress into a single database. For consistency, the emphasis is on changes in data fields as processes diverge into separate databases. Conceptually, the data tracking analysis for accuracy and consistency is quite similar. The approach to analysis is discussed in this section.

For currency, the emphasis is on the date and time information. The approach for currency is discussed in Section 4.

### 3.1 Step 3: Identifying Errors

Many changes may occur in data fields that lead to inaccuracies and/or inconsistencies. Changes in data fields are categorized into the following three types:

- Normalizational: changes such as insertion or deletion of delimiters, spaces, etc. in order to meet differing format criteria used by different computer systems throughout the process. Examples of normalizational changes are shown on Figure 4 in data fields CKR and CKL at the ASR stage.
- Translational: changes due to the use of different languages throughout the process, such as the USO language used by AT\&T and the ISI language used in the telecommunications industry between interexchange carriers (e.g., AT\&T, MCI, Sprint) and Local Exchange Companies (e.g., NJ Bell, Il Bell). Examples of translational changes are shown on Figure 4 in data fields IX IND and S25 at the ASR stage.

Spurious-operational: changes that occur when one stage incorrectly changes a data item to a different value. Spurious-operational changes may be further classified as follows:
-- A correction of an error produced at an earlier stage.
-- An error in the stage at which the change occurred.
In either case, an spurious-operational change indicates an error in the process. Examples of spuriousoperational changes are shown in Figure 4 in data fields BAN and LSO SECLOC at the FOC stage.

### 3.2 Step 4: Summarizing Results

The collected data can be analyzed at three levels of detail, as follows:
the "metrics" level, where the goal is to determine the overall health of the process,
the "localization" level, where the goal is to determine subprocesses (more precisely, pairs of subprocesses) and field combinations that present the greatest problems, and

- the "control" level, where the goal is to establish statistical control and provide the platform for improvement.

Information from higher-level analyses is used to determine what lower-level analyses are needed and all analyses are aimed at helping the process owner ${ }^{5}$ determine what actions to take. In the next two subsections, we illustrate the types of plots we have found useful for doing so.

[^46]Figure 4：A＂Tracked＂Record

| 1100m | Imo． | $\cdots$ | $\bigcirc$ | mx | －man | － 0 ＋10m | MxM |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| － | secsman |  | Ocemam | O0，rome an | － | －6am | ase nemom |
|  |  |  |  |  |  |  |  |
| $\cdots$ |  |  | －nvas | monom | －urom |  | momen |
| ＝nm |  |  | ＊ | ， | ， |  | ！ |
| Leocu |  |  | nexa | smom． | mroce |  | mocen |
| － |  | $\cdots$ | $\cdots$ | $\cdots$ | ． |  | $\pm$ |
| 0 |  |  | nomer | $\cdots$ | － |  | cose |
| memor |  |  | －ame | come | mose |  | move |
| $\cdots$ |  |  | $\stackrel{*}{*}$ | $\cdots$ ． | $\cdots$ |  | nowe． |
| Ecios |  |  | aman mourco | coma nowacos | mandraros | cone mavop | morm name 0 om |
| $\cdots$ |  | － | ＊ | $\pm$ | $\cdots$ |  | － |
| Emo | － | 1 | 1 | 1 | 1. | － | 1 |
| ＊ |  |  | mmen | 315．menesen | memone | 873．801．8008 | ＊11．enction |
| （4）1／ |  |  |  |  | 星 |  |  |
| 4 | 3mm | mem | －${ }^{1}$ | mm | Ene | （nex | amo |
| 1 n | 成 | $\stackrel{1}{4}$ | ＊ | $\pm$ | $\ldots$ | anter | － |
| ct |  |  | nom | ne | $\underline{m}$ |  | nem |
| 4omac | $\pm 8$ | $=\pi$ | 0 m | 99080 | $\pm 8$ | Herz\％ | eos are |
| E1 | $\cdots$ | － | 1 | ＊ | ． | $\cdots$ | ＊ |
| \％ |  |  | －mimum |  | momenis |  | Noneme |
| a | $\square$ | 20el | － | － | － | $\cdots$ | － |
| $\pm$ |  |  | － | － | ＊＊ |  | －nc |
| 107w | －nm | 500m | mom 3m |  | Wiom om | － | vomem |
| 내에 | － | －\％omen | ＋om | mee 112 | － | ＊＊＊＊ | mos |
| 迷吅吅 |  | M迷 | $\cdots \mathrm{mm}$ | －xe 110 | － 0 베 |  | Nomo |

Changes to the record are highlighted in bold－face type．The change in the CKR field from the USO to ASR stage is an example of a normalizational change；the change In the S25 field from USO to ASR is a translational change；the change in LSO SECLOC from FO to FOC is a spurious－ operational change．

## 3．2．1 Accuracy

In this subsection，we focus on those aspects of the access process depicted in Figure 1.
Figure 5 is an example of a useful plot at the metrics level．In this plot，the average numbers of all three types of changes－normalizational，translational，and spurious－operational，are plotted against time．The root causes for normalizational and translational changes have been investigated and are varying format and computer language requirements across subprocesses．As these changes are less serious，${ }^{6}$ we will not pursue them further herein．We will concentrate on the spurious－operational changes．

Figure 5：Total Changes in the Access Process：USO－Inventory DB


The average number of each type of change per record is plotted against time for the USO－Inventory DB portion of the access process．

[^47]As stated, the goal of the localization stage is to determine which subprocess, field combinations offer the greatest opportunity for improvement. There are any number of ways to proceed. One way is to first plot the proportion of records experiencing spurious- operational changes for each data field on a Pareto chart. Figure 6 is such a plot and it confirms that the greatest opportunities for improvement lie with LSO SECLOC and BAN. ${ }^{7}$ We will concentrate on BAN, saving LSO SECLOC for discussion in the consistency section.

Figure 6: Spurious - Operational Changes: USO - Inventory DB


The percent of fields exhibiting spurious - operational changes throughout the USO Inventory DB portion of the access process.

Next, one could plot the percent of changes between subsequent steps for these fields. This can also be done on a Pareto chart. We have found utility in plotting such numbers directly on the process flowchart. The resultant chart, Figure 7, is referred to as a "process overview of change" chart.

Figure 7: The Process Overview of Changes: BAN


The percent of records for which BAN changes between subsequent steps in the USO Inventory DB portion of the access process. Here $z \gg x, y, w$ (all zero or very small).

[^48]In the third level of analyses, p-charts (Statistical Quality Control Handbook) are used. Figure 8 is the p-chart at the FO- FOC stage for BAN. On such charts, the solid center line (CL) is the average proportion defective for all the samples. The dotted lines above and below the center line are the upper and lower control limits (UCL and LCL), respectively. Control limits are used to determine, among other things, the out-of-control points, which are highlighted with star marks. The uneven control limits shown on these charts are due to unequal sample sizes. More information on selection of appropriate sample sizes, calculation of control limits, determination of the out-of-control points, and the statistical interpretation thereof, is available in texts such as Statistical Quality Control Handbook, or Wadsworth, Stephens, and Godfrey (1986). Control charts may be generated, along with related statistical analysis, using commercially available software tools, such as SQC Troubleshooter.

As noted, Figure 8 is the p-chart for changes in the data field BAN at the FO-FOC stage. The process exhibits a high average proportion defective. Figure 8 further shows many points out of control. In seeking a root cause, many plots against various variables were generated. Figure 9 shows a regional breakdown of the BAN changes. The plots shows considerable variation in BAN changes over various administrative regions. This variation proved to be the key to making substantial improvement - the details of how BAN was treated in Region 6 proved to be fully transportable to other regions and, in ensuing months, great improvements were made.

Figure 8: p-Chart: BAN Changes at FOC

p-chart: Changes in BAN between Firm Order and Firm Order Confirmation.

### 3.2.2 Consistency

We now shift our attention to those aspects of the access process depicted in Figure 2. The analysis procceds along much the same lines as for accuracy. First, at the metrics level, the total numbers of inconsistencies in each of the normalizational, translational, and spurious-operational change categories can be plotted against time. Figure 10 is such a plot. Again, we focus on the spurious-operational changes. At the localization level, a Pareto chart (Figure 11) and a process overview of changes chart (Figure 12) confirm that LSO SECLOC has the greatest number of inconsistencies and that the root cause lies at the FO-FOC interface.

Figure 9: Spurious - Operational BAN Changes at FOC by Region


The average number of spurious-operational BAN changes per record between Firm Order and Firm Order Confirmation for regions of the country.

Figure 10: Total Changes Between Expense, Revenue DBs


The average number of each type of change per record is plotted against time for the USO - Expense, Revenue DBs portion of the access process.

Figure 11: Spurious - Operational Changes Between Expense, Revenue DBs


The percent of fields exhibiting spurious-operational changes throughout the USO - Expense, Revenue DBs portion of the access process.

Figure 12: Process Overview of Changes: USO - Expense, Revenue DBs


USO - Universal Service Order
ASR - Access Service Request
FO - Firm Order
LEC - Telephone Company Activities
FOC - Firm Order Confirmation
EXP - Expense Database
REV - Revenue Database

The Process Overview of Changes Chart for LSO SECLOC for the USO - Expense, Revenue DBs portion of the access process. Here $c>z \gg b>a>x, y$.

Figure 13 is the p-chart for LSO SECLOC at the FO-FOC interface. This figure suggests that data field LSO SECLOC at the FO-FOC interface, appears to be "statistically in control," although the defect rate is high. That is, no special causes of variation seem to be present. (Although sample number 11 is marked as out-of-control, no special cause has been identified.) The apparently "random" behavior of this process suggests that this defect level is inherent to the process, as it is currently defined. This inference was subsequently confirmed to be true: as a result of a detailed study, it was found that LSO SECLOC is not transmitted from the FO subprocess to the FOC subprocess. Instead, LSO SECLOC is independently determined a second time as part of the FOC subprocess. In particular, with respect to LSO SECLOC the access process is discontinuous at the FO- FOC interface, as shown in Figure 14. As a result, whether or not the two LSO SECLOCs match is random, resulting in the behavior as noted on Figure 13.

Figure 13: p-Chart: LSO SECLOC Changes at FOC

p-chart: Changes in LSO SECLOC between Firm Order and Firm Order Confirmation. Point 11 is marked as out-of-control (it fails the "two of three beyond 20 test:"), but no special cause of variation could be found. Later analyses showed that LSO SECLOC is independently determined during the Firm Order and Firm Order confirmation subprocesses.

Figure 14: Functioning of Access Process for LSO SECLOC


The actual functioning of the USO - Expense, Revenue DBs portion of the access process for LSO SECLOC. The figure illustrates the discontinuity between the FOC and LEC activities for LSO SECLOC.

In order to eliminate this source of inconsistencies, the process owner must close the discontinuity.

## 4. CURRENCY

We now turn our attention to those aspects of the access process shown in Figure 3. Here the emphasis is on data currency. In particular, the focus is on the timeliness and the speed of the process.

One feature of the first step in the process (the USO step) is that a customer can specify (within certain limits) when be/she wants service made available. Using this information, the USO step schedules due dates/times for each of the value-added activities. As Figure 3 depicts, let $d_{s}$ denote the time the USO stage is to be completed, $d_{3}$ the time the ASR is to be completed, and so forth (i.e., due times). It is important to note that a customer can change his/her order right up until the service due date/time.

Data tracking can provide the dates/times that each of the activities are actually initiated and completed. In particular, as Figure 3 depicts, let $t_{\delta}$ denote the actual completion time of the USO stage, $t_{2}$ the actual start time of the ASR stage, and so forth.

Two types of analyses involving the actual and due date/times are of particular interest. The first is the process timeliness with respect to the due dates. For example, the timeliness of the delivery of the service to the customer is measured by $t_{10}-d_{10}$. Positive values (i.e., $t_{10}>d_{g 0}$ ) are of particular concern because they indicate that service is not available to customers when requested.

The other interesting analysis involves differences between successive actual times. The difference between the actual exit and entry times for a given activity indicate the "duration" of the activity. The difference between the actual entry time for a given activity and the exit time for the previous activity indicates "queue time."

Figure 15: Ideal Functioning of Access Process: Currency


Times
LEGEND
USO - Universal Service Order
ASR - Access Service Request
FO
LEC
FOC

- Trm Order
INY

With a new scheduler, all the queue time of the access process is located at the front end of the process. The subsequent activities are scheduled to be performed one after another and with the service is turned up when requested.

Timeliness, duration, and queue time data can once again be analyzed using the three-ticred analysis procedure previously described and illustrated. We do not repeat these analyses here. Instead, we note that, having analyzed "duration" and "queue time," one can schedule the process in a different way. Here, the goals are:

1. to eliminate as much inter-activity queue time as possible, putting it at the front end of the process instead.
2. Lo time the start of value-added activity such that the work is completed at the date/time requested by the customer.

The process would ideally function as shown in Figure 15.

## 5. CONCLUSIONS

This paper describes data tracking, a new technique which is proving useful in making sustainable improvements to the quality of data in large databases. The basic idea behind data tracking is to focus attention on so-called information processes, which introduce data into these databases. In data tracking, sampling is employed to mark a fraction of the records traversing the process. Sampled records are then tracked - that, is their contents are written out as each step of the process is completed. Subsequent analysis can reveal ways to improve data accuracy, data consistency between two databases, and the speed of the overall process (data currency). In addition, statistical control can be employed to ensure that improvements are sustained.

## ACKNOWLEDGEMENTS

The authors gratefully acknowledge the support and encouragement of their colleague and partners in AT\&T Bell Labs and AT\&T's Network Services Division. Steve Borbash, Errol Caby, Young Huh, Arnold Lent and Anany Levitin have contributed many ideas and helped flesh out others. And Michele Bentley, Darrell Link, Patti McKnight, June Slonin, Lynda Snyder and Beverly Weber, have put many of the ideas described in this paper into practice. There applications have led to numerous improvements in the methods described herein. The authors also wish to acknowledge the help and support of Monica Falkenthal, Kerin Montgomery, and Tim Neet. Lastly, the authors wish to thank Scott Williamson, whose vision made this work possible.

## REFERENCES

Ballou, D.P., and Tayi, G.K. (1989). Commun. ACM 32, 3, 320-329.
Caby, E. A Framework for Data Consistency, manuscript in preparation.
Cochran, W.G. (1977). Sampling Techniques, Third Edition, New York: John Wiley \& Sons.
Huh, Y.U., Keller, F.R., Redman, T.C., and Watkins, A.R. (1990). Data Quality, Information and Software Technology. 32, 8, 559-565.

Huh, Y.U., Pautke, R.W., and Redman, T.C. Data Quality Control, submitted to Journal of Quality Technology.
Johnson, J.R., Leitch, R.A., and Neter, J. (1981). Characteristics of Errors in Accounts Receivables and Inventory Audits, Account. Rev., 56, 2, 270-293.

Laudon, K.C. (1986). Data Quality and Due Process in Large Interorganizational Record Systems, Commun. ACM 29, 14-18.

Levitin, A.V., and Redman, T.C. The Notion of Data and Its Quality Dimensions, manuscript in preparation.
Liepens, G.E. (1989). Sound Data are a Sound Investment, Quality Progress, 22, 9, 61-64.
Liepens, G.E., Garfinkel, R.S., and Kunnathur, A.S. (1982). Error Localization for Erroneous Data: A Survey, TIMS/Studies in the Management Sciences, 19, 205-219.

Morey, R.C. (1982). Estimating and Improving the Quality of Information in a MIS, Commun. ACM 25, 5, 337 . 342.

Pritzker, L., Ogus, J., and Hansen, M.H. Computer Editing Methods - Some Applications and Results, Bulletin of the Intemational Statistical Instirute, 41, 442-465.

Process Control Management and Improvement Guidelines, Issue 1.1, (1988). AT\&T.
Shewhart, W.A. (1939). Statistical Method from the Viewpoint of Quality Control, Graduate School of the Department of Agriculture, Washington, D.C.

Statistical Quality Control Handbook, (1956). AT\&T.
SQC Troubleshooter, (1990). AT\&T.
Svanks, M.I. (1988). Integrity Analysis, Information \& Software Technology 30, 10, 595-605.
Terry, M.E. (1963). The Principles of Statistical Analysis Using Large Electronic Computers, Bulletin of the International Statistical Institute, 40, 547-552.

Wadsworth, H.M., Stephens, K.S., and Godfrey, A.B. (1986). Modem Methods for Qualiry Control and Improvement, New York: John Wilcy \& Sons.

SPECIAL INVITED LECTURE

# METHODS FOR ESTIMATING THE PRECISION OF SURVEY ESTIMATES WHEN IMPUTATION HAS BEEN USED 

C.-E. Särndal'


#### Abstract

In almost all large surveys, some form of imputation is used. This paper develops a method for variance estimation when single (as opposed to multiple) imputation is used to create a complete data set. Imputation will never reproduce the true values (except in truly exceptional cases). The total error of the survey estimate is viewed in this paper as the sum of sampling error and imputation error. Consequently, an overall variance is derived as the sum of a sampling variance and an imputation variance. The principal theme is the estimation of these two components, using the data after imputation, that is, the actually observed values and the imputed values. The approach is model assisted in the sense that the model implied by the impulation method and the randomization distribution used for sample selection will together determine the appearance of the variance estimators. The theoretical findings are confirmed by a Monte Carlo simulation.


KEY WORDS: Single value imputation; Variance estimation; Imputation model; Model assisted inference.

## 1. DIFFERENT TYPES OF IMPUTATION

This paper reports work carried out in connection with the development of Statistics Canada's Generalized Estimation System (GES). Variance estimates are roulinely calculated in the different estimation modules that define the GES, as described in Lavallee, Leblond and Reinhardt (1990). There was a need to develop suitable methods for variance estimation when the data set contains imputed values, which is the case in practically all surveys. I am indebted to M. Hidiroglou, P. Lavallée, Y. Leblond, H. Lee, and G. Reinhardt for their interest and collaboration in the work that led to this paper.

Two principal approaches to estimation with missing data are weighting and imputation. In the recent literature, the weights used to compensate for nonresponse are usually viewed as the inverse of the response probabilities associated with an assumed response mechanism. Since the response probabilities are ordinarily unknown, they need to be estimated from the available data. Imputation, on the other hand, has the advantage that it yields a complete data matrix. Such a matrix simplifies data handling, but it does not imply that "standard estimation methods" can be used directly. The imputed values are sample-based, thus they have their own statistical properties, such as a mean and a variance.

In our age, imputation is an extensively used tool. It is interesting to note that Pritzker, Ogus and Hansen (1965) say about imputation policy at the U.S. Bureau of the Census: "Basically our philosophy in connection with the problem of ... imputation is that we should get information by direct measurement on a very high proportion of the aggregates to be tabulated, with sufficient control on quality that almost any reasonable rule for ... imputation will yield substantially the same results ... With respect to imputation in censuses and sample surveys we have adopted a standard that says we have a low level of imputation, of the order of 1 or 2 percent, as a goal".

[^49]Ideally, we should still strive for the goal of only one to two percent imputation. But in our time most surveys carried out by large survey organizations show a rate of imputation that is much higher. Clearly, if $30 \%$ of the values are imputed, the effects of imputation can not be ignored. Imputation can create systematic error (bias) in the point estimate; this is perhaps the most series concern. But even if an imputation method can be found such that there is no appreciable systematic error, one must not ignore the often considerable effect that imputation has on the precision (the variance) of the point estimate. There is a need for simple yet valid variance estimation methods for survey data containing imputations, so that the coefficients of variation of the survey estimates can be properly reported.

A variety of imputation methods have been proposed in recent years. These can be classified in different ways. One way to classify is imputed for a missing value. A complete data matrix is obtained, in which the imputed values are flagged. Estimates are calculated with the aid of the completed set. In multiple imputation, two or more values are imputed for each missing value. Several completed data sets are thus obtained. Estimates are calculated with the aid of the completed data sets.

Imputation methods also differ with respect to the modelling underlying the imputation. Some imputation methods use an explicit model, as when the imputed value is obtained by a regression fit, a ratio or mean imputation. In other methods, the model is only implicit, as in hot deck imputation and nearest neighbour donor imputation. The distinctions just made are important for this paper.

Statistics Canada currently uses imputation methods such as nearest neighbour donor, current ratio, current mean, previous value, previous mean, auxiliary trend. All of these are single imputation methods. The imputed values originate in the Generalized Edit and Imputation System (GEIS), from where they enter into the Generalized Estimation System (GES), where the point estimates and the variance estimates are calculated in a number of different estimation modules. This paper deals in particular with current ratio imputation, which represents a case of explicit modelling.

## 2. SOME THOUGHTS ON MULTIPLE IMPUTATION

Multiple imputation was suggested by D.B. Rubin around 1977. His ideas are exposed in a number of papers, of which Herzog and Rubin (1983) and Rubin (1986) are expository, and in a book, Rubin (1987). Multiple imputation has advantages as well as disadvantages; the same is true for single imputation.

Rubin (1986) sees as a disadvantage of single imputation that "... the one imputed value cannot in itself represent uncertainty about which value to impute: If one value were really adequate, then that value was never missing. Hence, analyses that treat imputed values just like observed values generally systematically underestimate uncertainty, even assuming the precise reason for nonresponse are known."

Multiple imputation is attractive because it communicates the idea that imputation has variability. It is precisely this variability -- the variability within and between the several completed data sets -- that is exploited in the variance estimation methods proposed under multiple imputation. These methods make powerful use of basic statistical concepts. (On the other hand, one can argue that sample selection also has variability, but most surveys cannot afford more than a single sample, and estimation must be carried out with this unique sample.)

Simple examples show that treating imputed values just like observed values can lead to severe underestimation of the true uncertainty; survey samplers have long been aware of this. And it is a fact that users sometimes treat imputed values just like observed values, with wrong statement of precision as a result. With modern computers, it is easy to impute by some rule or another, but not so easy to obtain valid variance estimates.

The citation above seems to conclude that because a single imputed value does not display variation, we cannot obtain reasonable variance estimates; we are necessarily led to underestimation. I do not share this opinion. The methods that I discuss show that valid variance estimation is indeed possible with single imputation.

A method for variance estimation in the presence of imputed values should have the following properties: (a) a sound theoretical backing; (b) robustness to the assumptions underlying the imputation; (c) it must be practical, easy to carry out, and readily accepted by users.

While multiple imputation has the ingredients (a) and (b), it is clear that, in some applications at least, it does not have the property (c). In the development of GES we must depend on procedures that are easy to administer and easy to accept by the user. The user of a data set (someone who is not primarily a statistician) can easily understand that the statistician imputes once, with the objective to fill in the best possible value for one that is missing. While it is true that for some purposes, such as secondary analyses, it might be interesting to have several completed data matrices, the costs of storage of multiple data sets will often rule out this option.

Multiple imputation may well be useful in other contexts and for other reasons than those that are essential to the development of the GES. Multiple imputation methodology has shown us one way of bandling the problem of understatement of the variance when some of the data are imputed. Multiple imputation is not the only answer. Let us consider what can be done with single imputation methods. The method described below is based on Särndal (1990).

## 3. IMPUTATION VARIANCE AND SAMPLING VARIANCE

An imputation rule corresponds to an (explicit or implicit) model for the relationship among variables of interest to the survey. That is, when the analyst has fixed an imputation rule, he or she has in fact chosen a model. The principle for the developments that follow is that if this rule is considered good enough for the point estimates (no systematic error), the rule is also good enough for the corresponding estimates of variance. In other words, the model maker simultaneously takes responsibility for control of the bias and for the appropriateness of the variance estimates. If the model maker cannot live up to this challenge, he or she should not impute.

Let $U=\{1, \ldots, k, \ldots, N\}$ be a finite population; let $y$ denote one of the study variables in the survey. The objective is to estimate the population total of $y, t=\Sigma_{v} y_{k}$. (If $C$ is any set of population units, where $C=J$, $\Sigma_{C}$ is used as shorthand for $\Sigma_{k e C}$, for example, $t=\Sigma_{V} y_{k}$ means $\Sigma_{k e U} y_{k}$.) A probability sample $s$ is selected with a given sampling design. The inclusion probabilitics are known, and ordinary design-based variance estimates would be obtained if all units $k \in s$ are observed. However, there are missing data. Let $r$ be the subset of $s$ for which the values $y_{k}$ are actually observed. For the complement, $s-r$, imputations are calculated. The data after imputation consist of the values denoted $y_{* k}, k \in s$, such that

$$
y_{4 k}=\left\{\begin{array}{cl}
y_{k} & \text { if } k \in r \\
y_{i m p, k} & \text { if } k \in s-r
\end{array}\right.
$$

where $y_{k}$ is an actually observed value, and $y_{i m p, k}$ denotes the imputed value for the unit $k$. The case $r=s$ implies no imputation; all data are actual observations.

Let us write the estimator of $t$ that would be used if all data are actual observations (that is, $r=s$ ) as $\hat{t}=\Sigma_{k e s} w_{k} y_{k}=\sum_{s} w_{k} y_{k}$, where $w_{k}$ is the weight given to the observed value $y_{k}$. For example, in simple random sampling without replacement (SRSWOR) of $n$ units from $N, w_{k}=N / n$ for all $k \in s$ when the expanded sample mean is used to estimate $t$, and $w_{k}=\left(\overline{z_{U}} / \overline{z_{s}}\right)(N / n)=\left(\sum_{U} z_{k}\right) /\left(\Sigma_{s} z_{k}\right)$ for all $k \in s$ when the ratio estimator is used with $z$ as an auxiliary variable.

When the data contain imputations, the estimator of $t$ is $\hat{i}_{k}=\Sigma, w_{k} y_{* k}$. That is, we assume that the weights $w_{k}$ are identical to those used when all data are actual observations. This principle is used in the estimation modules of the GES. It embodies an assumption that imputation by the chosen rule causes little or no systematic error in the estimates.

The variance of an estimated total is increased by imputation, because imputation does not (except in truly exceptional circumstances) reproduce the true value $y_{k}$. Concrete evidence of this is the fact that if the imputation rule is applied to the actually observed sample units, there will always be error. If the rule is not without error for the responding units, it is not without error for the nonresponding units either. In Section 4 we express the variance of $\hat{\boldsymbol{f}}_{\mathrm{b}}$ as a sum of two components, a sampling variance, and a variance due to imputation,

$$
V_{\text {tot }}=V_{\text {sam }}+V_{\text {tmp }} .
$$

The imputation variance $V_{i m p}$ is zero if all data are actually observed values, or if the imputation procedure is capable of exactly reproducing the true value $y_{k}$ for every unit requiring imputation. (Neither case is likely in practice.) The procedure given in Section 4 uses the data after imputation, $y_{* k}, k \in s$, to obtain estimates of each of the two components, leading to

$$
\hat{V}_{\text {tox }}=\hat{V}_{\text {sam }}+\hat{V}_{i m p} .
$$

The component $\hat{V}_{\text {sam }}$ is calculated in two steps: (1) Compute the standard design-based variance estimate using the data after imputation. (For example, if SRSWOR is used, and $r=s$, the standard unbiased variance estimate of $N \bar{y}_{s}$ is $N^{2}(1 / n-1 / N) \Sigma_{s}\left(y_{k}-\bar{y}_{s}\right)^{2} /(n-1)$. This formula, calculated on the data after imputation, yields $N^{2}(1 / n-1 / N) \Sigma_{s}\left(y_{\phi k}-\overline{y_{\theta s}}\right)^{2} /(n-1)$, where $\overline{y_{\phi g}}$ is the mean of the $n$ values $y_{\phi k}$.) (2) Add a term to correct for the fact that many imputation rules give data with "less than natural" variability, which would lead to understatement of the sampling variance unless corrective action is taken. Finally, the component $\hat{V}_{\text {imp }}$ is readily computed from the data after imputation. The user will easily accept the argument that the variance obtained by the standard formula is not sufficient in itself; something must be added because the imputation rule is less than perfect.

The method has the good property that if no impulation is required, that is, $r=s$, then $\hat{V}_{\text {imp }}=0$ and $\hat{V}_{\text {sam }}$ equals the "standard variance estimator" that one would have used with $100 \%$ actually observed values.

## 4. THEORETICAL DEVELOPMENTS

The total error of $\hat{i}_{0}$ is decomposed as

$$
\hat{i}_{*}-t=(\hat{i}-t)+\left(\hat{t}_{*}-\hat{i}\right)=\text { sampling error }+ \text { imputation error. }
$$

The imputation error is the difference between the unknown estimate that would have been calculated if the data had consisted entirely of actual observations and the estimate that can be calculated on the data after imputation. The imputation error is

$$
\hat{i}_{*}-\hat{i}=-\Sigma_{s-r} w_{k} e_{k}
$$

where

$$
e_{k}=y_{k}-y_{i m p, k}
$$

is an imputation residual which cannot be observed for a unit $k \in s-r$. The magnitude of $e_{k}$ depends on how well the imputation model fits. The residuals are small if the imputation method gives nearly perfect substitute values. To pursue the argument, different directions may be taken. Here, we use a model assisted approach in which three different probability distributions are considered. The corresponding expectation symbols are written as $E_{\xi}, E_{s}$, and $E_{r}$. Here, $\xi$ indicates "with respect to the imputation model"; $s$ indicates "with respect to the
sampling design", and $r$ indicates "with respect to the response mechanism, given $s$ ". The model is implied by the imputation rule, so it is known; the sampling design is the given probability sampling distribution, so it is also known; the response mechanism is an ordinarily unknown distribution governing the response, given the sample $s$. The response mechanism need not be known for our method to work, and it can be one for which the response probability is systematically related to the $y$-variable, as in the case of non-ignorable mechanisms.

Assume that $\hat{f}_{4}$ is overall unbiased in the sense that $E_{\xi} E_{s} E_{F}\left(\hat{r}_{*}-t\right)=0$. The overall variance is then given by

$$
V_{\text {tot }}=E_{\xi} E_{s} E_{\gamma}\left\{\left(\hat{f}_{\psi}-r\right)^{2}\right\}
$$

This can be described as an "anticipated variance", for the given sampling design and response mechanism. We obtain

$$
\begin{align*}
V_{\text {ort }} & =V_{\ell s r}\left(\hat{i}_{t}\right) \\
& =E_{\ell} E_{s} E_{r}\left\{\left(\hat{i}_{t}-\hat{t}\right)^{2}\right\}  \tag{4.1}\\
& =E_{\ell} E_{s} E_{r}\left\{(\hat{i}-t)+\left(\hat{i}_{t}-\hat{i}\right)\right\}^{2} \\
& =E_{\xi} V_{s}+E_{s} E_{\gamma} V_{\S c}
\end{align*}
$$

where $V_{s}=E_{s}\{(\hat{t}-t)\}^{2}$ is the design-based variance of $\hat{t}$, supposing $\hat{t}$ is design unbiased for the total $t$. (For an estimator with some slight design bias, $V$, is the design-based mean square error of $\hat{i}$.) Note that ( $\hat{i}-\boldsymbol{i}$ ) depends on $s$ only, and not on r. Moreover,

$$
V_{\xi c}=E_{\xi}\left\{\left(\hat{r}_{\phi}-\hat{i}\right)^{2} \mid s, r\right\}
$$

is the model variance of the imputation error, given $s$ and $r$. The subscript $c$ stands for "conditional". The derivation of (4.1) rests on two assumptions: (1) The expectation $E_{\xi}$ can be moved inside $E_{g} E_{p}$; (2) the mixed term

$$
\begin{equation*}
2 E_{\xi} E_{s}\left[(\hat{t}-\hat{i}) E_{r}\left(\left(\hat{t}_{+}-\hat{i}\right) \mid s\right\}\right] \tag{4.2}
\end{equation*}
$$

vanishes or is sufficiently close to zero that we can ignore it. This condition would be fulfilled, for example, if the expected imputation error is zero or negligible under the response mechanism, conditionally on the realized probability sample $s$. Even if ( 4.2 ) is not exactly zero for the mechanism that actually determines the response to the survey, we can in many cases approximate (4.2) by zero and still use the method below to obtain a variance estimate that is much better than pretending naively that imputed data are as good as actually observed data.

If we set $V_{\text {sam }}=E_{q} V_{s}$ and $V_{i m p}=E_{s} E_{r} V_{g c}$ in (4.1), then

$$
V_{\mathrm{tot}}=V_{\mathrm{sam}}+V_{\mathrm{imp}}
$$

or

$$
\text { overall variance }=\text { sampling variance }+ \text { imputation variance. }
$$

The objective is to estimate the overall variance, so that a valid confidence interval for the unknown $t$ can be calculated. Our approach is to obtain separate estimates, $\hat{V}_{m a m}$ and $\hat{V}_{i m p}$, of the two components $V_{s m m}=E_{g} V_{g}$ and $V_{i \mu \psi}=E_{g} E_{g} V_{i c}$. The data available for this estimation are $y_{* k}, k \in s$.

The argument for obtaining $\hat{V}_{\text {sem }}$ and $\hat{V}_{\text {inv }}$ is as follows:
(i) Estimation of the sampling variance component. Let $\hat{V}$, be the standard (design-unbiased or nearly design-unbiased) estimator of the design variance $V_{s}$. Denote by $\hat{V}_{*}$, the quantity obtained by calculating $\hat{V}_{s}$ from the data after imputation, $y_{* k} k \in S$. For many imputation rules, $\hat{V}_{* s}$ underestimates $V_{\text {seam }}$. The underestimation is compensated in the following way. Evaluate the expectation

$$
E_{\xi}\left(\hat{V}_{s}-\hat{V}_{t z}\right)=V_{d v} .
$$

Then find a model unbiased estimator, denoted $\hat{V}_{\text {dr }}$, of $V_{\text {dp }}$. This may require the estimation of certain parameters of the model $\xi$. Consequently,

$$
E_{\xi}\left(\hat{V}_{d v}\right)=E_{\xi}\left(\hat{V}_{s}-\hat{V}_{t s}\right) .
$$

Then

$$
\hat{V}_{s m}=\hat{V}_{\phi s}+\hat{V}_{d t}
$$

is overall unbiased for the component $\boldsymbol{V}_{\text {sam }}=E_{\xi} V_{s}$, as the following derivation shows:

$$
\begin{aligned}
E_{\S} E_{s} E_{r}\left(\hat{V}_{\text {sem }}\right) & =E_{s} E_{r}\left\{E_{\xi}\left(\hat{V}_{t s}\right)+E_{q}\left(\hat{V}_{\text {dif }}\right)\right\} \\
& =E_{s} E_{r}\left\{E_{q}\left(\hat{D}_{s}\right)\right\}=E_{\xi} E_{s}\left(\hat{V}_{s}\right) \\
& =E_{\ell} V_{s}=V_{s a m} .
\end{aligned}
$$

(ii) Estimation of the imputation variance component. Simply find an estimator, $\hat{V}_{\text {f }}$, that is model unbiased for $\boldsymbol{V}_{\xi c}$. That is, $E_{\xi}\left(\hat{V}_{\xi c}\right)=V_{\xi c}$. Again, this may require the estimation of unknown parameters of the model $\xi$. Then $\hat{V}_{i c}$ is overall unbiased for the imputation variance component $V_{i m p}$, since

$$
E_{s} E_{r} E_{\xi}\left(\hat{V}_{\xi c}\right)=E_{s} E_{r} E_{\xi c}=V_{i m p} .
$$

Finally, an overall unbiased estimator of $\boldsymbol{V}_{\text {bot }}$ is given by

$$
\hat{V}_{\text {bot }}=\hat{V}_{\text {sam }}+\hat{V}_{\text {isup }}
$$

where $\hat{V}_{\text {sam }}=\hat{V}_{\text {es }}+\hat{V}_{\text {dif }}$ and $\hat{V}_{\text {imp }}=\hat{V}_{\text {bc }}$. Note that the role of $\hat{V}_{\text {dif }}$ is to correct for the fact that the data after imputation may display "less than natural" variation. This often happens when $y_{i m p, k}$ equals the predicted value from a fitted regression, that is, "the value on the line". The variation around the line is not relected in the predicted value.

It is an advantage of this variance estimation method that it requires only minimal assumptions about the relation of the response propensity to the study variable $y$. If the chosen imputation model holds approximately, the method works even if there is a systematic relationship, for example, that units with large values $y_{k}$ are less likely to respond.

EXAMPLE. The sample $s$ is drawn with SRSWOR; $n$ units are drawn from $N$. Denote the number of respondents by $m$. The respondent mean is imputed for units requiring imputation. That is, $y_{\theta k}=y_{k}$ if $k \in r$ and $y_{* k}=\bar{y}_{r}$ if $k \in s-r$, and $\hat{f}_{t}=(N / n) \sum_{s} y_{* k}=N \bar{y}_{p}$. Then $\hat{V}_{* s}=N^{2}(1 / n-1 / N)((m-1) /(n-1)\} S_{y r}^{2}$,
$\hat{V}_{d i f}=N^{2}(1 / n-1 / N)\{(n-m) /(n-1)\} S_{y r}^{2}$, and $\hat{V}_{i m p}=N^{2}(1 / m-1 / n) S_{y m}^{2}$, where $\quad S_{y}^{2}=\sum_{r}$ $\left(y_{k}-\bar{y}_{p}\right)^{2} /(m-1)$. Thus, $\hat{V}_{s o m}=N^{2}(1 / n-1 / N) S_{y r}^{2}$, and $\hat{V}_{p o t}=N^{2}(1 / m-1 / N) S_{y r}^{2}$. The following table shows the contribution of the three terms to the total $\hat{V}_{\text {wo }}$, for different rates of imputation, assuming that $N$ is large compared to $m$ and $n$, and $(m-1) / m \approx(n-1) / n=1$.

| imputation rate in \% | $\%$ contribution to $\hat{V}_{\text {not }}$ |  |  |
| :---: | :---: | :---: | :---: |
| $\mathbf{1 0 0 ( 1 - \mathrm { m } / \mathrm { n } )}$ | $\hat{V}_{\text {es }}$ | $\hat{V}_{\text {dj }}$ | $\hat{V}_{\text {imp }}$ |
| 10 | 81 | 9 | 10 |
| 20 | 64 | 16 | 20 |
| 30 | 49 | 21 | 30 |

The table illustrates the dangers of acting as if imputations are real data: with $30 \% \mathrm{impute}$ values, the standard variance estimator $\hat{V}_{\phi s}$ in this example covers less than half of the correctly estimated total variance. Imputation by the respondent mean is useful as an example; the results are particularly simple. But usually in practice, respondent mean imputation is neither justified nor efficient. It is not justified because the underlying model is not sophisticated enough to avoid systematic error in the point estimates. It is usually inefficient because the residuals $e_{k}$ are large.

## 5. APPLICATION TO IMPUTATION BY THE CURRENT RATIO METHOD

The method assumes that a positive auxiliary value $x_{k}$ is known for every unit $k \in s$. If $k \in s-r$, we impute $y_{\text {imp.k }}=\hat{B} x_{k}$ with $\hat{B}=\left(\sum_{r} y_{k}\right) /\left(\sum_{r} x_{k}\right)$. The data after imputation are

$$
y_{* k}=\left\{\begin{array}{cl}
y_{k} & \text { if } k \in r \\
\hat{B} x_{k} & \text { if } k \in s-r
\end{array}\right.
$$

The model behind current ratio imputation is

$$
\begin{equation*}
y_{k}=\beta x_{k}+\epsilon_{k} \tag{5.1}
\end{equation*}
$$

where the $\epsilon_{k}$ are uncorrelated model errors such that

$$
\begin{equation*}
E_{\xi}\left(\epsilon_{k}\right)=0, \quad V_{\xi}\left(\epsilon_{k}\right)=\sigma^{2} x_{k} \tag{5.2}
\end{equation*}
$$

Suppose that the sample $s$ is selected by SRSWOR. Let the respective sizes of $s, r$, and $s-r$ be $n$, $m$, and $n-m$. If no imputation was needed, the estimator of $t=\sum_{U} y_{k}$ would be $\hat{t}=N \bar{y}_{s}$. Using the data after imputation, we get

$$
\begin{equation*}
\hat{t}_{t}=(N / n) \sum_{s} y_{* k}=N \bar{x}_{s} \bar{y}_{r} / \bar{x}_{r} \tag{5.3}
\end{equation*}
$$

(Overbar and subscript $s, r$, or $s-r$ indicates "straight mean", for example, $\overline{y_{r}}=\sum_{r} y_{k} / m, \overline{x_{g-r}}=\sum_{g, r} x_{k} /(n-m)$, etc. $)$. Using the results of the preceding section, we have $V_{\text {sor }}=V_{s a m}+V_{i m p}$ with $V_{s a m}=E_{\xi}\left\{N^{2}(1 / n-1 / N) S_{y U}^{2}\right\}$ and $V_{i e q}=E_{s} E_{r}\left\{N^{2}(1 / m-1 / n) C_{1} \sigma^{2}\right\}$, where
$S_{y U}^{2}=\Sigma_{U}\left(y_{k}-\overline{y_{U}}\right)^{2} /(N-1)$ and $C_{1}=\overline{x_{g}} \overline{x_{s-r}} / \bar{x}_{g}$, a known constant. The mixed term (4.1) is exactly zero in this application. Our method of variance estimation gives $\hat{V}_{\text {rot }}=\hat{V}_{\text {sam }}+\hat{V}_{\text {imp }}$, where

$$
\begin{gather*}
\hat{V}_{s m m}=N^{2}(1 / n-1 / N)\left\{S_{y * s}^{2}+C_{0} \hat{\sigma}^{2}\right\}  \tag{5.4}\\
\hat{V}_{\text {imp }}=N^{2}(1 / m-1 / n) C_{1} \hat{\sigma}^{2} \tag{5.5}
\end{gather*}
$$

where $S_{y * s}^{2}=\Sigma_{s}\left(y_{* k}-\overline{y_{4 s}}\right)^{2} /(n-1)$ is the variance calculated on data after imputation, and we have chosen to estimate $\sigma^{2}$ by the model unbiased formula

$$
\hat{\sigma}^{2}=\frac{1}{\bar{x}_{r}\left\{1-\frac{1}{m}\left(c v_{x}\right)^{2}\right\}} \frac{\sum_{r}\left(y_{k}-\hat{B} x_{k}\right)^{2}}{m-1}
$$

where $c \nu_{x r}=S_{x} / \bar{x}_{r}$ is the coefficient of variation of $x$ in the response set $r$. The constant $C_{0}$ is obtained as

$$
C_{0}=\frac{1}{\sigma^{2}} E_{\xi}\left(S_{\chi s}^{2}-S_{y{ }^{2} s}^{2}\right)
$$

where

$$
s_{y s}^{2}=\frac{1}{n-1} \sum_{s}\left(y_{k}-\bar{y}_{s}\right)^{2}
$$

is the (unknown) sample variance based on data with $100 \%$ actual observations. After evaluation,

$$
C_{0}=\frac{1}{n-1}\left\{\Sigma_{s-r} x_{k}-\frac{\Sigma_{s-r} x_{k}^{2}}{\sum_{r} x_{k}}+\frac{1}{n} \frac{\Sigma_{s-r} x_{k} \Sigma_{s} x_{k}}{\Sigma_{r} x_{k}}\right\} .
$$

If $m$ is not too small, the approximations $\hat{\sigma}^{2} \approx\left(\sum_{r} e_{k}^{2}\right) /\left(\sum_{r} x_{k}\right)$ with $e_{k}=y_{k}-\hat{B} x_{k}$ and $C_{0} \approx(1-m / n) \bar{x}_{g-r}$ are sufficiently good for most applications.

We can write the imputation variance component as

$$
\hat{V}_{i m p}=N^{2}(1 / m-1 / n) A \bar{x}_{s} \hat{\sigma}^{2}
$$

where $A=\bar{x}_{s-r} / \bar{x}_{r}$. The constant $A$ reflects the selection effect due to nonresponse. If large units are less inclined to respond than small units, then $A$ may be considerably greater than unity, and, for a given a sample $s$ and a given number $m$ of respondents, the component $\hat{V}_{i m p}$ tends to be large, relative to a case where, say, all units are equally likely to respond. This tendency makes good sense intuitively.

Two special cases are noted: (1) If all $x_{k}=1$, the estimated total variance becomes simply

$$
\hat{V}_{\text {tox }}=\hat{V}_{\text {sem }}+\hat{V}_{\text {inpp }}=N^{2}(1 / m-1 / N) S_{y r}^{2}
$$

where $S_{y r}^{2}$ is the variance of the $m$ actual observations $y_{k}$. This agrees with the variance obtained under a twophase sampling design with SRSWOR in each phase. (2) If no imputation is required, that is, if $s=r$, then $\hat{V}_{\text {ive }}=0$, and

$$
\hat{V}_{\text {voc }}=\hat{V}_{s a m}+N^{2}(1 / n-1 / N) S_{y e}^{2}
$$

That is, our method yields the well known variance estimator for SRSWOR.

A Monte Carlo study with 100,000 repeated response sets $r$ was carried out to confirm the above results for current ratio imputation. A finite population of size $N=100$ was generated according to the model consisting of (5.1) and (5.2). The typical response set $r$ was obtained as follows: Draw a SRSWOR sample $s$ of size $n=30$; given $s$, generate $r$ by a response mechanism in the form of independent Bernoulli trials, one for each $k \in s$, with probability $\theta_{k}$ for the outcome "response". Three different response mechanisms were used: Mechanism 1: $\theta_{k}$ increases with $y_{k}$ in such a way that $\theta_{k}=1-\exp \left(-a_{1} y_{k}\right) ;$ Mechanism 2: $\theta_{k}$ increases as $y_{k}$ decreases in such a way that $\theta_{k}=\exp \left(-a_{2} y_{k}\right)$; Mechanism 3: $\theta_{k}$ is constant at 0.7 . The constants $a_{1}$ and $a_{2}$ in the first two response mechanisms (which can be described as non-ignorable) were fixed to obtain an average response probability of 0.7 . The sizes of the realized response sets $r$ thus varied around a mean of 21 for all three mechanisms. For each $r$, the point estimate $\hat{t}_{\text {, given by }}(5.3)$ was calculated as well as three different variance estimators, $\hat{V}=\hat{V}\left(\hat{t}_{\phi}\right)$. These were: (1) the model assisted variance estimator $\hat{V}_{\text {mot }}=\hat{V}_{\text {sam }}+\hat{V}_{\text {tmp }}$ equal to the total of (5.4) and (5.5); (2) the two-phase sampling variance estimator $N^{2}(1 / n-1 / N) S_{y r}^{2}+N^{2}(1 / m-1 / n) \Sigma_{r} e_{k}^{2} /(m-1)$, an estimator which follows from standard two-phase sampling theory with an assumption of SRSWOR subsampling of $m$ respondents from the $n$ units in the initial sample (Rao, 1990); and (3) the standard unadjusted variance estimator $N^{2}(1 / n-1 / N) S_{y+s}^{2}$ obtained by acting as if imputations are as good as actual data. The results are shown in the following table.

|  | Relative bias of $\hat{\boldsymbol{V}}$ in \% |  |  |
| :--- | :---: | :---: | :---: |
| Estimator $\hat{\boldsymbol{V}}$ | Mechanism 1 | Mechanism 2 | Mechanism 3 |
| Model assisted | -0.20 | -4.64 | -3.99 |
| Two-phase | 9.95 | -12.49 | -1.11 |
| Standard unadiusted | -2573 | -3790 | 33.21 |

 mean of the 100,000 values of $\hat{V}$ and $\operatorname{var}\left(\hat{i}_{\theta}\right)$ is the variance of the 100,000 values of $\hat{i}_{\theta}$. The simulation confirms that the model assisted method gives nearly umbiased variance estimation regardless of the response mechanism. This is not surprising because our population conforms to the model underlying the simulation. The two-phase estimator works well for the response mechanism 3 ("data missing at random"); otherwise it is biased. Finally, to act as if imputed data are as good as actual data leads to a dramatic understatement of the true variance for all three mechanisms.

## 6. IMPUTED VALUES THAT HAVE AN ADDED RESIDUAL

We can distinguish two types of imputed values: (1) the imputed value $y_{i m p, k}$ consists of a predicted value only, $y_{\text {prod, },}$, as when the value on a fitted regression line or surface is used. For example in the current ratio
imputation method as used above, $y_{i m p, k}=y_{\text {prod,k }}=\hat{B} x_{k}$ with $\hat{B}=\left(\Sigma_{r} y_{k}\right) /\left(\Sigma_{r} x_{k}\right)$; (2) the imputed value $y_{\text {imp,k }}$ consists of a predicted value and a residual, so that $y_{i m p, k}=y_{\text {prod,k }}+e_{k}^{\dot{k}}$. The residual term, whose purpose is to make imputed values more like actual observations, may be obtained by sampling the residuals $e_{k}=y_{k}-y_{\text {pred } k}$ calculated for the responding units $k \in r$. A scheme for this is given below. This type of imputation is sometimes recommended in the literature as a means of preserving the distributions of the imputed data; see, for example, the discussion in Little (1988). The imputation process then requires more effort to complete, and for the purposes of the GES (whose principal aim is valid estimation of the precision of survey estimates), it is not clear that the advantages gained are worth the extra effort.

Let us, however, indicate one scheme for imputation by "predicted value plus residual" in the case where the current ratio imputation model is taken as the point of departure: For $k \in r$, calculate $e_{k}=y_{k}-\hat{B} x_{k}$ with $\hat{B}=\left(\sum_{r} y_{k}\right) /\left(\Sigma_{r} x_{k}\right)$, then $\tilde{e}_{k}=e_{k} / \sqrt{x_{k}}$. This gives a supply of $m$ "standardized residuals" $\tilde{e}_{k}$. Then for a unit $k \in s-r$, calculate $e_{k}^{0}=\sqrt{x_{k}} \tilde{e}_{k}$, where $\tilde{e}_{k}$ is drawn by SRSWR from the supply, and $x_{k}$ belongs to the unit requiring imputation. Then large $x$-value units tend to obtain larger residuals $e_{k}^{0}$, which is consistent with the model. Then set $e_{k}^{0}=e_{k}^{0}-\left(\sum_{s-r} e_{k}^{0}\right) /(n-m)$. For $k \in s-r$, impute $y_{\text {imp,k }}=\hat{B} x_{k}+e_{k}^{*}, k \in s-r ;$ for $k \in r$, we have actual observations, $y_{k}$. Since the $e_{k}^{*}$ were made to sum to zero over $s-r$, the point estimator is given by $\hat{i}_{+}=(N / n) \sum_{z} y_{* k}=N \bar{x}_{g} \bar{y}_{r} \mid \bar{x}_{r}$ as in Section 4, but its variance is different. It can be shown that $E_{\S} E_{s} E_{q} E_{\xi}\left(S_{y+s}^{2}-S_{y}^{2}\right) \approx 0$, where $E_{q}$ denotes average with respect to the random selection of a standardized residual. That is, the difference between the variance calculated on data after imputation, $S_{y, s}^{2}$, and the unknown variance of a sample consisting entirely of actual observations, $S_{y s}^{2}$, is approximately zero on the average. We can use $\hat{V}_{\text {sam }}=N^{2}(1 / n-1 / N) S_{y+s}^{2}$ as an approximately overall unbiased estimator of the sampling variance component. There is no need now to add a correction $\hat{V}_{\text {dr }}$. However, an estimator of "e imputation variance $V_{\text {imp }}=N^{2}(1 / m-1 / n) C_{1} \sigma^{2}$ must still be calculated and added to $\hat{V}_{\text {sam }}$.

## 7. CONCLUDING REMARKS

The continued work on the variance estimation techniques outlined in this paper has the following objectives: (1) study of the robustness of the variance estimators in Section 4, that is, the sensitivity to breakdown of the assumptions of the imputation model; (2) extensions to imputation procedures based on models that are implicit only, in particular the nearest neighbour donor method; (3) extensions to the case where there is a mixture of several imputation procedures in the same survey; (4) extensions to other types of estimators than those considered in the example in this paper. A forthcoming publication will contain results for the case where the Horwitz-Thompson estimator, $\hat{i}=\Sigma_{s} y_{k} / \pi_{k}$, serves as the prototype. The estimator using data after imputation is then

$$
\begin{aligned}
\hat{t}_{t} & =\sum_{r} y_{k} / \pi_{k}+\left(\sum_{s-r} x_{k} / \pi_{k}\right)^{\prime} \hat{B} \\
& =\sum_{s} y_{k} / \pi_{k}-\sum_{s-r} e_{k} / \pi_{k}
\end{aligned}
$$

where $e_{k}=y_{k}-x_{k}{ }_{k} \hat{B}$ is the imputation residual for unit $k$.

## REFERENCES

Herzog, T.N., and Rubin, D.B. (1983). Using Multiple Imputations to Handle Nonresponse in Surveys. In W.G. Madow, I. Olkin and D.B. Rubin (eds.), Incomplete Data in Sample Surveys. New York: Academic Press, 209-245.

Lavallée, P., Leblond, Y., and Reinhardt, G. (1990). Generalized Estimation Systems, statement of requirements. Report, Business Survey Methods Division, Statistics Canada.

Little, RJ.A. (1988). Missing-Data Adjustments in Large Surveys (with discussion). Joumal of Business and Economic Statistics, 6, 287-301.

Pritzker, L., Ogus, J., and Hansen, M.H. (1965). Computer editing methods: some applications and results. Bulletin of the Intemational Statistical Institute, 41, 1, 442-466.

Rao, J.N.K. (1990). Variance Estimation Under Imputation for Missing Data. Manuscript seen by courtesy of the author.

Rubin, D.B. (1986). Basic Ideas of Multiple Imputation for Nonresponse. Survey Methodology, 12, 37-47.
Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: Wiley \& Sons.
Särndal, C.E. (1990). Estimation of Precision in the Generalized Estimation System when Imputation is Used. Report, Informatics and Methodology Field, Statistics Canada.


## CLOSING REMARKS

Procecdings of Statistics Canada Symposium 90
Measurement and Improvement of Data Quality
October 1990

# CLOSING REMARKS 

G.J. Brackstone

That brings us to the end of Symposium 90. We have covered a lot of ground - from John Early's rousing talk on the first morning which set the tone and context for the Symposium, to Carl Sarndal's excellent review of the impact of imputation on variance and variance estimation and the discussion which followed.

We have explored many aspects of data quality, some of them traditional areas for survey statisticians, some of them novel. We have considered total survey design in which we try to balance cost and quality across all operations of a survey, and we have addressed means for improving efficiency and productivity within individual survey operations. We have dealt with the perennial issue of census coverage, and we have explored recent developments in variance estimation. The special quality problems associated with administrative records have received atiention, as have quality issues inherent in maintaining survey frames - which themselves often depend on administrative sources. We were introduced to the challenge of integrating data from different sources and of varying quality, and, of course, we heard about quality assurance programs.

Where does all this leave us? We could go away depressed that there are so many quality problems to address, many of them very difficult, or we could leave energized to apply and explore some of what we have heard in the past three days to our own problems. I hope it will be the latter. To those of you from Statistics Canada, we certainly welcome proposals that this Symposium may have prompted that will provide needed quality improvements in our programs at reasonable cost, or that will enable us to maintain our quality at lower cost.

We have heard a lot about total quality, total quality management and total quality programs. But there are, I think, at least three different senses in which the word "total" is being used here. Firstly, there is the recognition that the concept of quality extends beyond just variance or even mean squared crror. It embraces timeliness, relevance, and even the level of service with which the data are disseminated. In this sense we have to consider total quality. Secondly, there is the idea that the survey designer must consider all the various stages of a survey when trying to reach the optimum cost-quality trade-off. In this sense, quality usually refers to variance or mean squared error - it is "total" in the sense that it embraces the contributions to error from all stages of the survey. And thirdly, "total" can refer to the scope of a quality program in the sense that quality is to be pursued in all the programs and products of an Agency. Each of these ideas is important, but they are different.

It is important to maintain a broad and balanced view of quality. Ideas of "fitness for use" or "what the user wants" are very relevant to attempts to improve quality. At a recent meeting with users, when asked what improvements they would most like to see in our outputs, the immediate response was "get them out sooner." Mean squared error was not on their minds. We need to put our scarce quality improvement resources in those areas where there is a real potential to improve relevant aspects of quality that users will notice and appreciate.

Now a few words about this Symposium itself. Registration was close to 350 . We were especially pleased to welcome participants from Australia, Finland, Macau, Sweden and the United States in addition to representatives from many Canadian universities, businesses and governments. I would like to thank the Organizing Committec of Normand Laniel, Robert Lussier, Mary March and Jeff Smith for the excellent job they have done in planning the agenda for this Symposium and in ensuring that everything ran smoothly over the past few days. In this they have been ably assisted by Rosie Arena-Ryan, Suzanne Bonnell, Ann Brown, Suzanne Johnston, Christine Larabie, Carole Morin and Benita Thérriault of Statistics Canada and by Gillian Murray of Carleton University. I would also like to thank Don Royce who organized the session on Coverage Measurement in Censuses of Population. Finally, I must thank all those presenters and chairpersons whose contributions served to make this Symposium a success.
M. lisert pan ! !

We are starting to plan next year's Symposium. The theme will be Spatial Issues in Statistics. We are intending to focus on subjects that bring together the theories and application of survey methodology and geography. We expect to cover the use of geographic methods, tools and techniques in survey design and in data collection, the processing of geographically-referenced data, the integration and analysis of spatial data, issues in the display and dissemination of geographic data, and related research.

Symposium 90 is closed. Thank you all for your participation and support.

STAISTICS CANADALIRRARY
BもLIOTHQUE S\& ATNIE CANADA

1010207269



[^0]:    ' In cases of joint authorship, the name of the presenter is shown boldface.

[^1]:    ${ }^{1}$ I.P. Fellegi, Chief Statistician, Statistics Canada, Tunney's Pasture, 26-A, R.H. Coats Building, Ottawa, Ontario, Canada, K1A 0T6.

[^2]:    I John F. Early, Vice President, Research and Development, Juran Institute, Inc., 11 River Rd., Wilton, CT 06897, USA.

[^3]:    ${ }^{1}$ S. Linacre, Australian Bureau of Statistics, P.O. Box 10, Belconnen, ACT, 2616, Australia.

[^4]:    1 J.-F. Gosselin, Director, Operations Research and Development Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6.

[^5]:    * The options expressed in this paper are solely the author's.

    1 R.D. Tortora, Chief, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C. 20233. U.S.A.

[^6]:    ${ }^{2}$ The full quality matrix includes columns for the errors of nonobservation and errors of observation for the other 3 stages of the survey.

[^7]:    1 M. Colledge, Business Survey Methods Division, $11^{\text {th }}$ floor, R.H. Coats Building, Statistics Canada, Ottawa, Ontario Canada K1A 0T6.

[^8]:    ${ }^{1}$ J. G. Robinson and H. Hogan, U.S. Bureau of the Census, Washington, D.C. 20233, U.S.A.

[^9]:    ${ }^{2}$ A more extensive discussion of the development and quality of the historical coverage estimates is given in Robinson (1988).
    ${ }^{3}$ For the demographic estimates in 1980, the reliability is conditioned by the treatment of the undocumented alien population. Further, we are investigating a potential bias in the 1940 Birth Registration Test results for Blacks, where an underestimation of birth registration completeness led 10 an overestimate of "corrected" births and overstated estimates of net undercount. This bias mainly affects the estimates for Blacks, aged 35 to 54 in 1980.

[^10]:    4. Adjustment for the apparent bias in the 1940 birth registration completeness factors for Blacks (See Footnote 3) will result in lower Black/White differences for 1940-1980 than those shown in Table 1.
[^11]:    J. Gregory Robinson, (1986)

[^12]:    5 See U.S. Bureau of the Census (1988), Chapters 5-7, for a discussion of the 1980 PEP sets of estimates.

[^13]:    6 Research by Schaefer (1989) has called into question the accuracy of Series 14.

[^14]:    7 This estimate is derived by adding the 2492.9 thousand people duplicated identified in the PEP to 214.0 thousand people duplicated in the Whole Household Usual Home Elsewhere Program. The PEP only measured duplicates with a search area around the enumeration. The estimate thus does not include postcensus day movers who may have been counted once at their census day address and a second time at their new address.

[^15]:    1 R.G. Carter, Social Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, K1A OT6.

[^16]:    1 N. Cressie, Department of Statistics, Iowa State Universily, Ames, IA 50011, USA.

[^17]:    ${ }^{1}$ H.F. Woltman and K.F. Thomas, Statistical Support Division, US Bureau of the Census, Washington, DC 20233.

[^18]:    1 R.R. Sitter, Assistant Professor, Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, K1S 5B6. This research composed a portion of the author's Ph.D. thesis at the University of Waterloo, Waterloo, Ontario, under the supervision of C.F.J. Wu.

[^19]:    ${ }^{1}$ Y. Leblond, Business Survey Methods Division, $11^{\text {th }}$ floor, R.H. Coats Building, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6.

[^20]:    I G.S. Werking and R.L. Clayton, Monthly Industry Employment Statistics Division, Bureau of Labor Statistics, Washington, D.C., USA, 20212.

[^21]:    ${ }^{1}$ J.-M. Berthelot and M. Latouche, Business Survey Methods Division, Statistics Canada, $11^{\text {th }}$ Floor, Section "A", R.H. Coats Building, Ottawa, Ontario, Canada, K1A 0T6. FAX (613) 951-1462.

[^22]:    ${ }^{1}$ R. Fecso, Research and Applications Division, National Agricultural Statistics Scrvice, U.S.D.A., Washington, DC 20250-2000.

[^23]:    1 J.M. Leyes, Small Area and Administrative Data Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6.

[^24]:    2 Unmarried persons who (a) declare themselves to be single, (b) are under the age of 30, (c) reside with their parents and (d) file a tax return are defined to be "filing children".

[^25]:    ${ }^{3}$ To minimize the T1FF data processing costs, most of the T1FF data in this paper are based on samples.

    * There is not an exact reference period relationship between T1FF and the population estimates. The closest is to use T1FF for one year and the population estimates for the next year. For example, the 1988 T1FF data are compared to the 1989 population cstimates.

    5 In processing the 1986 tax file, a somewhat carlier file was used than in other years. As a result, the coverage was lower than in other years. Had this not occurred, the coverage in 1986 would have been higher than $93.7 \%$.

[^26]:    ${ }^{6}$ With more experience, the record linkage procedures were improved and more single parent families were linked to other single parent families to create an increased number of common law couples.

[^27]:    3 The T1 does contain some information on dependent children, namely, relationship to taxfiler and birthdate. The T1 does, however, contain the birthdates of children for Family Allowance children (covers about $85 \%$ of children). A project is in process to assign ages to imputed children.

    * The taxfiling rate for the $65+$ population increased from $55.6 \%$ in 1985 to $70.9 \%$ in 1988.

[^28]:    9 The SCF is an annual supplement to the Canadian Labour Force Survey. The SCF is similar to the March supplement to the Current Population Survey (CPS) in the United States.

[^29]:    ${ }^{10}$ The most recent refundable tax credit for the Federal Sales Tax Credit was $\$ 70$ per person eligible. The current information indicates that the GST refundable tax credit will be paid quarterly and that the payment will be about $\$ 380$ per person, although there are provisions for the amount according to prescribed criteria.

[^30]:    ${ }^{1}$ G. Kalton, Survey Research Center, University of Michigan, Ann Arbor, Michigan 48106-1248, U.S.A.

[^31]:    ${ }^{1}$ C.Z.F. Clark, Agriculture Division, U.S. Bureau of the Census, Washington, D.C. 20233 U.S.A.

[^32]:    1 P.S. Hanczaryk and M.L. Trager, U.S. Department of Commerce, Bureau of the Census, Washington, D.C. 20233 U.S.A.

    2 Businesses with paid employees in the current and/or the prior year.

[^33]:    ${ }^{3}$ We will develop a mechanism to account for such cases in future automated-coding programs. Certain combinations of key words should be coded, despite the presence of an additional key word in the business name.

[^34]:    ${ }^{\prime}$ L. Granquist, Statistical Methods, Research \& Development Department, Statistics Sweden, S-115 81 Stockholm, Sweden.

[^35]:    L. Draper and T. Petkunas, Statistical Research Division; B. Greenberg, Business Division, Bureau of the Census, Department of Commerce, Washington, D.C. 20233, U.S.A.

[^36]:    ${ }^{1}$ L.-P. Rivest and E. Rouillard, Département de mathématiques et de statistique, Université Laval, Cité universitaire, Québec, G1K 7P4, Canada.

[^37]:    ' K. O'Conor, Internal Revenue Service, P.O. Box 2608, Washington D.C. USA 20013 B. K. Atrostic and R. Gillette, Office of Tax Analysis.

[^38]:    1 R.D. Burgess, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6.

[^39]:    1/ the value and appropriateness of release criteria from the perspectives of the user and the Agency;
    2 / the nature or form of release criteria that might or should be applied for statistical programs;

[^40]:    ${ }^{1}$ B.L. Khuong, Director of Mcthodology, Bureau de la statistique du Québec, 117 ruc St-André, $3^{\circ}$ étage, Québec, Québec G1K 3Y3.

[^41]:    ' D.N. Williams and Associates, 418 Edgewood Avenue, Ottawa, Ontario, K1Z 5K5.

[^42]:    ${ }^{1}$ E. Yashchin, IBM Corporation, Thomas J. Watson Research Center, Department of Mathematical Sciences, Box 218, Yorktown Heights, NY 10598.

[^43]:    R.W. Pautke, T.C. Redman, AT\&T Network Services Division, AT\&T Bell Laboratories. Crawfords Corner Road, Holmdel, New Jersey, USA 07733.
    ${ }^{2}$ Specific AT\&T data quality measures are proprietary and are not presented here. All examples are real insofar as they accurately reflect the main points of this paper.

[^44]:    ${ }^{3}$ Accuracy is sometimes confused with precision. According to Shewhart (1939) accuracy is a measure of correctness, while precision "is a measure of the degree of reproducibility."

[^45]:    - The access process meets these criteria.

[^46]:    s Herein, take the process owner to be that person or collection of persons responsible for the process.

[^47]:    ${ }^{6}$ We do not wish to convey the message that normalizational and translational changes are not of concern． Adjusting formats and／or making translations adds time and expense both to software development and to processing and provides opportunities for errors．

[^48]:    ${ }^{7}$ BAN stands for Billing Account Number, LSO SECLOC is the telephone company office providing the service.

[^49]:    1 C.-E. Särndal, Université de Montrêal, C.P. 6128, Succursale A, Montréal, Québec H3C 3 J 7.

