$A_{i}: 191989$

LIERAFY

## EIISLIOTHÈQUE



## Canadä

# SURVEY METHODOLOGY 

## A JOURNAL OF STATISTICS CANADA DECEMBER 1988

Published under the authority of the Minister of Regional Industrial Expansion and the Minister of State for Science and Technology
${ }^{\circ}$ Minister of Supply and Services Canada 1988

Extracts from this publication may be reproduced for individual use without permission proyided the source is fully acknowledged. However, reproduction of this publication in whole or in part for purposes of resale or redistribution requires written permission from the Programs and Publishing Products Group, Acting Permissions Officer, Crown Copyright Administration,

Canadian Government Publishing Centre,
Ottawa, Canada K1A 0S9
March 1989
Price: Canada, $\$ 30.00$ a year
Other Countries, \$35.00 a year
Payment to be made in Canadian funds or equivalent
Catalogue 12-001, Vol. 14, No. 2
ISSN 0714-0045
Ottawa

# SURVEY METHODOLOGY 

## A Journal of Statistics Canada

The Survey Methodology Journal is abstracted in The Survey Statistician and Statistical Theory and Methods Abstracts and is referenced in the Current Index to Statistics.

## MANAGEMENT BOARD

Chairman
G.J. Brackstone

Members
B.N. Chinnappa
R. Platek
G.J.C. Hole
D. Roy
C. Patrick
F. Mayda (Production Manager)
M.P. Singh

## EDITORIAL BOARD

Editor
M.P. Singh, Statistics Canada

Associate Editors

K.G. Basavarajappa, Statistics Canada<br>G. Kalton, University of Michigan<br>D.R. Bellhouse, U. of Western Ontario<br>L. Biggeri, University of Florence<br>M.N. Murthy, Applied Statistics Centre, India<br>W.M. Podehl, Statistics Canada<br>D. Binder, Statistics Canada<br>E.B. Dagum, Statistics Canada<br>W.A. Fuller, Iowa State University<br>J.F. Gentleman, Statistics Canada<br>M. Gonzalez, U.S. Office of<br>Management and Budget<br>D. Holt, University of Southampton<br>J.N.K. Rao, Carleton University<br>D.B. Rubin, Harvard University<br>I. Sande, Statistics Canada<br>C.E. Särndal, University of Montreal<br>F.J. Scheuren, U.S. Internal Revenue Service<br>V. Tremblay, Statplus, Montreal<br>K.M. Wolter, U.S. Bureau of the Census

## Assistant Editors

J. Armstrong, J. Gambino and J.-L. Tambay, Statistics Canada

## EDITORIAL POLICY

The Survey Methodology Journal publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

## Submission of Manuscripts

The Survey Methodology Journal is published twice a year. Authors are invited to submit their manuscripts in either of the two official languages, English or French to the Editor, Dr. M.P. Singh, Social Survey Methods Division, Statistics Canada, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Two nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

## Subscription Rates

The price of the Survey Methodology Journal (Catalogue No. 12-001) is $\$ 30.00$ per year in Canada, $\$ 35.00$ per year for other countries (payment to be made in Canadian funds or equivalent). Subscription order should be sent to: Publication Sales, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6. A reduced price of US $\$ 16.00$ ( $\$ 20.00$ Can.) is available to members of the American Statistical Association, the International Association of Survey Statisticians, and the Statistical Society of Canada. Please subscribe through your organization.

# SURVEY METHODOLOGY 

A Journal of Statistics Canada<br>Volume 14, Number 2, December 1988

## CONTENTS

In This Issue ..... 135
Special Section - Census Coverage Error
R.D. BURGESS
Evaluation of Reverse Record Check Estimates of Undercoverage in the Canadian Census of Population ..... 137
A. ROMANIUC
A Demographic Approach to the Evaluation of the 1986 Census and the Estimates of Canada's Population ..... 157
C.Y. CHOI, D.G. STEEL, and T.J. SKINNER
Adjusting the 1986 Australian Census Count for Under-Enumeration ..... 173
N. CRESSIE
When Are Census Counts Improved by Adjustment? ..... 191
D.B. RUBIN, J.L. SCHAFER, and N. SCHENKER
Imputation Strategies for Missing Values in Post-Enumeration Surveys ..... 209
D.J. FEIN and K.K. WEST
The Sources of Census Undercount: Findings from the 1986 Los Angeles Test Census ..... 223
M.H. MULRY and B.D. SPENCER
Total Error in the Dual System Estimator: The 1986 Census of Central Los Angeles County ..... 241
A.M. ZASLAVSKY
Representing Local Area Adjustments by Reweighting of Households ..... 265

## CONTENTS - Continued

Software Developments
J. LORIGNY
QUID, A General Automatic Coding Method ..... 289
M.J. WENZOWSKI
ACTR: A Generalized Automated Coding System ..... 299
W. MUDRYK
Quality Control Processing System for Survey Operations ..... 309
Y. DeGUIRE
Postal Address Analysis ..... 317
D.N. EMERY
A Brief Note on SQL ..... 327
G. NATHAN
A Bibliography on Randomized Response: 1965-1987 ..... 331
Corrigendum ..... 347
Acknowledgements ..... 349

## In This Issue

Eight papers in this issue deal with Census Coverage Error. These papers, together with the four papers on this topic that appeared in the June 1988 issue, provide the reader a good overview of some of the latest methods available for dealing with census coverage error. A great deal of attention has recently been directed at this problem by both policy makers and statisticians. In many countries, studies are carried out during or following each census to measure coverage error. In Canada, the Reverse Record Check (RRC) is the most important study undertaken to measure undercoverage. A Post-Enumeration Survey (PES) is conducted in the United States and Australia.

The papers by Burgess and Romaniuc deal with coverage problems in the Canadian Census of Population. Burgess describes the RRC methodology, and considers some of its limitations that lead to errors in estimates of undercoverage. Romaniuc, on the other hand, takes a demographic approach to the study of the accuracy of the census. The results obtained in this way are contrasted with those based on the RRC. In addition, Romaniuc looks at the quality of data for components of change (births, deaths, migration) used in the demographic approach.

Choi, Steel and Skinner's paper deals with the 1986 Australian PES. Like Romaniuc, the authors consider demographic estimates of under-enumeration. Based on their analysis, the authors conclude that PES-based adjustments should continue to be used in the 1991 Census, but emphasize that investigation of bias problems should continue.

Cressie uses a model for undercount errors to investigate the adjustment of census counts. He considers synthetic estimation, Bayes and empirical Bayes approaches, and uses risk to compare estimators. A "usual empirical Bayes" estimator is found to have the smallest risk. Cressie notes that the results depend on the assumption that a sufficiently large number of households are chosen in the PES.

The paper by Rubin, Schafer and Schenker on imputation for missing values in a PES also has a Bayesian flavour. The authors review the imputation methods discussed by Schenker in the previous issue of Survey Methodology. They propose two model-based methods, and conclude that the method that does not ignore the missing data mechanism is preferable. The authors caution that, although their approach looks promising, more work is needed.

Fein and West present a systematic classification of the causes of undercount and conclude that partial household omission is the biggest contributor to the undercount. Methodological analysis of total error in the dual system estimator (an estimator that was examined by authors in the June 1988 issue) is discussed by Mulry and Spencer. Using a Bayesian approach, the authors combine the error components to obtain a final interval estimate of net undercount rate.

Zaslavsky deals with the undercount problem by using block-level undercount estimates to reweight households in the block. An advantage of this approach is that the "character" of each block is preserved. The details of the method are interesting and will look familiar to readers acquainted with raking methods.

The development of new computer systems designed to process large amounts of information is a topic of increasing interest to survey statisticians. Five of the papers in this issue describe Software Development related to survey methodology.

Automated coding systems developed by central statistical agencies are described in two papers. Lorigny deals with the QUID system used at the Institut National de la Statistique et des Études Économiques. Wenzowski's paper is a guide to the ACTR system, developed at Statistics Canada. Both QUID and ACTR are designed to handle any type of classification system efficiently.

Readers will be interested in comparing the approaches taken in the two systems. Some performance data are also given.

Mudryk describes a computer system for quality control currently used as part of Statistics Canada's overall quality assurance program. The objectives of the system are both to exercise error prevention in survey processing operations and to reduce inspection levels progressively as the quality of processing improves and stabilizes.

Deguire describes a system, designed to analyze the syntax of postal addresses, currently under development at Statistics Canada. The software produces address search keys consisting of standardized address components that can be used during computerized matching operations such as those involved in the construction of a national Address Register.

Emery describes SQL (Structured Query Language), the most popular query language associated with relational database management systems. The strengths and weaknesses of the language are highlighted.

In the final paper in this issue, Nathan provides a comprehensive list of over 250 books, theses and papers dealing with randomized response. A subject classification is also included.

The Editor

# Evaluation of Reverse Record Check Estimates of Undercoverage in the Canadian Census of Population 

R.D. BURGESS ${ }^{\mathbf{1}}$


#### Abstract

Estimates of undercoverage in the Canadian Census of Population have been produced for each Census since 1961, using a Reverse Record Check method. The reliability of the estimates is important to how they are used to assess the quality of the Census data and to identify significant causes of coverage error. It is also critical to the development of methods and procedures to improve coverage for future Censuses. The purpose of this paper is to identify potential sources of error in the Reverse Record Check, which should be understood and addressed, where possible, in using this method to estimate coverage error.


KEY WORDS: Matching; Mobility; Nonresponse bias; Response error; Reverse record check; Sampling error; Tracing.

## 1. INTRODUCTION

The Census of Canada is conducted every five years; the most recent was in 1986. Starting with the 1971 Census, the main data collection methodology has been self- enumeration: less than $4 \%$ of the population are enumerated using the canvasser method. In geographic areas where self-enumeration is used, each dwelling is listed and a questionnaire dropped off by an enumerator just prior to Census Day (June 3 in 1981 and 1986). In larger urban areas the respondent household is asked to return the completed questionnaire by mail to the local supervisor of the enumeration. In rural areas and smaller urban areas the questionnaires are picked up by the enumerator.

The enumerator is to perform basic checks of coverage and response quality for his/her assignment and follow up on missing and incomplete questionnaires. Supervisory checks and quality control of the enumerator's work are also carried out. However, there is no independent and rigorous check of the listing of dwellings. Further, there is only limited opportunity to verify the number of persons listed on the questionnaire by the respondent household.

Not unexpectedly there are overcoverage and undercoverage errors in the Census. Such errors are important because of the various uses of Census data; representation in the Parliament of Canada is determined using Census population counts: various federal- provincial government financial agreements incorporate formulae that have population count or distribution as a factor (Statistics Canada 1983b). In turn the quality of estimates of coverage error is an important issue: for the use of Census data; in considering adjustment of population and dwelling counts to compensate for the coverage error; and in attempting to improve coverage quality for future Censuses by identifying significant causes or areas of coverage error.

Since 1961, Statistics Canada has produced and published an estimate of undercoverage for each Census of Population. The method used to produce these estimates has been a Reverse Record Check (RRC) study which involves five general activities or stages:

[^0](i) frame preparation - identification of a set of nonoverlapping lists that together are to cover the total population that should be enumerated in the Census;
(ii) sample design and selection - selection of a random sample of persons from the lists;
(iii) tracing - determination of the address of usual place of residence on Census Day for each selected person (or verification that he/she died or emigrated prior to the Census);
(iv) searching - review of Census returns to determine whether the selected person had been enumerated or missed in the Census; and
(v) weighting and estimation - weighting up of sample results to produce an estimate of the number of persons missed in the Census.

A more detailed description of this methodology can be found in Gosselin 1976 or Statistics Canada 1984.

Other methodologies - post Census re-enumeration, demographic analysis and administrative record checks - could also be used to estimate Census undercoverage. In the Canadian context, however, each of these methodologies would likely produce results less reliable than those of the RRC. Re-enumeration studies show a tendency to miss the same households or persons as the Census itself. Demographic methods are model-based and suffer from a lack of reliable emigration estimates, measure only change in net coverage between censuses, do not identify individual cases and causes of coverage error, and are weakened sub-nationally by error in internal migration estimates. Administrative record checks are limited by the absence of a national administrative system that either has more complete coverage than the Census or has coverage errors independent of Census coverage error - a condition that would allow an incomplete administrative file to be used. Even if such a complete system existed, its use would be another version of a reverse record check, unless it were completely up to date in coverage and addresses, as of Census Day.

For these reasons the reverse record check has been the preferred methodology in Canada, though demographic analysis methods have been used for corroborative analysis. However, the RRC itself has deficiencies. The purpose of this paper is to describe some of the sources of error or limitations in the RRC method, in the context of the Canadian Census of Population. In Section 2 aspects of the survey methodology of the RRC that can lead to error in the final results are reviewed. The results of some analysis of RRC estimates, in conjunction with data from other sources, have raised unresolved problems related to the use of RRC results in population estimation. These results are presented in Section 3. Some concluding remarks are given in Section 4.

## 2. LIMITATIONS OF THE REVERSE RECORD CHECK METHODOLOGY

A limitation, in the context of this paper, is anything that restricts the applicability of the Reverse Record Check estimates or the confidence with which they can be used. Limitations can arise because of: differences between what is conceptually required by users and what the RRC attempts to measure; shortfalls in the design of the Reverse Record Check in attempting to meet its objectives; or sampling, response and other errors. Some of these limitations might be eliminated or reduced through modification of specific aspects of the Reverse Record Check. Others will persist or, by their nature, cannot be addressed.

### 2.1 Applicability of Reverse Record Check Estimates

The objective of the Reverse Record Check is to provide estimates, for each of the ten provinces, of undercoverage in the Census of Population. Net coverage error is not estimated and the Yukon and Northwest Territories are excluded from the study.

The RRC estimates the proportion of the population missed in the Census - i.e., the proportion of the population that was not enumerated but should have been. Overcoverage (persons enumerated more than once, and persons enumerated who should not have been or were ficticious) is not estimated by the RRC. Thus net coverage error, undercoverage minus overcoverage, is not estimated by this vehicle. Even if the amount is small, the potential importance of overcoverage lies in its size and distribution relative to undercoverage. For example, overcoverage of $0.2 \%$, one tenth the level of undercoverage in 1976 and 1981, would be very important if the rate for a particular province is as high as $0.5 \%$.

The two Canadian territories have not been included in the RRC because the size of their populations is small but they have exceptionally high rates of intercensal in and out migration. In terms of sampling error, to produce reliable estimates for the territories, a proportionally large sample of the territorial population would have to be selected - of the order of a $5 \%$ sample or 3,750 persons. The territories have in and out intercensal migration rates of a third or more. Therefore, 1,250 of the 3,750 persons (on average) in the minimum sample should be intercensal in-migrants, assuming a proportional sample is required. The RRC uses lists for which the address of residence for the majority of persons was obtained five years earlier and in-migrants to the territories can only be identified during the conduct of the study. This in itself is not a problem. However, the RRC uses only a $0.15 \%$ sample. The in-migrants to the territories, therefore, would be expected to be sampled at this latter rate and not at the required $5 \%$ rate. This would result in a sample of in-migrants to the territories of only 30 persons. Thus, within the current framework of the RRC, and without prohibitive additional expense, it is not possible to select a meaningful sample to represent that third or more of the territorial population who are intercensal in-migrants.

### 2.2 The Reverse Record Check Methodology

Each of the five stages of the Reverse Record Check is a known or potential source of error.

### 2.2.1 Frame

The sample for the RRC is selected from four lists or frames:
(i) Census: persons enumerated in the previous Census - for example, the 1981 Census was used for the 1986 Reverse Record Check;
(ii) Birth: intercensal births, obtained from vital statistics records;
(iii) Immigrant: intercensal immigrants, obtained from records of Employment and Immigration Canada; and
(iv) Missed: persons missed in the previous Census - which is available as a sample only from the previous Reverse Record Check (no complete list exists for this group).

These lists are intended to include or represent, without duplication of individuals on or between lists, all persons who should be enumerated (in one of the ten provinces) in the current Census.

Some people, however, are not represented on these lists. Included among these are: (a) intercensal and never enumerated illegal aliens; (b) certain classifications of refugee; (c) certain

Canadians "abroad" at the time of the previous Census who returned prior to the current Census; (d) persons who move from the territories to one of the provinces in the intercensal period; and (e) persons not enumerated in any Census covered by the application of the RRC, but who were usual residents of Canada prior to 1961.

It is assumed, without direct evidence, that the number of persons in category (e) has become small enough to be irrelevant. For the 1981 Census the size of category (d) was estimated to be of the order of 18,000 persons. Most of these persons were usual residents of the territories at the time of the previous (1976) Census. There were probably also a few of what would be Birth frame and Immigrant frame persons among the 18,000 .

Category (c) includes some Canadians working, studying or travelling abroad who did not maintain a usual place of residence in Canada during their absence and may also include children born outside Canada to parents in this category. It does not include persons in the Canadian military, in External Affairs or other government service (and their families) living abroad. They are included in the Census frame and the Missed frame. For the 1981 Census, the size of this returning 'abroad" group was estimated to be approximately 67,000 persons.

Refugee applicants and illegal aliens in Canada are to be enumerated in the Census, assuming they do not have a usual place of residence outside of Canada, and are not holders of work or student visas. For the 1981 and 1986 RRC studies, persons applying from abroad and entering Canada as refugees were included in the Immigrant frame. Persons applying within Canada were included in the Immigrant frame only if they had been granted refugee status. As of April 1985, there were 12,500 applications from within Canada under consideration PLAUT 1985. The number of illegal aliens in Canada is not known or reliably estimated. Some illegal aliens may be represented in the Census frame or even the Missed frame. Amnesty programmes in the 1970's and 80's will have resulted in some illegal aliens being entered in the Immigrant frame.

Under the current RRC methodology the exclusions to the frames are important to the extent that such persons are not counted in the current Census. Since the Immigrant frame tends to have a high undercoverage rate ( $8.5 \%$ compared to $2.0 \%$ overall in 1981), it is not unreasonable to expect a high undercoverage rate for the refugee status claimants. It is possible that the majority of illegal aliens were not counted in the Census. These elements of undercoverage could be significant relative to the estimated number of persons missed (approximately 500,000 in 1981). The refugee status claimants and the illegal aliens may have been clustered in a few urban centres within only certain provinces. This would increase the impact of such exclusions on the reliability of estimates.

The lists can also be expected to include some amount of overcoverage; e.g., persons enumerated in the previous Census who should not have been or who were enumerated more than once, fictitious persons and processing errors. Some overcoverage is detected during the course of the RRC operations. In estimating undercoverage, however, the effect of overcoverage in the frames would be consequential only if it approaches or exceeds the undercoverage in the Census in size.

### 2.2.2 Sample Size and Design

Error due to sampling is a major limitation of the RRC results. While the potential size of this error is dependent upon sample size and design, the sample size is the more important element. It, along with the available lists, limits the design options.

The basic 1981 and 1976 RRC undercoverage estimates for provinces and their corresponding estimates of standard error are presented in Table 1. The coefficients of variation (standard error divided by estimated undercoverage) varied from $4.5 \%$ at the Canada ( 10 provinces)

Table 1
Estimated Population Undercoverage in the 1981 and 1976 Census, by Province, showing Provinces with Significant Differences in Population Undercoverage (with $95 \%$ confidence)

| Province | Population Undercoverage |  | Province with a Significantly Different Undercoverage Rate |
| :---: | :---: | :---: | :---: |
|  | Rate | S.E. |  |
|  | (\%) | (\%) |  |
| 1981 Census <br> Canada (10 Provinces) |  |  |  |
|  | 2.01 | 0.09 |  |
| 1. Newfoundland | 1.74 | 0.45 | 10 |
| 2. Prince Edward Island | 1.17 | 0.54 | 9 and 10 |
| 3. Nova Scotia | 1.05 | 0.34 | 5,6,9 and 10 |
| 4. New Brunswick | 1.81 | 0.30 | 10 |
| 5. Québec | 1.91 | 0.21 | $3,7,8$ and 10 |
| 6. Ontario | 1.94 | 0.14 | 3, 7, 8 and 10 |
| 7. Manitoba | 0.98 | 0.35 | 5, 6,9 and 10 |
| 8. Saskatchewan | 0.99 | 0.37 | 5, 6,9 and 10 |
| 9. Alberta | 2.54 | 0.36 | 2, 3, 7 and 8 |
| 10. British Columbia | ${ }^{3.16}$ | 0.33 | all but 9 |
| 1976 Census <br> Canada (10 Provinces) |  |  |  |
|  | 2.04 | 0.10 |  |
| 1. Newfoundland | 1.10 | 0.39 | 5 and 10 |
| 2. Prince Edward Island | 0.38 | 0.25 | 4, $5,6,8,9$ and 10 |
| 3. Nova Scotia | 0.86 | 0.34 | 4, 5 and 10 |
| 4. New Brunswick | 2.16 | 0.37 | 2, 3, 7 and 10 |
| 5. Québec | 2.95 | 0.25 | 1, 2, 3, 6, 7, 8 and 9 |
| 6. Ontario | 1.52 | 0.17 | 1, 2, 5, 7 and 10 |
| 7. Manitoba | 1.07 | 0.33 | 4, 5 and 10 |
| 8. Saskatchewan | 1.33 | 0.34 | 2, 5 and 10 |
| 9. Alberta | 1.49 | 0.26 | 2, 5 and 10 |
| 10. British Columbia | 3.13 | 0.31 | all but 5 |

level, up to $13.6 \%$ at the regional (Atlantic, Québec, Ontario, Prairie and British Columbia) level and up to $46 \%$ at the provincial level. Sub-provincial coefficients of variation were typically higher. For an Electoral District of average size ( 86,323 persons in 1981) with an estimated $\mathbf{2 \%}$ undercoverage, the coefficient of variation would be approximately $50 \%$. For smaller geographic areas and small population groups the coefficient of variation could be much higher.

The sampling error, of course, has an effect on attempts to differentiate among provincial, and among other undercoverage rates. In turn this affects attempts to identify specific causes or areas of undercoverage, and undermines the validity of adjusting for coverage error as a means to improve Census counts. Those provinces with a significantly different undercoverage rate are also shown in Table 1. The undercoverage rates for the provinces appear to fall into six groupings, for 1981, based on both rate of undercoverage and provinces with which the rate is significantly different. For 1976, with eight groups, there was less similarity between provinces. No group in either Census, however, can be shown to be completely different from all others, and may not be.

This general situation is not dissimilar to that for applications of the Reverse Record Check for the 1966 and 1971 Censuses. From 1966 onward only the province of British Columbia has had an undercoverage rate significantly above the Canada level. The variation from Census to Census for most provinces, in large part, could be due to sampling error. Why it is not for British Columbia is a major concern for both the Reverse Record Check and the Census.

The need to use a sample of "missed" persons from the previous RRC also places a limitation on the design and sample size. There is no direct control of the size of this segment of the sample. Any limitations of the previous Reverse Record Check, to the degree that these were reflected in the estimate of "missed" persons, will be passed. (See Sections 2.2.4, 2.2.5 and 3).

### 2.2.3 Tracing

Given the nature of the lists or frames used for sample selection, addresses and other information may be up to five years out of date. Attempts are made to update addresses prior to Census Day using administrative files. (This was first carried out extensively for the 1986 RRC.) After Census Day, the Census questionnaire corresponding to the original address, or the update if available, is searched as a first attempt to determine whether the selected person was enumerated in the Census. Every selected person not found enumerated in the first search must be traced. The selected person, or a reliable source, must be contacted either to obtain an updated or confirmed address, or to determine the selected person's status, i.e., as deceased, emigrated, abroad.

Despite extensive tracing activities, not all selected persons can be traced. This may result in a form of nonresponse bias. In the 1981 RRC $3.4 \%$ of all selected persons were not traced. With overall undercoverage in the Census estimated to be $2.0 \%$ this "not traced" rate represents an important uncertainty in the RRC estimates.

A weight adjustment is carried out to account for these "not traced" cases. The effect of the weight adjustment for the 1981 Census was to impute an undercoverage rate of $3.27 \%$ for the "not traced" cases from the Census and Missed frames (jointly), $1.46 \%$ for the Birth frame and $11.94 \%$ for the Immigrant frame. Overall, the proportion of "not traced" weights "imputed" by the weight adjustment to "missed" was 1.6 times the initial (weighted) proportion represented by the "missed" cases among all traced selected persons. This suggests a relationship between "not traced" and "missed". It is not known, of course, if the 1.6 rate was too high, too low or correct. To the extent that it is not correct, there may be some distortion in provincial estimates of undercoverage as well as a bias in overall estimates of undercoverage.

Since the rates of intercensal interprovincial in and out-migration vary from one province to another, there may be some distortion among provincial estimates. This will occur if the proportion of interprovincial movers within weighting groups is not the same among the cases traced and not traced.
Intercensal-interprovincial-movers (applicable for Census and Missed frames only) have a high undercoverage rate. This rate was estimated to be $6.13 \%$ for the 1981 Census, based upon mobility data from the 1981 RRC derived by comparing of the 1976 Census and 1981 Census addresses. The estimated undercoverage rate for intercensal migrants within a province (i.e., between Census Subdivision (CSD) or municipality movers) was $3.83 \%$. For intercensal nonmigrant movers (within CSD or municipality) the undercoverage rate was estimated to be $\mathbf{2 . 8 3 \%}$. Given these rates and the distribution of mobility characteristics, the "imputed" undercoverage rate for the "not traced" cases from the Census and Missed frames put together would be expected to be at least $3.52 \%$ rather than the actual $3.27 \%$. That is, given persons not traced almost always have moved. It is, in turn, assumed that these "not traced" cases included
proportionally at least as many migrants, within and between provinces, and had not less than the same undercoverage rates, by mobility status, as traced cases. (The distribution of mobility status of the enumerated population 5 years and older, estimated through the 1981 RRC was approximately: (i) Non-movers - $55 \%$; (ii) Non-migrant Movers - $17 \%$; (iii) Migrants Same Province - 21.7\%; (iv) Migrants Different Province - 5\%; and (v) Migrants From Outside Canada-2\%.)

Given the tracing methods used, it is not unreasonable to speculate that the proportion of migrants, and thus the undercoverage rate, was much higher for the "not traced" cases. If they were, then there could be a significant downward bias in the estimates of undercoverage. For example, if the "true" undercoverage rate among the cases not traced was close to $5.0 \%$, then the bias in the undercoverage estimate at the Canada ( 10 provinces) level would exceed the sampling error.

### 2.2.4 Searching and Classification

After all tracing attempts have been made and any interviews conducted, each selected person is classified to one of six categories:
(1) enumerated;
(2) missed;
(3) deceased;
(4) emigrated or abroad;
(5) overcoverage in a list or frame; and
(6) not traced.

As outlined above, to determine whether a selected person has been enumerated or missed the Census questionnaire corresponding to the selected person's address must be searched. For the search to result in the correct classification of the selected person, it is necessary that the address being searched be the correct address, and that the selected person be correctly identified on the Census questionnaire and in RRC documentation; i.e., that there be no response error or nonresponse for the relevant items.

If the selected person is correctly identified (complete name, correct age and sex, etc.,) and there are no processing errors, then no selected person who was missed in the Census will be classified as "enumerated". The converse is not true. If a selected person has been enumerated in the Census at some address other than that which is obtained from the list of selection, some other administrative source or a directory, then to be classified as "enumerated" that address must be provided by the selected person or some other contact. If the selected person does not or can not provide that address (for example, recall error or can not remember), then he or she will be classified as "missed" or "not traced". Generally, when the selected person (or a parent, spouse or other reliable source) gives an address, or set of addresses, where he/she should have been or may have been enumerated, this address information is accepted as correct. Selected persons will be classified as "enumerated" or "missed" based on this address information. It is not known how accurate such address information actually is for persons classified as "missed".

On the other hand there may be a higher probability of classifying a person missed as "not traced'' than a person enumerated in the Census. Before a person can be classified as missed he/she (or a reliable source) must be interviewed to confirm the address and to obtain possible alternative addresses and certain Census data for him/her and the household. This procedure will eliminate some classification error. At the same time, if the information about a person missed is doubted this can only be resolved through the contact with him/her (or a parent,
spouse, etc.). If the doubt is not resolved the case will be classified as "not traced". Conclusive information is not always necessary for a person who was enumerated. With exhaustive searching it may be possible to transform a selected person, who was enumerated, from 'not traced" to "enumerated", even if the address obtained is incomplete or incorrect. Such searching is much less likely to alter the outcome for persons missed in the Census.

The selected person is not always adequately identified. In accepting a selected person as matched; i.e., found enumerated on a Census questionnaire - name is not always identical on the Census and RRC documents. Sometimes only the first person listed on a Census questionnaire has a complete name and in a few cases no names are given. If the identity of the selected person cannot be determined from the list or frame, then the case will be classified as "not traced" at the outset. Included among these will be persons "assigned" for absent households and refusals in the previous Census. Date of birth and other data are not always present, complete or found identical in matching. For the majority of cases the quality of matching is unquestioned, but a minority of cases raise doubts. Doubtful cases accepted as matched potentially are misclassified as "enumerated". Those rejected as matched potentially are misclassified as $\checkmark$ "'missed", though most will be classified as "not traced". Different rules for acceptance/rejection as matched, of course, may yield different estimates of undercoverage.

Some overcoverage in the frames can be detected. This will include: some foreign residents enumerated in the previous Census; persons "created" by processing error in the previous Census; immigrants who have not yet resided in Canada; births in Canada to non-resident parents; and fictitious or out of scope "persons'" listed on the questionnaire from the previous Census. In 1981 these cases represented less than $0.1 \%$ of selected persons.

Overcoverage in the form of duplication in a frame will not be detected. Fictitious selected persons may go undetected and be classified among the "not traced" cases.

The final classifications of the selected persons from the 1981 RRC are presented in Table 2 (from Burgess 1986).

### 2.2.5 Weighting and Estimation

At the time of sample selection, a basic weight equal to the inverse of the sampling fraction is assigned to each selected person record. Two types of weight adjustment are made to this basic weight - one to account for "not traced" cases, the other to account for deviations in

Table 2
1981 RRC Final Classification of Selected Persons

| Final Classification | Frame |  |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Census |  | Birth |  | Immigrant |  | Missed |  | Total |  |
|  | Cases | \% | Cases | \% | Cases | \% | Cases | \% | Cases | \% |
| Traced | 29,761 | 97.1 | 3,211 | 92.3 | 1,392 | 96.1 | 807 | 96.1 | 35,171 | 96.6 |
| Enumerated | 27,541 | 89.8 | 3,096 | 89.0 | 1,113 | 76.8 | 696 | 82.9 | 32,446 | 89.1 |
| Deceased | 1,056 | 3.5 | 33 | 0.9 | 5 | 0.3 | 26 | 3.1 | 1,120 | 3.1 |
| Emigrated/Abroad | 299 | 1.0 | 34 | 1.0 | 111 | 7.7 | 24 | 2.8 | 468 | 1.3 |
| Missed | 865 | 2.8 | 48 | 1.4 | 163 | 11.3 | 61 | 7.3 | 1,137 | 3.1 |
| Not Traced (incl. Overcoverage | 895 | 2.9 | 267 | 7.7 | 57 | 3.9 | 33 | 3.9 | 1,252 | 3.4 |
| TOTAL | 30,656 | 100.0 | 3,478 | 100.0 | 1,449 | 100.0 | 840 | 100.0 | 36,423 | 100.0 |

the representativeness of the sample, after elimination of "not traced" cases, relative to the lists of selection.

A 'not traced" case represents a person enumerated or missed in the Census, a deceased person, an emigrant, a person abroad or overcoverage. The weights of the "not traced" cases, therefore, are redistributed among the "traced" cases. The adjustment is carried out within groups defined by various demographic and geographic characteristics, and frame.

The weight adjustment for the "not traced" cases is carried out in two stages. First, an adjustment is made for those cases for which no tracing was undertaken because there was inadequate information for matching and tracing. These cases are weighted into all other selected persons. Second, an adjustment is made for all other "not traced" cases. These are weighted into specific groups of the remaining selected persons. How the "not traced" adjustment is carried out is restricted by the information available on the "not traced" selected persons. Ideally, how a selected person was traced and whether he/she had moved and how far, as well as demographic characteristics, should be taken into consideration in defining weighting groups. To date only demographic characteristics and minimal mobility data have been used in the weight adjustment. (Persons selected in the Census frame who have not moved in the intercensal period and who were classified as "enumerated" are excluded from this weight adjustment.) By their nature it is difficult to categorize most "not traced" cases beyond the fact that they were not found enumerated at the address given on the list of selection.

For the second type of adjustment, totals for relevant sub-groups of the population are obtained from each frame (except for the Missed frame for which only a sample is available). Using these "known totals", an adjustment to the RRC weights is made within the corresponding subgroups of the sample. This is done to reduce the error in the estimates by ensuring that totals from the sample, for basic population characteristics for which undercoverage rates are published, correspond to the totals in the frames.

Neither adjustment deals at all with the various exclusions to the lists used for sample selection. In the calculation of any proportion of persons missed in the Census the published Census count of enumerated persons is used in the denominator in order to minimize sampling error. (The covariance of the estimate of "enumerated" persons and the estimate of "missed" persons tends to be negative.) Since the RRC does not represent all elements of the true population, the effect of using the Census count is to assume that the undercoverage rate for the exclusions is zero.

The estimator, which takes the general form defined as:
Estimated proportion of persons missed

$$
=\frac{\text { Estimated no. of missed persons }}{\text { no. of persons counted in the Census }+ \text { Estimated no. of missed persons }}
$$

is discussed further in Appendix 2.

### 2.3 Reducing Potential for Error and Methodological Limitations

Experimental work and evaluation of methods in the RRC may make it possible to eliminate or reduce the impact of some sources of error or limitations.

Overcoverage might be estimated by means of an independent study. Such a study is being conducted, on an experimental basis, for the 1986 Census. However, the cost to produce estimates of adequate quality at the province level may be very high.

The production of estimates for the Yukon and Northwest Territories requires a set of lists other than those used for the RRC. Such a set would have to be current and have no significant duplication that could not be removed or estimated. With such a set of lists, the basic

RRC methods could be applied. Some experimental work in this regard has been done and more is planned.

The lists used for the RRC could be augmented to eliminate some of the exclusions, for example, refugee status claimants and migrants from the territories to the provinces. These people, however, will be difficult to trace. Sampling these groups may do little more than change the nature of the problem.

A sample of "abroad" persons could be obtained by using the previous Reverse Record Check. Such a sample, however, would be very small, would not represent the entire group in question and the selected persons would be difficult to trace.

Other than illegal aliens the "never enumerated" group will become smaller and smaller over time. Intercensal illegal aliens, and other illegal aliens never enumerated in Canada, will remain excluded.

The impact of sampling error can be reduced by increasing the sample size. The question is to what size, at what cost, based upon what criteria? An increase in the RRC sample from its current 36,500 persons to 100,000 should be sufficient-to bring the provincial standard error estimates, for the undercoverage rates, down below $0.2 \%$. However, this may not be sufficient for purposes of adjusting the Census counts, depending upon the level and distribution of undercoverage estimates actually obtained. A reduction of the standard error to $0.1 \%$ for each province - the level yielded by the 1981 and 1976 RRC studies for the Canada ( 10 province) level estimate of undercoverage of $2 \%$ - would require a sample for Canada of approximately 350,000 persons, assuming the 1981 provincial levels of undercoverage, type of sample design and design effects. To conduct a high quality RRC operation for such a large sample, given the controls and quality checks required, would be much more costly than the mere increase in sample size suggests, and might be operationally unrealizable. Increasing the sample size, of course, would not reduce any bias in the estimates.

Tracing methods are examined before and after each RRC. Major changes were made for 1986 and changes and improvements are being contemplated for 1991. It must be expected, however, that there will again be a non-negligible percentage of "not traced" cases. These cases will continue to be dealt with by weighting or by imputation and weighting.

Evaluative studies can be conducted to assess the quality of matching and of address information provided by respondents or reliable sources. The potential impact of the matching algorithm or criteria can also be assessed to some extent. However, even if such studies identify a problem, solutions may not be readily forthcoming.

Modifications to the weighting procedures can be tested in an attempt to better deal with mobility and other characteristics when adjusting for "not traced"' cases (Burgess 1986). Additional information for this purpose might be available from administrative sources. Some minor refinements using existing information can also be made. For example, the adjustment for "not traced" persons contacted, but from whom the necessary Census Day address information could not be obtained, might be different from that for "not traced" persons who potentially may be "deceased", "emigrated" or "abroad".

Adjustments using current Census totals of enumerated persons could be tested as well. For this to reduce any bias associated with "not traced" cases and persons not represented in the RRC sample, however, the basic classification of cases to "missed" must be without bias and there must be no interprovincial distortion of the proportion "missed". These types of modifications to the weighting would not in themseives eliminate bias.

## 3. ANALYSIS OF REVERSE RECORD CHECK RESULTS

The RRC not only provides estimates of the number of persons missed in the Census, but also independent estimates of the number of persons enumerated in the Census, and the number of intercensal deaths, emigration and persons who have moxed abroad but who have not emigrated. These estimates are used in validating RRC estimates. Some of the results of this validation process serve to illustrate limitations discussed in Section 2.

Analysis has also been carried out to correlate geographic variation in undercoverage to variation in the distribution of Census population and household characteristics.

### 3.1 Independent Estimates

The Reverse Recored Check estimates of persons enumerated in the Census, of intercensal deaths, and of persons leaving Canada in the intercensal period can be compared to estimates from other appropriately chosen sources - for example, estimates of enumerated persons to Census counts and estimates of deaths to Vital Statistics data. If there are no significant biases in the RRC estimates, then any differences between these estimates will usually be explainable by the corresponding sampling error of the RRC estimate. If there are significant differences, then these might be due to biases in the RRC estimates. The overall quality of these estimates, revealed by the comparisons, likely will be a reflection of the quality of the estimates of "missed" persons.

RRC estimates of emigrants $(296,727)$ and of persons "abroad" $(57,909)$ compared favourably with estimates based upon demographic analysis. The RRC estimate for emigrants, for example, is in the mid range of the five demographic analysis values examined - ranging from 197,000 to 372,000 , with a mean value of 266,400 . The RRC estimate of deceased persons $(846,378)$ is very close to the value $(840,689)$ published by Statistics Canada 1976 to 1981.

Comparisons of estimates for enumerated persons do indicate some problems. Some of these comparisons are presented in Table 3. For Canada ( 10 provinces) and for two of the ten provinces, the number of persons enumerated in the Census, as estimated by the RRC, is significantly different from the published Census count. The discrepancy of 209,911 at the aggregate level can be explained in part by exclusions from the lists or frames of the RRC. The discrepancies among provinces is difficult to explain. That in particular makes the discrepancy important. The 209,911 aggregate discrepancy must be considered in the context of the RRC estimate of 497,277 persons missed in the Census; similarly, the discrepancy for British Columbia of 80,304 in the context of an estimated 89,445 persons missed and the discrepancy for Alberta of 86,244 persons in the context of an estimated 58,335 persons missed.

An estimated 67,000 non-immigrants who had been "abroad" at the time of the previous Census arrived in Canada legally, and an estimated 18,000 persons moved from the territories to a province in the intercensal period. Assuming none of these people was missed in the Census, the discrepancy would be reduced to approximately 125,000 persons. This difference would remain at the outer limits of what would be reasonably accepted as due to sampling error only. Further, all of these $85,000(67,000+18,000)$ persons would have had to have moved to Alberta and British Columbia to reduce the discrepancies for these provinces to within $95 \%$ confidence intervals - a clearly unreasonable supposition.

The remainder of the difference $(125,000)$ could be made up of various (potential) errors in the RRC or the Census: (i) sampling error in the RRC estimate of enumerated persons; (ii) an increase in overcoverage in the 1981 Census - compared with the 1976 Census; (iii) RRC exclusion of illegal aliens and refugee claimants enumerated in the Census itself; (iv) underestimation of persons missed in the 1976 Census - these persons make up 1981 Missed

Table 3
Reverse Record Check Estimates of the Number of Persons Enumerated in the 1981 Census by Province

|  | RRC Estimate <br> of Persons <br> Enumerated | S.E. of <br> RRC <br> Estimate | Census <br> Published <br> Count | Persons <br> Enumerated. <br> RRC-Census | RRC Estimate <br> of Persons <br> Missed |
| :--- | ---: | ---: | ---: | ---: | ---: |
| Canada (10 provinces) | $24,064,376$ | 62,193 | $24,274,287$ | $-209,911^{2}$ | 497,277 |
| Newfoundland | 568,696 | 8,256 | 567,681 | 1,015 | 10,039 |
| Prince Edward Island | 116,012 | 3,005 | 122,506 | $-6,494$ | 1,456 |
| Nova Scotia | 837,045 | 11,185 | 847,442 | $-10,397$ | 9,034 |
| New Brunswick | 685,332 | 8,167 | 696,403 | $-11,071$ | 12,864 |
| Québec | $6,410,662$ | 38,648 | $6,438,403$ | $-27,736$ | 125,180 |
| Ontario | $8,629,374$ | 52,802 | $8,625,107$ | 4,267 | 171,010 |
| Manitoba | $1,028,162$ | 15,133 | $1,026,241$ | 1,921 | 10,203 |
| Saskatchewan | 973,450 | 11,740 | 968,313 | 5,137 | 9,712 |
| Alberta | $2,151,480$ | 24,238 | $2,237,724$ | $-86,244^{2}$ | 58,335 |
| British Columbia | $2,664,163$ | 19,798 | $2,744,467$ | $-80,304^{2}$ | 89,445 |

I Statistics Canada 1982.
2 Greater than 3 standard errors.
frame; and/or (v) over-estimation of persons missed in the 1981 Census. The extent to which each of these sources might have contributed to the difference is not known. The fact that a large part of the difference seems to be associated with British Columbia and Alberta is perhaps in some degree due to under-estimation of intercensal migrants. Migration to these provinces was particularly high between 1976 and 1981 (Statistics Canada 1979; 1983a).

There may also be some bias in the estimates of emigrated, abroad and/or deceased persons. If these are over-estimated for reason other than "not traced" bias, there should also be a tendency to under-estimate the persons missed, since the last address in Canada is sought and used in searching. Persons who emigrated, died or went abroad after Census Day may have been reported as such at the time of tracing, perhaps several months after Census Day. At the same time, the fact that deceased persons do not appear to have been under-estimated despite the exclusions to the RRC frames suggests a lower mortality rate for the exclusions (as is the case for immigrants - see Table 2) than for the entire population and/or over-estimation of this group.

The data in Table 4 show that intercensal migrants were under-estimated for all provinces except Saskatchewan. This may be in part associated with the "not traced" cases. The underestimation for British Columbia may explain the discrepancy for this province shown in Table 3. On the other hand, the under-estimation for Alberta does not adequately explain the discrepancy for that province and, thus one or more of the factors (i) to (v) noted above must be contributing to this discrepancy.

Under-estimation of migrants may cause a distortion of undercoverage estimates among the provinces; i.e., the large differences shown in Table 4 by province might be indicative of substantial biases in provincial under-enumeration rates. Further, as noted in Section 2.2.3, migrants have higher than average levels of undercoverage. If the enumerated persons within this group are under-estimated, while in general non-migrants are not under-estimated, relative to the Census, then estimates of undercoverage may be too low.

Table 4
Reverse Record Check Estimates of Migrants ${ }^{1}$ Enumerated in the 1981 Census, by Province

| Province | Estimate of Migrants |  |  | Census Estimate of InterProvincial Migration |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | RRC | Census published estimate ${ }^{2}$ | Difference RRC-Census | In | Out | Out/In |
| Canada | 4,670,311 | 5,046,500 | -376,239 | 1,124,970 | 1,122,370 | - |
| Newfoundland | 61,499 | 72,100 | -10,601 | 18,430 | 38,265 | 2.08 |
| Prince Edward Island | 13,257 | 20,530 | - 7,273 | 9,945 | 9,950 | 1.00 |
| Nova Scotia | 125,949 | 137,865 | -11,916 | 54,455 | 62,880 | 1.16 |
| New Brunswick | 96,607 | 109,955 | -13,348 | 41,460 | 49,965 | 1.21 |
| Québec | 1,092,919 | 1,145,085 | -52,166 | 61,310 | 203,035 | 3.31 |
| Ontario | 1,572,504 | 1,725,225 | -152,721 | 250,570 | 328,640 | 1.31 |
| Manitoba | 143,391 | 165,105 | -21,714 | 54,030 | 97,620 | 1.81 |
| Saskatchewan | 204,937 | 192,840 | 12,097 | 63,395 | 69,220 | 1.09 |
| Alberta | 669,995 | 691,970 | -21,975 | 336,830 | 139,180 | 0.41 |
| British Columbia | 689,253 | 785,825 | -96,622 | 234,545 | 123,615 | 0.53 |

[^1]Discrepancies between the RRC estimate of enumerated persons and the Census count have also occurred for earlier Census. The value of the RRC estimate minus the Census count was 289,000 for 1971 , and $-324,000$ for 1976 . For both of these Censuses, the RRC estimates of persons deceased and emigrated/abroad were consistent with other sources. The large change from 1971 to 1976, coincident with the large negative values for two consecutive Censuses, cannot emanate from a single source. Changes in the size of overcoverage, larger than the size of the discrepancies, would be required between Censuses. This by itself, however, would not be consistent with the results of demographic analysis for these three Censuses (Statistics Canada 1987).

Remaining consistent with the demographic estimates, the differences would be explained in part by the presence of a large downward bias in the 1971 RRC estimate of persons missed. The 1971 unbiased estimate would have to be of the order of $3.8 \%$ rather than the estimated $1.9 \%$. This would have to be accompanied by a not as large decrease in overcoverage between 1966 and 1971 followed by an increase in overcoverage for 1976 and a decrease for 1981. There would have to be also some under-estimation of missed persons for 1976.

Such a scenario is speculative, however, and no reason was found for such changes occurring. Other scenarios may also be possible. The occurrence of the discrepancies, however, does raise questions about the reliability of the RRC estimates and the potential effect of overcoverage on net coverage error.

The provincial distribution of the discrepancy between the RRC estimate of enumerated persons and the Census count differ among Censuses, further confounding its effects and potential sources. These results for the 1976 Census are given in Table 5.

Table 5
Difference Between Reverse Record Check Estimates of Persons Enumerated and the 1976 Census Counts

|  | Difference in <br> Population <br> Enumerated <br> Province <br> (RRC-1976 Census) | Percent <br> Difference |
| :--- | :---: | :---: |
| Canada (10 provinces) | $-323,500$ | -1.4 |
| Newfoundland | 21,900 | 3.9 |
| Prince Edward Island | -500 | -0.4 |
| Nova Scotia | $-4,500$ | -0.5 |
| New Brunswick | $-15,000$ | -2.3 |
| Québec | $-56,200$ | -0.9 |
| Ontario | $-207,000$ | -2.5 |
| Manitoba | $-6,600$ | -0.6 |
| Saskatchewan | 1,400 | 0.1 |
| Alberta | $-43,400$ | -2.4 |
| British Columbia | $-12,800$ | -0.5 |

### 3.2 Variation in Geographic Distributions

The RRC estimates of undercoverage can be used as general indicators of the coverage quality of the Census. They are also intended to be used to direct the development and testing of coverage improvement procedures for future Censuses. Under ideal circumstances, they would be used to model undercoverage to produce estimates for small areas and as part of a coverage adjustment "correction'' procedure. For these uses, geographic variation in coverage quality, indicated by the RRC results, is of particular concern. Variation in Census data distributions have been examined to determine whether they are correlated to the apparent variation in undercoverage among provinces. To date these investigations have not yielded satisfactory models or explanations.

A lack of success modelling undercoverage or explaining the variation between provinces may be due to, or confounded by: (i) bias and/or sampling error in the RRC estimates; (ii) undercoverage not strongly correlated to the Census characteristics of individuals, households and/or families; (iii) undercoverage correlated to a perhaps complex combination of Census and other characteristics; and/or (iv) a multitude of sources of undercoverage that must be considered separately; for example, undercoverage of individuals considered separately from undercoverage of entire households.

## 4. CONCLUSION

The RRC is thought to be the best vehicle developed to date for estimation of undercoverage in the Census in Canada. Its estimates provide basic measures to monitor and assess the quality of Census counts.

There are conceptual, theoretical and practical limitations to the RRC Check method as currently applied to the Canadian Census. The frames or lists used, while covering the large
majority of the population to be enumerated, are not comprehensive. Specific geographic areas are excluded as are certain segments of the population. The sample size is limited, but not necessarily to its present size, by constraints of tracing and matching, and by the demands for accuracy in operations. The "not traced" cases are a source of bias. The proportion of cases not traced, relative to the proportion of 'missed" cases, in particular, adds an important uncertainty to the estimates, as does the inconsistency of RRC estimates of enumerated persons with corresponding Census counts.

In some instances the degree or impact of error, or limitations, could be evaluated in greater depth. Modifications and alternative procedures or methods that have a reasonable likelihood of improving the quality and applicability of the estimates can be applied. Potentially, alternatives can be developed. Such changes, however, would have varying costs and degrees of effectiveness associated with them. Also, it remains to be shown whether such changes would do more than enhance the status of the RRC estimates as general indicators of coverage quality in the Census.

## ACKNOWLEDGEMENTS

The author would like to thank Gordon Brackstone, Geoff Hole, Judy Clarke and staff of Social Survey Methods for assistance and comments during preparation of the paper. The comments of the referees and editors were helpful and appreciated.

## Appendix 1

## Further Results From the Reverse Record Check

Results from the 1986 Census Reverse Record Check have been published'(Statistics Canada 1988). The following extract displays the undercoverage rates for the 1981 and 1986 Censuses for demographic characteristics. Analysis of the 1986 undercoverage estimates by province, age, sex, marital status, mother tongue and other groupings is continuing.

## 1981 and 1986 Reverse Record Check Undercoverage Rates for Selected Population Characteristics - 10 Provinces

| Characteristic | 1981 Estimated Population Undercoverage |  | 1986 Estimated Population Undercoverage |  |
| :---: | :---: | :---: | :---: | :---: |
|  | Rate | S.E. | Rate | S.E. |
|  | \% | \% | \% | \% |
| Sex |  |  |  |  |
| Male | 2.37 | 0.13 | 3.91 | 0.16 |
|  | 1.65 | 0.12 | 2.87 | 0.16 |


| Age Group |  |  |  |  |
| :--- | :--- | :--- | :--- | :--- |
| $0-4$ | 1.21 | 0.22 | 2.28 | 0.48 |
| $5-14$ | 1.23 | 0.21 | 2.12 | 0.26 |
| $15-19$ | 2.96 | 0.52 | 3.89 | 0.60 |
| $20-24$ | 5.51 | 0.29 | 9.06 | 0.45 |
| $25-34$ | 2.31 | 0.28 | 4.76 | 0.32 |
| $35-44$ | 2.20 | 0.26 | 2.40 | 0.32 |
| $45-54$ | 0.81 | 0.23 | 1.77 | 0.28 |
| $55-64$ | 0.91 | 0.29 | 2.09 | 0.31 |


| Marital Status |  |  |  |  |
| :--- | :--- | :--- | :--- | :--- |
| Married/Separated | 1.22 | 0.11 | 1.89 | 0.15 |
| Divorced | 5.10 | 1.03 | 7.07 | 1.07 |
| Widowed | 0.64 | 0.39 | 2.68 | 0.51 |
| Single/Never Married | 2.86 | 0.16 | 4.91 | 0.21 |


| Mother Tongue |  |  |  |  |
| :--- | ---: | ---: | ---: | ---: |
| English | 1.86 | 0.11 | 3.12 | 0.13 |
| French | 1.80 | 0.20 | 3.10 | 0.33 |
| Other | 3.08 | 0.26 | - | - |


| Urban/Rural Population <br> Size Group |  |  |  |  |
| :--- | :--- | :--- | :--- | :--- |
| Urban Areas | 2.08 | 0.11 | 3.28 | 0.13 |
| $500,000 \&$ over | 2.29 | 0.17 | 38 | 0.15 |
| 100,000 to 499,999 | 1.86 | 0.31 | 2.94 | 0.33 |
| Less than 100,000 | 1.80 | 0.23 | - | - |

## Appendix 2

## Equations Used to Assess RRC Estimates and Estimator

The 1981 Reverse Record Check estimates have been assessed and discussed based upon four equations. The first simply defines the RRC population or frames. The second redefines the RRC sample in terms of the outcome or estimates of the study. The third defines the population enumerated in the Census in terms of the RRC estimate of enumerated persons. The fourth defines the error components for the estimate of missed persons.

## Equation 1:

The RRC population size $=C_{76}+M_{76}-e\left(\hat{M}_{76}\right)+I_{76 / 81}+B_{76 / 81}$, where

```
\(C_{76} \quad=\) number of persons counted, or enumerated, in one of the ten provinces in the
    1976 Census,
    \(M_{76} \quad=\) number of persons missed in one of the ten provinces in the 1976 Census,
    \(e\left(\hat{M}_{76}\right)=\) error (under or ( - ) over estimation of persons) associated with \(M_{76}\), the Missed
        frame sample; i.e., \(M_{76}=\hat{M}_{76}+e\left(\hat{M}_{76}\right)\),
    \(I_{76 / 81}=\) number of registered 1976 to 1981 intercensal immigrants to one of the ten
        provinces,
    \(B_{76 / 81}=\) number of registered 1976 to 1981 intercensal births in one of the ten provinces.
```


## Equation 2:

The RRC estimates $=\hat{C}_{81}+\hat{C}_{f 781}+\hat{M}_{81}+\hat{M}_{f 781}+\hat{L}_{76 / 81}+\hat{A}_{81}+\hat{D}_{76 / 81}+\hat{O}_{f 81}$ where
$\hat{C}_{81} \quad=$ estimated number of persons in an RRC frame who were enumerated in one of the ten provinces in the 1981 Census,
$\hat{C}_{f T 81}=$ estimated number of persons in an RRC frame who were enumerated in one of two territories in the 1981 Census,
$\hat{M}_{81} \quad=$ estimated number of persons in an RRC frame who were missed in one of the ten provinces in the 1981 Census,
$\hat{M}_{f T 81}=$ estimated number of persons in an RRC frame who were missed in one of the two territories in the 1981 Census,
$\hat{L}_{76 / 81}=$ estimated number of persons in an RRC frame who were 1976 to 1981 intercensal emigrants,
$\hat{A}_{81} \quad=$ estimated number of persons in an RRC frame who were abroad and had no usual place of residence in Canada at the time of the 1981 Census,
$\hat{D}_{76 / 81}=$ estimated number of persons in an RRC frame who died in the 1976 to 1981 intercensal period,
$\delta_{f 81}=$ estimated overcoverage (number of 'persons') in the Census, Birth and Immigrant frames which was detectable in the 1981 RRC operations.

## Equation 3:

The estimate $\hat{C}_{81}$ should $=C_{81}-\hat{C}_{81}\left[e\left(\hat{M}_{76}\right)\right]-\hat{C}_{c / r e 81}-R_{81}-T_{76 / 81}-S_{76 / 81}$ $+M_{n 81}-O_{81}+\hat{C}\left(O_{n f 81}\right)$,
where


Thus,
$\hat{C}_{81}-C_{81}=-\hat{C}_{81}\left[e\left(\hat{M}_{76}\right)\right]-\hat{C}_{c / r e 81}-S_{76 / 81}+M_{n 81}-O_{81}+\hat{C}\left(O_{n f 81}\right)$

- $R_{76 / 81}-T_{76 / 81}$,
assuming no error in $\hat{O}_{n f 81}$.


## Equation 4:

$M_{81}-\hat{M}_{81}=\hat{M}_{81}\left[e\left(\hat{M}_{76}\right)\right]-\hat{M}_{81}\left(\hat{O}_{n f 81}\right)+\hat{M}_{c / r e 81}+M_{n 81}=e\left(\hat{M}_{81}\right)$.
where
$\hat{M}_{c / r e 81} \quad=$ under or (-) over-estimation of "missed" persons in an RRC frame because of classification, response, sampling and "no trace" error in the 1981 RRC,
$\hat{M}_{81}\left[e\left(\hat{M}_{76}\right)\right]=$ that component of $e\left(\hat{M}_{76}\right)$, represented in $\hat{M}_{81}$,
$\hat{M}_{81}\left(\hat{C}_{n f 81}\right)=$ estimated overcoverage (number of "persons'") in the Census, Birth and Immigrant frames which was not detected in the 1981 RRC operations and is represented in $\hat{M}_{81}$.

Note: There is a classification, response, sampling and "no trace" error component associated with each item of equation $2 ;$ e.g., $\bar{C}_{c / r e 81}$ and $\hat{M}_{\text {c/re81 }}$. These taken in total sum to zero. In the above equations these error components exclude error caused by overcoverage and overcoverage which results in a "not traced"; e.g., non-existent persons enumerated in the previous Census. The effect of overcoverage is included, for example, in $\hat{C}\left(\delta_{n f 81}\right)$ and $\hat{M}\left(\hat{O}_{n f 81}\right)$.

Similarly,

$$
e\left(\hat{M}_{76}\right)=\hat{M}_{76}\left[e\left(\hat{M}_{71}\right)\right]-\hat{M}_{76}\left(\hat{O}_{n f 76}\right)+\hat{M}_{c / r e 76}+M_{n 76}
$$

Error and part of the difference $\hat{C}-C$ can be passed from one RRC to another through the Missed frame and through overcoverage in the Census frame. This error could account for a large part of the difference $\hat{C}_{81}-C_{81}$. The effect on $\hat{C}-C$ may be much greater than on $\hat{M}$.

The rate of net coverage error in the 1981 Census, for the ten provinces, would be equal to:

$$
\frac{M_{81}+M_{n 81}-O_{81}}{C_{81}+M_{81}+M_{n 81}-O_{81}}
$$

and the rate of undercoverage would be:

$$
\frac{M_{81}+M_{n 81}}{C_{81}+M_{81}+M_{n 81}-O_{81}} .
$$

The estimator used in the RRC is

$$
\frac{\hat{M}_{81}}{C_{81}+\hat{M}_{81}}
$$

Even a relatively small value of $M_{n 81}-O_{81}$ could contribute significant bias to the results of the RRC, if these results are used as estimates of net coverage error. A relatively small value of $e\left(\hat{M}_{81}\right)$ could contribute significant bias to the RRC undercoverage estimates: two potential elements of bias coming from the previous RRC; one from any misclassification within the RRC; and one from "missed" persons among those not included in an RRC frame. There may be, of course, some cancellation among these elements.

An alternative estimator would be to use $\hat{C}_{81}$ instead of $C_{81}$ in the denominator. There are specific and not unlikely circumstances under which the use of $C_{81}$ would produce estimates with less bias at the national level. These circumstances, which involve the relative sizes of $C_{81}-C_{81}, O_{81}$ and $M_{n 81}$ do not hold, however, for provinces or estimates for which the Census count of enumerated is less than the RRC estimate of enumerated.

## REFERENCES

BURGESS, R.D. (1986). Major issues and implications of tracing survey respondents. International Symposium on Panel Surveys, Washington D.C.
GOSSELIN, J.-F. (1976). The Methodology of the 1971 Reverse Record Check. Survey Methodology, 2, 180-193.
PLAUT, G. (1985). Refugee determination in Canada. House of Commons Standing Committee on Labour, Manpower and Immigration.
STATISTICS CANADA (1976 to 1981). Vital Statistics Volume III: Death/Mortality Catalogue 84-206, Statistics Canada.
STATISTICS CANADA (1979). Mobility Status and General Population Characteristics. Catalogue 92-834, Statistics Canada.

STATISTICS CANADA (1982). Population Counts - 1976 and 1981 - Federal Electoral Districts. Catalogue 99-908, Statistics Canada.
STATISTICS CANADA (1983a). Mobility Status. Catalogue 92-907, Statistics Canada.
STATISTICS CANADA (1983b). Reverse Record Check Tabulations, 1981 Data Quality Project, 1981 Census of Population, unpublished.
STATISTICS CANADA (1984). Public Use Reverse Record Check Tape Users Guide. 1981 Census of Population, Statistics Canada.
STATISTICS CANADA (1987). Population Estimation Methods, Canada. Catalogue 91-528E, Statistics Canada.
STATISTICS CANADA (1988). Undercoverage Rates from the 1986 Reverse Record Check. User Information Bulletin, Number 2, Statistics Canada.

# A Demographic Approach to the Evaluation of the $\mathbf{1 9 8 6}$ Census and the Estimates of Canada's Population 

ANATOLE ROMANIUC ${ }^{\mathbf{1}}$


#### Abstract

A significant increase in coverage error in the 1986 Census is revealed by both the Reverse Record Check and the demographic method presented in this paper. Considerable attention is paid to an evaluation of the various components of population growth, especially interprovincial migration. The paper concludes with an overview of two alternative methods for generating postcensal estimates: the currently-in-use, census-based model, and a flexible model using all relevant data in combination with the census.


KEY WORDS: Census undercoverage; Population estimates; Demographic component method.

## 1. INTRODUCTION

The accuracy of the census, and of the postcensal population estimates based thereon, is an important issue in its own right. The use of population numbers in the formulae for calculating revenue transfers between various levels of government, makes the question of accuracy all the more critical and politically sensitive (Fellegi 1980; Romaniuc and Raby 1980). The intense debates on whether or not to adjust population counts for census undercoverage in Canada and the USA, and several judicial litigations fought in the latter country in recent years, are indications of both the political importance and the technical complexity of the issue.

Yet, in spite of all that has been written on the subject, the elaborate arguments marshalled by both those for and those against adjustment, the debates remain inconclusive (Keyfitz 1979 and 1981; Kish 1980; Spencer 1980; Freedman and Navidi 1986; Stoto 1987). Eventually Statistics Canada decided (as did the US Department of Commerce) against adjustment for census undercoverage, while at the same time reaffirming its long-standing commitment to the policy of data quality evaluation (Wilk 1981). By making public both the evaluation results and the underlying methodology, the users can make adjustments to suit their particular needs, in full knowledge of the strengths and limitations of the census counts and estimates. It is in the spirit of this policy on quality evaluation that this paper has been written.

There are basically two approaches to the evaluation of the accuracy of census counts. One is the "micro"' approach, involving individual verification, case-by-case record matching, in order to identify persons who have been missed, enumerated more than once, or enumerated even though, by definition, they are not part of the census universe. To this type of evaluation belong the US Bureau of the Census Post-Enumeration Program and Statistics Canada's Reverse Record Check (RRC).

[^2]The second is the "macro" evaluation approach involving an analysis at aggregate levels, such as comparison of the census counts with figures derived from independent sources or with estimates arrived at by means of statistical and demographic methods. Following the pioneering work by Ansley Coale (1955), the demographic techniques of analysis have been used by the US Bureau of the Census to evaluate census coverage concurrently with the Post-Enumeration Program (see most recent report by Fay, et al. 1988). Some earlier attempts of this kind in Canada were also made (Lapierre 1970). The essence of the demographic method, as we shall see later, is that it brings to bear the formal relationship between population and its growth components - namely births, deaths and migration.

The evaluation of the 1986 Census coverage through the Reverse Record Check (RRC) has been carried out and reported upon elsewhere (Carter 1988; Statistics Canada 1988). It suffices to say that the RRC-based estimates of undercoverage are subject to sampling error which can be quite significant for provinces with a small population - and to biases of unknown magnitudes (difficulties in tracing persons or matching individual records). Furthermore, the RRC has been designed primarily to measure undercoverage. The measurement of overcoverage has been attempted on an experimental basis, but at the time of writing, the results were unavailable. For these and similar reasons, an alternative assessment of the accuracy of the census counts becomes all the more important.

This paper evaluates, by means of demographic analysis, the accuracy of the three most recent censuses, with emphasis on the 1986 Census. A three-step operation is followed. First, census counts and population estimates are compared with each other. Second, demographic techniques are used to generate alternative estimates of census undercoverage which are, in turn, compared with those based on the Reverse Record Check. As a third and final step, the focus of evaluation is shifted from census counts to intercensal change in population. Two sets of independent estimates of intercensal population change are produced. One is based on the two consecutive censuses, while the other is obtained directly from data on births, deaths and migration.

Before proceeding with the actual evaluation, a word of caution is in order. Though of acceptable quality for most of the uses they serve, neither census counts nor population estimates are perfect. Indeed, there is no one set of data deemed to be perfect enough to serve as a benchmark for the validation of other data. The statistical reality is that data are imperfect in varying degrees. The fine tuning and high precision that would be required for particular uses - such as government allocations and revenue transfers referred to earlier - might not be attainable under the present state of the art. However, we hope that this evaluation, using a combination of statistical tools, imperfect as they may be, will enable us to get some sense of the direction and magnitude of errors and biases affecting census population counts and various components of population estimates. Such an undertaking will hopefully set the stage for improvements as we work toward the 1991 Census and the post-1991 population estimation methodology.

## 2. CENSUS COUNTS VERSUS POPULATION ESTIMATES: ERROR OF CLOSURE

The postcensal estimates of population are obtained, as per equation 1, by the so-called component method, whereby births and immigrants are added to, and deaths and emigrants are subtracted from, the base census population. The net interprovincial migration is then added to estimate population by province. The procedure is repeated annually over the five-year period
to the next census. The current estimation methodology calls for postcensal estimates to be revised retrospectively so as to bring them in line with the latest census counts (Statistics Canada 1987). The difference, as per equation 2, between estimates thus arrived at and census counts is termed "the error of closure" (EC).

$$
\begin{gather*}
\hat{P}_{t}=R_{t-5}+\left[B_{t-5, t}-D_{t-5, t}+I_{t-5, t}-\hat{E}_{t-5, t}+\hat{N}_{t-5, t}\right]  \tag{1}\\
E C(\%)=\frac{\hat{P}_{t}-R_{t}}{R_{t}} \times 100 \tag{2}
\end{gather*}
$$

where:

$$
\begin{array}{ll}
\hat{P}_{t} & =\text { estimated population at time } t ; \\
R & =\text { census counts at time } t \text { or } t-5 \text { as the case may be; } \\
B \quad & \text { number of births; } \\
D \quad & =\text { number of deaths; } \\
I \quad & \text { number of immigrants; } \\
\hat{E} \quad \text { = number of emigrants as estimated; } \\
\hat{N} \quad \text { = net interprovincial migration as estimated; } \\
t-5, t \quad \text { indicates the five-year period during which the events occurred. }
\end{array}
$$

Table 1 presents the error of closure for the last four censuses for Canada, provinces and territories. On the whole, agreement between the census counts and the population estimates is fairly good even for provinces. This is all the more remarkable considering the fact that, in the absence of direct records, both emigration from Canada and interprovincial migration have to be estimated from administrative data (family allowance and income tax files).

Despite the high level of agreement, there are two salient features in the error of closure. One such feature is the jump to nearly one percent error of closure in 1986, a relatively large error when compared to that in the previous censuses. For the 1971 and 1976 censuses the error stood at slightly over one-half of one percent and only at one-quarter of one percent in 1981. The other feature is the negative error of closure in 1981. Whereas in the other three censuses, the estimates exceeded the census counts, in 1981 the former fell short of the latter. Almost all of this shortfall originated in the province of Alberta.

Turning to the provinces, one notes a consistently positive error of closure in 1986, whereas the sign of the error varied in the previous three censuses. Furthermore, for most of the provinces, the magnitude of the error has increased in 1986 as compared to the previous three censuses. The larger errors of closure were found in the Maritime Provinces and Quebec, and the smaller in Ontario and in the Western Provinces, with the exception of Saskatchewan.

The 1981 case of Alberta, referred to above, calls for some further remarks. In 1981, this province had to contend with an unusually large negative error of closure: the estimates fell short of the census count by 53,886 individuals or $2.41 \%$. There are two possible explanations for this outcome. One is that the 1981 Census in this province may have suffered from a

## Table 1

Error of Closure: Canada, Provinces and Territories, June 1971, 1976, 1981 and 1986

| Geographic Area | Percent Error $^{\prime}$ |  |  |  |
| :--- | ---: | ---: | ---: | ---: |
|  | 1971 | 1976 | 1981 | 1986 |
|  | 0.51 | 0.58 | -0.25 | 0.95 |
| Newfoundland | 0.32 | -0.19 | 1.25 | 2.02 |
| Prince Edward Island | -0.76 | 1.58 | -0.31 | 1.06 |
| Nova Scotia | -2.45 | 0.93 | -0.03 | 1.28 |
| New Brunswick | -0.44 | 1.51 | -0.28 | 1.57 |
| Quebec | 0.08 | 0.10 | -0.58 | 1.34 |
| Ontario | 1.41 | 1.07 | 0.37 | 0.73 |
| Manitoba | -0.01 | 1.21 | 0.83 | 0.57 |
| Saskatchewan | 0.21 | 0.91 | -0.52 | 1.06 |
| Alberta | 0.31 | -0.09 | -2.41 | 0.81 |
| British Columbia | 0.47 | -2.07 | -0.22 | 0.58 |
| Yukon | -6.63 | -0.92 | -2.11 | -4.66 |
| Northwest Territories | 3.14 | -5.60 | -1.32 |  |

${ }^{1}$ Population Estimate - Census Count $\times 100$
Census Count
Source: Demography Division, Statistics Canada.
relatively large "overcount'". Prompted by the booming oil-based economy, a great number of transient job-seekers from other provinces made their way to Alberta, some of whom may have been incorrectly enumerated as this province's usual residents. Yet, the fact that for 1981 Alberta showed an above-average undercount ( $\mathbf{2 . 5 4 \%}$ ) only adds to the puzzle. The other possible explanation is that the flow of in-migrants to Alberta, in those days of its economic prosperity and demographic boom, was not fully captured by the family allowance and taxation files - the basis of interprovincial migration estimates. In other words the large shortfall in the 1981 estimates of population might have resulted from an understatement of the net migration to Alberta.

Having demonstrated that the gap between estimates and counts widened significantly in 1986, the question to be addressed in the subsequent sections is whether this is due to the deterioration of: (a) the census coverage or (b) the data on the components of population growth over the last intercensal period.

## 3. DEMOGRAPHICALLY-DERIVED UNDERCOVERAGE RATE

By adjusting the census base population for undercoverage as estimated from the RRC, and by adding the net population increase (births, deaths and migrants) over the subsequent postcensal period, one obtains, as per equation 3, the population at the time of the next census. We shall call this the expected population, to differentiate it from the estimated and enumerated populations dealt with in the previous section.

$$
\begin{equation*}
P_{t}^{\prime}=\left[R_{t-5}+\hat{U}_{t-5}\right]+\hat{G}_{t-5, t} \tag{3}
\end{equation*}
$$

where:
$P_{t}^{\prime}=$ expected population at time $t$;
$R_{t-5}=$ enumerated population at time $t-5$;
$\hat{U}_{t-5}=$ the number of individuals missed in the census $t-5$, as estimated through the Reverse Record Check (RRC);
$\hat{G}_{t-5}=$ estimates of net population change over the intercensal period $t-5, t$ (births, deaths and migrants in equation (1)).

The difference, $U_{t}^{\prime}$, between the expected population, $P_{t}^{\prime}$, and the enumerated population, $R_{t}$, as per equation 4, can be taken here as a coverage error. We shall call this the demographic estimate of coverage error.

$$
\begin{equation*}
U_{t}^{\prime}=P_{t}^{\prime}-R_{t} \tag{4}
\end{equation*}
$$

And the rate of coverage error, $u_{t}^{\prime}$, is simply the ratio of the demographically estimated error of coverage, $U_{t}^{\prime}$, to the expected population, $P_{t}^{\prime}$ :

$$
\begin{equation*}
u_{t}^{\prime}=\frac{P_{t}^{\prime}-R_{t}}{P_{t}^{\prime}}=\frac{U_{t}^{\prime}}{P_{t}^{\prime}} \tag{5}
\end{equation*}
$$

For comparison, the undercoverage rate as estimated through the RRC stands as follows:

$$
\begin{equation*}
\hat{u}_{t}=\frac{\hat{U}_{t}}{R_{t}+\hat{U}_{t}} \tag{6}
\end{equation*}
$$

How do the demographically estimated error of coverage and the RRC-estimated undercoverage compare? First, it should be stressed that both are subject to error and bias. The former is affected by: (a) the lack of an estimate of overcoverage; (b) the biases in the RRC-based undercoverage $\hat{U}$ at $t$ and $t-5$ censuses, and; (c) the biases involved in the estimates of intercensal net population change $\hat{G}_{t-s, t}$, particularly its migration component. The RRC estimate of undercoverage is affected by: (a) sampling error, and; (b) various biases due to tracing of individuals, record matching, etc. Furthermore the undercoverage rate, $\hat{u}$, as per formula (6), is slightly downwardly biased because $R_{r}$ in the denominator includes an overcount of unknown quantity. Hence, alone on these grounds, comparison between the two coverage measurements is far from being straightforward.

But there are conceptual differences as well. The RRC estimate is a pure undercoverage measurement. Demographically estimated coverage error is a more complex, difficult to define unequivocally, entity. It is neither an undercoverage nor a net undercoverage. In order, to better grasp the relationship between the two, the equation (3) of the expected population, $P_{f}^{\prime}$, may be rewritten as per (7). Note that the enumerated population, $R$, is now expressed in terms of its two components: those who were correctly enumerated, $R^{\prime}$, and those who were overcounted, $O$.

$$
\begin{equation*}
P_{t}^{\prime}=\left[\left(R_{t-5}^{\prime}+O_{t-5}\right)+\hat{U}_{t-5}\right]+\hat{G}_{t-5, t} \tag{7}
\end{equation*}
$$

The undercoverage rate estimated by the demographic method as expressed in equation (5) now becomes:

$$
\begin{equation*}
u_{t}^{\prime}=\frac{\left[\left(R_{t-5}^{\prime}+O_{t-5}\right)+\hat{U}_{t-5}+\hat{G}_{t-5, t}\right]-\left(R_{t}^{\prime}+O_{t}\right)}{\left(R_{t-5}^{\prime}+O_{t-5}\right)+\hat{U}_{t-5}+\hat{G}_{t-5, t}} \tag{8}
\end{equation*}
$$

It follows from (8) that the overcoverage affects both the expected and the enumerated populations. Consequently, the demographic rate of undercoverage reflects the combined effect of the undercoverage per se and the difference in the overcoverage, 0 , of the base census, $t-5$, and terminal census, $t$. Assuming that both (a) the RRC-based undercoverage, $\hat{U}$ at $t$ and $t-5$, and (b) the population change (the net sum of the components) for intercensal period, $\hat{G}_{t-5, t}$, are correctly estimated, then the demographic coverage rate, $u_{t}^{\prime}$, and the RRC rate, $\hat{u}_{t}$, will vary numerically depending on the level of the overcoverage of censuses at time, $t-5$ and $t$, so that if $O_{t} \geq O_{t-5}$ then $\hat{u}_{t} \frac{<}{>} u_{t}^{\prime}$.

Having clarified the conceptual particularities of the two measures of coverage error, we now turn to Table 2 which presents for Canada the coverage estimates for the 1981 and 1986 censuses. Both estimates reveal a significant increase in the coverage error in the 1986 Census. However, the demographically-derived rate of coverage error is consistently lower than the RRC rate of undercoverage: $2.82 \%$ and to $3.21 \%$ for 1986 , and $1.70 \%$ and $2.01 \%$, for 1981 , respectively. This could mean that the overcoverage was higher in 1981 than in 1976, and higher in 1986 than in 1981, on the condition that the assumptions underlying the identities are correct. But there are no data to either confirm or deny the validity of these assumptions.

The estimates of coverage error by the two methods - demographic and RRC - by province in Table 2 are portrayed by Figure 1(a) and 1(b). The explanation of the differences at the provincial level is liable to present even greater uncertainties because the error and biases,

Table 2
Demographic and Reverse Record Check Estimates of Undercoverage Rates: By Provinces, 1981 and 1986

| Geographic Area | Demographic Method |  | Reverse Record Check ${ }^{1}$ |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $\begin{gathered} 1981 \\ (\%) \end{gathered}$ | $\begin{aligned} & 1986 \\ & (\%) \end{aligned}$ | 1981 |  | 1986 |  |
|  |  |  | (\%) | (\%) | (\%) | (\%) |
| Canada |  |  |  |  |  |  |
| (Territories not included) | 1.70 | 2.82 | 2.01 | (0.09) | 3.21 | (0.12) |
| Newfoundland | 2.29 | 3.60 | 1.74 | (0.95) | 2.01 | (0.32) |
| Prince Edward Island | 0.05 | 2.10 | 1.17 | (0.54) | 2.16 | (0.80) |
| Nova Scotia | 0.82 | 2.22 | 1.05 | (0.34) | 2.63 | (0.38) |
| New Brunswick | 1.83 | 3.28 | 1.81 | (0.30) | 2.83 | (0.36) |
| Quebec | 2.31 | 3.13 | 1.91 | (0.21) | 3.06 | (0.29) |
| Ontario | 1.81 | 2.53 | 1.94 | (0.14) | 3.40 | (0.19) |
| Manitoba | 1.88 | 1.44 | 0.98 | (0.35) | 2.22 | (0.40) |
| Saskatchewan | 0.76 | 2.00 | 0.99 | (0.37) | 2.51 | (0.36) |
| Alberta | -1.18 | 3.09 | 2.54 | (0.36) | 2.75 | (0.33) |
| British Columbia | 2.62 | 3.55 | 3.16 | (0.33) | 4.49 | (0.39) |

[^3]referred to above, at these levels are expected to be larger than they are at the national level. This is true in particular for sampling error in the case of the RRC undercoverage estimates, and for the biases in the interprovincial migration affecting net intercensal population change in the case of the demographic estimates of coverage error.

With the above comments regarding the biases and conceptual differences in mind, let us see how consistent are the two coverage measures at the provincial level? To this end, the following criterion of consistency is posited: if the two measures of coverage were conceptually identical and empirically correct, their respective correlation points in space should line up along the $45^{\circ}$ bisectrix.

For the 1981 Census, disregarding the special case of Alberta referred to earlier (and also P.E.I. heavily affected by the sampling error), the correlation points follow closely the theoretical $45^{\circ}$ straight line. The discrepancies are small: in most cases they are not statistically significant given the standard deviation affecting the RRC estimates (see Table 2).

For the 1986 Census, six provinces out of ten (Saskatchewan, Nova Scotia, Prince Edward Island, Quebec, Alberta and New Brunswick) have their respective points falling within close range of the $45^{\circ}$ bisectrix and thus meet the consistency test. One, Newfoundland, falls far afield on the left side, suggesting a possible understatement of the RRC undercoverage rate for this province. Manitoba, Ontario and British Columbia fall well to the right side of the $45^{\circ}$ bisectrix suggesting a possible overstatement of the RRC undercoverage or understatement of demographic coverage rate.

It should be stressed once again that the analysis of the accuracy of census coverage has been hampered by the lack of information on overcoverage. Yet, it is fair to say that notwithstanding its limitations, the analysis strongly points to a deterioration of the 1986 census coverage.

## 4. CENSUS AND COMPONENT-BASED INTERCENSAL POPULATION CHANGE: A CHECK FOR CONSISTENCY

The task now at hand is to compare two sets of independent estimates of the intercensal net population change: one set based on demographic components (births, deaths and migration), the other set derived from two consecutive censuses, unadjusted and adjusted for undercoverage. Refer to the former as component-based estimates and to the latter as census-based estimates of intercensal net population change.

$$
\begin{align*}
& \hat{G}_{t-5, t}=B_{t-5, t}-D_{t-5, t}+I_{t-5, t}-\hat{E}_{t-5, t}+\hat{N}_{t-5, t}  \tag{9}\\
& \bar{G}_{t-5, t}=R_{t}-R_{t-5}  \tag{10}\\
& \bar{G}_{t-5, t}^{\prime}=\left(R_{t}+\hat{U}_{t}\right)-\left(R_{t-5}+\hat{U}_{t-5}\right) . \tag{11}
\end{align*}
$$

All the above notations have been made explicit in the previous formulae.
Two independently-produced estimates might be construed as reasonably trustworthy if they are similar for a given point in time. As seen in Table 3, the difference between census-based and component-based estimates is only about $5 \%$ for the 1976-81 period. For the 1981-86 period, the two estimates differ by a substantial margin of $19 \%$ if unadjusted, and by $8 \%$ if adjusted for undercoverage.


Figure 1. Relationship between Undercoverage Rates as Estimated by Reverse Record Check and Demographic Method, 1986 Census


Figure 2. Relationship between Undercoverage Rates as Estimated by Reverse Record Check and Demographic Method, 1981 Census

Table 3
Ratio Between Census and Component-Based Intercensal Change in Population: By Province, 1976-81 and 1981-86

| Geographic Area | Ratio between Census and Component-based Intercensal Population Change Multiplied by 100 |  |  |  |
| :---: | :---: | :---: | :---: | :---: |
|  | 1981-86 |  | 1976-81 |  |
|  | Not adjusted for Census Undercoverage | Adjusted for Census Undercoverage | Not adjusted for Census Undercoverage | Adjusted for Census Undercoverage |
| Canada | 80.9 | 108.4 | 104.5 | 106.1 |
| (Territories not included) |  |  |  |  |
|  |  |  |  |  |
| Prince Edward Island | 76.7 | 101.6 | 109.8 | 135.7 |
| Nova Scotia | 70.4 | 110.3 | 101.3 | 111.1 |
| New Brunswick | 55.6 | 86.7 | 111.3 | 99.2 |
| Quebec | 54.2 | 97.1 | 122.7 | 83.8 |
| Ontario | 88.1 | 115.2 | 92.0 | 103.2 |
| Manitoba | 89.2 | 117.3 | 35.6 | 29.0 |
| Saskatchewan | 79.4 | 110.3 | 112.0 | 105.5 |
| Alberta | 88.8 | 94.5 | 115.6 | 124.4 |
| British Columbia | 89.5 | 118.3 | 102.2 | 105.8 |

Note: The procedure cannot be applied for the period 1971-76 because, for this and earlier periods, emigration has been estimated residually from the two consecutive censuses and the remaining growth components (births, deaths and immigrants).
Source: Demography Division, Statistics Canada.

The comparison by province is a more delicate matter. On the components side, one has to contend with the reliability of the interprovincial migration estimates. On the census side, one must reckon with the variability of biases in undercoverage and overcoverage, and sampling errors in the RRC undercoverage estimates. Sampling errors alone could account for up to $15 \%$ of variations in the ratio between the two estimates of the intercensal population change for some provinces. Any variations beyond this level are more likely to have been induced by errors and biases from other than the sampling.

Hence, in the absence of a more trustworthy criterion, we have set $\pm 15 \%$ as a tolerance limit for the discrepancies between the two estimates. The tolerance limit thus set, has at least the merit of screening out highly questionable cases.

With these qualifications in mind, let's turn to Table 3, which compares by province, census and component-based population changes for the last two intercensal periods. Six provinces out of ten for the 1976-81 period, and four out of ten for the 1981-86 period meet the somewhat arbitrarily set tolerance test. In general, the discrepancies are wider for the 1981-86 period than for 1976-81. Particularly conspicuous in this regard are the provinces of Newfoundland, Quebec, and New Brunswick.

Newfoundland's census-based 1981-86 population change represents only $5 \%$ of that derived from the components. It is still only $19 \%$, even after adjustment for undercoverage. Such a low population growth would call for a net migration loss of about 26,000 over the 5 -year period. Yet, all the three sources of interprovincial migration (Family Allowance, Taxation and the census mobility question) place these losses in the range of 14,800 to 16,500 (see Table 5).

Similar inconsistencies are found in the case of Quebec. The census-based population growth for the period 1981-86, which represents only $64 \%$ of the component-based growth, would imply Quebec's loss through out-migration to be twice the amount estimated by Statistics Canada, that is, 160,000 instead of 80,000 . Yet again, all the three sources of information put the net-migration losses in the range of 63,000 to 81,000 over the 5 -year period. The gap between the two estimates of intercensal change is almost wiped out when the 1981 and 1986 census counts are adjusted for undercoverage.

The case of New Brunswick is similar to that in Quebec and Newfoundland. The censusbased estimate of population growth for the 1981-86 period suggests a net loss through outmigration of 11,200 , whereas the family allowance-based figure is 2,200 . The census mobility question and taxation figures are even lower, 1,376 and 65 , respectively. Adjustment for undercoverage would bring New Brunswick's two estimates of the intercensal population change well within the tolerance limit.

What, then, can be concluded from the above analysis regarding the intercensal population change? It appears that both the components and the census generate reasonably consistent estimates of population change for the 1976-81 period. The discrepancies are small, within a tolerable limit for Canada and for most of the provinces. This, however, is not the case for the most recent intercensal period, 1981-86. Something seems to have deteriorated and the question remains as to whether it is the census or the components of population growth. As was seen in the preceding section, the 1986 Census experienced a significant increase in undercoverage estimated by two different methods. Adjustment for undercoverage, however, did not always produce better estimates of intercensal population growth, in fact the opposite happened in some cases. In the next section, we take a closer look at the components of population growth.

## 5. HOW GOOD ARE THE COMPONENTS OF POPULATION GROWTH?

What follows is a brief assessment of the quality of the data on births, deaths, immigration, emigration, and interprovincial migration. For a more complete account of the data on those components, and methodologies for estimating migration, the reader is referred to the 1987 Statistics Canada publication "Population Estimation Methods, Canada".

The registration of births and deaths is deemed to be complete in this country. Deaths or births that somehow escape registration must be by necessity very small in number in view of the prevailing regulations (need for a burial certificate) and the material (family allowance) incentives and legal requirements for registering births. Some late registration may occur, but the numbers are small. For the $1981-85$ period, 3,831 or $0.02 \%$ of all births and 2,528 , or $0.03 \%$, of all deaths were registered beyond the cut-off date. This makes a net of only 1,303 persons unaccounted for in the population estimates.

Immigration statistics are regarded as reasonably accurate to the extent one speaks here of landed immigrants. The distribution of immigrants by province is based on their intended destination rather than on where they actually settle. It is, however, noteworthy, as per Table 4, that this distribution closely agrees with the 1986 Census distribution of immigrants.

Compared to the three other components reviewed above - births, deaths and immigration - interprovincial migration and emigration are weaker links in equation (1) which is used for estimating population for postcensal years. There are indeed no direct records of internal migration or emigration. Such figures must be estimated indirectly from administrative files

Table 4
Percentage Distribution of Immigrants by Province Based on the 1981 Census and Immigration Records of Intended Destination in 1980

| Geographic Area | Immigration Records | Census |
| :--- | :---: | ---: |
|  |  |  |
| Newfoundland | 0.4 | 0.3 |
| Prince Edward Island | 0.1 | 0.1 |
| Nova Scotia | 1.1 | 1.0 |
| New Brunswick | 0.8 | 0.8 |
| Quebec | 43.7 | 15.0 |
| Ontario | 5.4 | 42.7 |
| Manitoba | 2.5 | 5.4 |
| Saskatchewan | 13.2 | 2.6 |
| Alberta | 17.2 | 14.5 |
| British Columbia + Yukon | 100.0 | 17.6 |
| + Northwest Territories |  | 100.0 |
| Canada |  |  |

Source: Demography Division, Statistics Canada.

- family allowance and income tax - which contain information on changes of residence. They deserve, therefore, more than a cursory consideration. In what follows, we shall focus on the significant methodological and data improvements achieved in recent years, as well as address certain persistent shortcomings inherent to these estimates. For a more complete account see Chapters IV and V of the Population Estimation Methods, Canada, 1987.

While family allowance data have been used since 1956 , the most significant innovation to the system for estimating interprovincial migration was the addition of personal income tax data in 1976. As of 1981, a "two-track" estimation system was implemented: the preliminary quarterly and annual estimates based on family allowance data, and the final annual estimates based on taxation data. Both these data sources have strengths and weaknesses.

The main advantage of the family allowance file lies in its timeliness and fairly high accuracy. The information on change of address is available two months after the fact. The accuracy of the file is contingent upon two factors. The first is the comprehensiveness of coverage of child population, as every child under 18 years of age, supported by a parent, is entitled to a monthly payment. The second is the financial incentive for the beneficiaries of family allowances to report any change of address as soon as it occurs. The family allowance file does not, however, provide information on adult migration. This has to be estimated indirectly, by applying a conversion factor, ' $f$ ', which is obtained by calculating the ratio of the adult migration rate to the child migration rate from the taxation data available for the most recent year.

Given the key importance of the $f$ factor in the estimation formulae, a few comments are called for. Prior to 1971, the value of $f$ was based on 5 -year migration data from the most recent census. As the annual age-specific data on migrants became available from income tax records, the decision was made to use such data since they have an advantage over census data in that they reflect a more recent age pattern of migration.

Another innovation is worth mentioning. Prior to 1981, the $f$ factor was calculated only by province of origin. However, with the availability of relevant data from taxation, it became evident that this factor also varies significantly by province of destination. Consequently, the decision was made to calculate the $f$ factor by both province of origin and province of destination.

Turning now to the personal income tax file as the data source for estimating interprovincial migration, the following assessment is in order. As compared to the family allowance file, the taxation file has the advantage of having a much broader demographic base: tax filers and their dependents represent roughly $90 \%$ of the population. However, there are various sources of potential errors and biases. Information on tax filers' dependents must be imputed from the dollar value of total exemptions. Various assumptions have to be made in imputing the migratory status of the tax filers' dependents, as well as that of persons who are neither filing income tax returns, nor are dependents upon those who do so, and therefore are not covered at all by the taxation system. This is particularly the case for young adults and the elderly, who may be more prone to neglect to file their tax-return or who may not earn the minimum income required for filing. Such differential age-related biases, if indeed present, affect the estimates of the age structure, and this in turn affects the value of the $f$ factor, used in the family allowance-based preliminary estimates of interprovincial migration.

Table 5 presents figures on net interprovincial migration for the intercensal 1981-86 period based on family allowance, taxation, and the census question on residence five years ago. Notwithstanding some significant variations in numbers, the three sources of data provide a consistent picture of level of interprovincial net migration over the 5 -year period, by province.

What has been said about interprovincial migration also holds for emigration - Canadians taking residence in another country. Prior to 1981, the aggregate emigration to countries other than the United States and the U.K. (for which data were available through the immigration services of the two countries) had to be estimated residually from consecutive censuses and the components of intercensal population growth. As of 1981, the estimation of the number of emigrants has been based on family allowance and income tax data. The procedure is similar to that described above for estimating interprovincial migration. Child-migration is estimated from family allowance data. To estimate adult emigration, and hence total emigration, a conversion factor, $f$, based on income tax data, is applied to child-emigration. This same procedure applies to both the preliminary and final estimates of emigration, except that in the latter case more complete data are used.

Table 5
Net Interprovincial Migration for the Period 1981-1986, Based on Specified Sources

| Geographic Area | Census ${ }^{1}$ | Family <br> Allowance | Income <br> Tax |
| :--- | ---: | ---: | ---: |
| Canada | 0 | 0 | 0 |
| Newfoundland | $-16,550$ | $-14,837$ | $-15,051$ |
| Prince Edward Island | 1,540 | 293 | 751 |
| Nova Scotia | 6,275 | 5,204 | 6,895 |
| New Brunswick | $-1,370$ | $-2,239$ | -65 |
| Quebec | $-63,295$ | $-76,040$ | $-81,254$ |
| Ontario | 99,355 | 115,497 | 121,767 |
| Manitoba | $-1,555$ | $-3,700$ | $-2,634$ |
| Saskatchewan | $-2,820$ | -668 | $-2,974$ |
| Alberta | $-27,665$ | $-34,073$ | $-31,676$ |
| British Columbia | 9,500 | 13,289 | 7,382 |
| Yukon | $-2,665$ | $-2,381$ | $-2,775$ |
| Northwest Territories | -755 | -345 | -366 |

[^4]Table 6
Estimates of Emigrants by Different Methods, Canada, 1981-86

| Method | $1981-86$ |
| :--- | :---: |
| Residual Method from Censuses <br> (a) Unadjusted for Undercoverage <br> (b) Adjusted for undercoverage |  |
| Revenue Canada Tax File <br> Family Allowance Method (current) (using the $f$ <br> factor from the tax file) | $\mathbf{4 7 6 , 4 0 6}$ |
| Family Allowance Method (proposed) (using the $f$ <br> factor from the immigration file) | 165,272 |
| Reverse Record Check ${ }^{1}$ | 235,481 |

1 Preliminary.
Source: Demography Division, Statistics Canada.
Table 6 compares, for the 1981-86 intercensal period, the estimates of emigration based on the family allowance files with the estimates produced by the various alternative methods. Note that the residually-derived emigration estimates, whether from adjusted or unadjusted census counts, are out of line with the more plausible estimates derived from the administrative files and the Reverse Record Check (RRC).

In brief, significant enhancements have been made to the system used to estimate interprovincial migration and emigration, particularly since 1981 . While it can be surmised that the overall quality of the estimates has improved as a result, no demonstrable proof can be adduced. The family allowance and income tax data are fraught with various shortcomings inherent in any data system that has been designed for administrative rather than for statistical purposes.

## 6. CONCLUSIONS AND EMERGING ISSUES

Statistics Canada's population estimation system rests on two building blocks: (1) Census population counts, and; (2) components of population change, namely births, deaths and migrants. Postcensal estimates are carried forward by adding the components of population change over the subsequent years, to the base population, provided by the census. They are revised retrospectively when the next census counts become available. Thus, the census counts are both the base for the postcensal estimates, and the standard for their post-facto validation. The system has produced timely, reliable and internally consistent population estimates, and over the years has enjoyed a remarkable stability.

Much of its stability can be attributed to the high quality of the Canadian censuses. For Canada as a whole, undercoverage as measured by the Reverse Record Check (RRC) remained almost unchanged, at close to $2 \%$, for three consecutive censuses - 1971, 1976 and 1981. Hence, even if the census fell somewhat short of the "true" population of Canada, it provided a highly reliable basis for gauging population growth.

The 1986 Census marks, however, a departure from the trend, as the rate of undercoverage, estimated by the Reverse Record Check, rose to $3.2 \%$. The 1986 Census understates the population increase over the 1981-86 period by about $20 \%$, if one accepts the component method as the standard of validation. Both the Reverse Record Check and the demographic analysis corroborate the deterioration of census coverage in 1986.

On the population components side of the equation - the other building blocks of the estimation system - records on births, deaths, and landed immigrants are fairly reliable. The interprovincial migration and emigration estimates have benefited from various data and
methodological enhancements, particularly since 1981, as was explained in the preceding section. But, as was also pointed out, they may suffer from various shortcomings inherent in any data sources - such as the family allowance and taxation files - that have been designed for administrative rather than statistical purposes. The estimates of interprovincial migration and emigration remain, along with census undercoverage and overcoverage, the prime sources of possible errors and biases in the postcensal estimates of population by province.

What does the future hold for the estimation system as described above? Can it continue working as it stands, or does it need some major reconceptualization? The apparently higher undercoverage rates of the 1986 Census, and its potential consequences for population estimates, has prompted the discussion of an alternative to the present census-based method of producing estimates. This alternative would no longer necessarily rely on the most recent census as a bench-mark, but instead would use relevant available information, including census counts, undercoverage and overcoverage, as well as administrative records, to generate the "best" possible estimates. In other words, the census counts remain an important ingredient of the estimation process, but not the overriding one; nor would the most recent census necessarily be used, if, say, the counts from the previous census were deemed to be more reliable.

After careful consideration, Statistics Canada has decided that the 1986 Census (unadjusted for undercoverage) would be used for the 1986 postcensal estimates and revision of the estimates for the 1981-86 intercensal period. In other words, the existing estimation procedures were reconfirmed. But at the same time, it was recognized that the evaluation of the census and estimates needed to be stepped up, and that an estimation strategy for the post-1991 Census period needed to be devised. Such an estimation strategy would have to take into account plans and realistic prospects for improvements and enhancements in the following four areas:
(1) 1991 Census coverage;
(2) Measurement of both undercoverage and overcoverage;
(3) Administrative records used for the purpose of population statistics: enhancement of the currently used sources - Family Allowance and Taxation - and the harnessing of new ones, such as Old Age Security and Provincial Health Care Files;
(4) Estimates of migration, particularly those concerning interprovincial migration, returning Canadian residents after a protracted stay abroad, and emigration from Canada.

These raise some fundamental issues concerning the philosophy and policy that ought to govern the working of a statistical system, thus transcending the rather narrow question of adjustment for undercoverage referred to at the outset of this paper. In the census-based conception, the emphasis is on the stability and internal coherence of the estimation system. In the conception of a census-divorced estimation model, a premium is placed on flexibility so as to increase the accuracy of the estimates through the utilization of the relevant available information, but possibly at the price of methodological consistency over time. The resolution of the dilemma between these two conceptions will be greatly influenced by the progress that is achieved in the four areas of statistical endeavour identified above.

## ACKNOWLEDGEMENTS

The author has benefited from discussions with his colleagues in the Demography Division: Gwenaël Cartier, Gilbert Lagrange, Ronald Raby, Robert Riordan, Edward Shin and Ravi Verma. Valuable comments received from an anonymous referee and K.G. Basavarajappa,

David Binder, Malcolm Britton, Dick Carter, Ivan Fellegi and M.P. Singh, are also gratefully acknowledged.

## REFERENCES

CARTER, R.G. (1988). Measuring coverage errors in the census population. Presented at the annual meeting of the Canadian Population Society, the University of Windsor, Windsor, Ontario.
COALE, A.J. (1955). The population of the United States in 1950 classified by age, sex and color. Journal of the American Statistical Association, 50, 16-54.

FAY, R.E., PASSEL, J.S., ROBINSON, G.J., and CONRAN, C.D. (1988). The coverage of population in the 1980 Census. U.S. Department of Commerce, Bureau of the Census.
FELLIGI, I.P. (1980). Should the census count be adjusted for allocation purposes - equity considerations. Proceedings of the 1980 Conference on Census Undercount, U.S. Bureau of the Census, 193-203.

FREEDMAN, D.A., and NAVIDI, W.C. (1986). Regression models for adjusting the 1980 Census, Statistical Science, 1, 3-39.
KEYFITZ, N. (1979). Information and allocation: Two uses of the 1980 Census, The American Statistician, 33, 45-50.
KEYFITZ, N. (1980). Issues in adjusting for the 1980 Census undercount. Paper presented at the Annual Meeting of the American Statistical Association, Detroit.
KISH, L. (1980). Diverse adjustments for missing data. Proceedings of the 1980 Conference on Census Undercount, U.S. Bureau of the Census, 193-203.
LAPIERRE-ADAMCYK, E. (1970). Estimation of net census underenumeration by age and sex using demographic analysis techniques. Working paper, Demographic Analysis and Research Section, Statistics Canada.

ROMANIUC, A., and RABY, R. (1980). The impact of census enumeration on selected Federal/Provincial transfer payments. Demography Division, Statistics Canada.
SPENCER, B. (1980). Issues of accuracy and equity in adjusting for census undercoverage. Paper presented at the Annual Meeting of the American Statistical Association, Detroit.
STATISTICS CANADA (1987). Population estimation methods, Canada. Catalogue No. 91-528E, Statistics Canada.

STATISTICS CANADA (1988). Undercoverage rates from the 1986 Reverse Record Check. User Information Bulletin, No. 2, Otttawa.
STOTO, M.A. (1987). Statement to the Subcommittee on Census and Population, Committee on Post Office and Civil Service, U.S. House of Representatives, San Francisco.
WILK, M.B. (1981). Letter to Mr. H. Breau, Chairman, Parliamentary Task Force on Federal-Provincial Fiscal Arrangements, July 3, 1981.

# Adjusting the 1986 Australian Census Count for Under-Enumeration 

C.Y. CHOI, D.G. STEEL and T.J. SKINNER ${ }^{1}$


#### Abstract

In Australia, population estimates have been obtained from census counts, incorporating an adjustment for under-enumeration in 1976, 1981 and 1986. The adjustments are based on the results of a Post Enumeration Survey and demographic analysis. This paper describes the methods used and the results obtained in adjusting the 1986 census. The formal use of sex ratios as suggested by Wolter (1986) is examined as a possible improvement of the less formal use made of these ratios in adjusting census counts.


KEY WORDS: Census under-enumeration; Post-enumeration survey; Demographic estimates; Sex-ratios.

## 1. INTRODUCTION

The population census provides the basic information from which estimates are made of the population of the nation, each of the eight States and sub-State local government areas. In Australia, these population estimates are required for the determination of the number of seats each State will have in the Federal House of Representatives, the allocation of funds to each State, and the funding of local government authorities. Population estimates are also used in their own right as indicators of population growth and distribution and as denominators for various demographic, social and economic indicators. Because population estimates are used in such important ways, a high level of accuracy is required.

In Australia, it is known that the level of under-enumeration at the census is significant and that this level is related to important variables such as birthplace, geographic area and age/sex. Because of this, an adjustment for under-enumeration is made to census counts used for population estimates.

The adjustment of census counts for under-enumeration is a recent practice in Australia. Prior to the 1976 Census, census counts without adjustment for under-enumeration were used directly for population estimation purposes. The need to make this adjustment was recognised when the 1976 Census count fell considerably below the population estimates for the 1976 Census date which were updated from the 1971 Census, and when the 1976 Post Enumeration Survey (PES) showed a high under-enumeration rate of 2.6 per cent compared with 0.5 per cent in 1966 and 1.3 per cent in 1971. The 1976 PES also showed significant variations in underenumeration between States and Territories, ranging from 4.2 per cent for the Northern Territory to 1.1 per cent for Tasmania. In 1986, the level of under-enumeration is estimated to be 1.9 per cent. As in 1976, there were significant variations between States and Territories. The adjustment of 1976 and subsequent census counts has been well received and no challenges have been raised to the appropriateness of doing so or the accuracy of the methods used. This is in contrast with the high level of controversy experienced in the United States of America on the appropriateness of making adjustments to the 1980 census counts for under-enumeration.

[^5]Data for the assessment of the level of under-enumeration are primarily derived from a census PES. Results of the PES are assessed by comparing these with estimates based on demographic statistics and other independent data such as statistics on school enrolments, on children whose parents receive government family allowances, and on persons registered with the government Medicare insurance system. In Australia, school enrolments for children aged 6-15 years are compulsory and until means-testing was introduced in November 1987, family allowances had been universally paid to mothers of all children of ages less than 17. Medicare insurance is also compulsory and universal for all residents. These independent statistics are therefore helpful as a check of the PES results and demographic estimates.

Although population estimates include an adjustment for under-enumeration, no adjustment is made for other census data. Census counts are published without adjustment.

## 2. THE 1986 POST-ENUMERATION SURVEY

In its five yearly population census, the Australian Bureau of Statistics (ABS) employs census collectors for the delivery of forms to each household and for the collection of completed forms from each household. The census is conducted on the basis of enumerating people where they are located on census night.

This collector-based field system allows the census collection phase to be completed two weeks after the census date. This allows a census PES to be conducted reasonably close to the census date - in 1986 within $4-5$ weeks of census night. Because the PES asks a number of questions requiring detailed answers referring to a person's location on census night, its conduct close to census date minimises recall error and also reduces the number of exclusions due to deaths and overseas travel.

As the PES provides the basis for adjusting the census counts for under-enumeration, it is important that the PES be statistically independent of the census. The Appendix describes the steps taken to ensure independence.

The basic approach adopted in the 1986 PES was to select a sample of people independently of the census through a multi-stage area sample of private dwellings. The information required of each person in the selected households was obtained by personal interview of any responsible adult by trained field staff from the ABS regular interview panel. Matching of PES and census records to determine whether each person in the sample should have been included in the census and how many times the person was in fact included was undertaken by clerical staff employed in the Census Data Transcription Centre. The procedures used are described in the Appendix.

From the survey, the ratio of the number of persons who should have been included in the census ( $x$ ) to the number of persons who were estimated to have been in fact included ( y ) can be estimated. This ratio is the net adjustment factor which accounts for both over and underenumeration of individuals.

This adjustment factor, after weighting, is then applied to the actual census count (Y) to produce an estimate of the population (X), i.e. $X=Y(x / y)$.

To allow for differences in expected and actual sample take in the PES, this procedure was applied at the age ( 5 year groups), sex and geographic area (capital city statistical division/rest of State) level. PES estimates are produced on both an actual location at the census date and usual residence basis. The estimation also includes an adjustment for the small level of noncontact and non-response in the PES. For example the estimate of usual residence population for geographic area (s) and age sex cell (a) is:

$$
X_{s a}=Y_{s a} x_{s a} / y_{s a}
$$

where

$$
x_{s a}=\sum_{g c} \frac{D_{g c}+d_{g c}}{D_{g c}} \cdot \frac{x_{s a g c}}{f_{g}}
$$

and

$$
y_{s a}=\sum_{g c} \frac{D_{g c}+d_{g c}}{D_{g c}} . \frac{y_{s a g c}}{f_{g}}
$$

In these estimation formulae the subscript $c$ denotes the response status of the PES dwelling in the census and the subscript $g$ denotes the geographic area in which the person was selected in the PES. $D_{g c}$ is the number of responding dwellings and $d_{g c}$ is the number of non-contact/non-responding dwellings in area $g$ and census response category $c$. The sampling fraction varies between states and is denoted $f_{g}$.

In this form the estimator is a post-stratified ratio estimate. Ignoring for the moment that people may be enumerated in the census incorrectly or more than once, the estimator is the estimator obtained from a dual-record system or a capture-recapture approach discussed, for example, in Bishop, Fienberg and Holland (1975, pp231-234). This is shown in the diagram below where under the assumption of independence the estimate of the total population is $Y$ ( $x / y$ ) which is the ratio estimate $X$.

## PES

Census

|  | Counted | Missed |
| :---: | :---: | :---: |
| Counted | y | Y |
| Missed |  |  |
|  | x |  |

The 1986 PES, however, was designed to collect information on both the number of persons missed by the census and the number of persons over-enumerated, i.e. included in the census erroneously or included more than once. The estimate $X$ takes into account both over and underenumeration at the same time. In this respect, the approach adopted is different from the traditional capture-recapture methodology.

Variance estimation was based on treating $X$ as a ratio estimate derived from a multi-stage sample. The relative standard errors on the PES estimates of the population are given in Table 1. From this table and tables 2 and 4 we see that standard errors are considerably less than the adjustments implied by the PES national age by sex estimates and State by sex estimates.

Table 1
1986 Census: Relative Standard Errors of PES Estimates
of the Population

| Age | Males | Females | Persons |
| :--- | :---: | :---: | :---: |
|  | $\%$ | $\%$ | $\%$ |
| $0-4$ | 0.29 | 0.36 | 0.24 |
| $5-9$ | 0.29 | 0.30 | 0.22 |
| $10-14$ | 0.28 | 0.29 | 0.21 |
| $15-19$ | 0.32 | 0.32 | 0.24 |
| $20-24$ | 0.49 | 0.43 | 0.34 |
| $25-29$ | 0.49 | 0.36 | 0.32 |
| $30-34$ | 0.36 | 0.30 | 0.27 |
| $35-39$ | 0.38 | 0.32 | 0.24 |
| $40-44$ | 0.37 | 0.30 | 0.26 |
| $45-49$ | 0.43 | 0.38 | 0.25 |
| $50-54$ | 0.41 | 0.30 | 0.30 |
| 55-59 | 0.43 | 0.37 | 0.25 |
| 60-64 | 0.53 | 0.41 | 0.29 |
| 65-69 | 0.47 | 0.39 | 0.29 |
| $70-74$ | 0.12 | 0.10 | 0.34 |
| 75+ | Males | Females | 0.31 |
| All ages | $\%$ | $\%$ | 0.08 |
| State | 0.21 | 0.18 | Persons |
|  | 0.23 | 0.21 | $\%$ |
| NSW | 0.27 | 0.24 | 0.14 |
| ICC | 0.27 | 0.20 | 0.16 |
| QLD | 0.29 | 0.25 | 0.19 |
| SA | 0.36 | 1.53 | 0.17 |
| WA | 1.65 | 0.74 | 0.19 |
| TAS | 0.61 |  | 1.25 |
| NT |  |  | 0.55 |
| ACT |  |  |  |

For a more detailed description of the 1986 Post-Enumeration Survey and the estimation procedures, see Appendix.

## 3. DEMOGRAPHIC ESTIMATES OF CENSUS UNDER-ENUMERATION

An alternative method for the estimation of census under-enumeration is through the use of past demographic data including those from previous censuses, births and deaths registers, and overseas migration statistics. For example, estimates of the population at a certain date can be made by updating a previous census using data on births, deaths and overseas migration. The more distant is the previous census which serves as the base, the longer is the time series of reliable vital and migration statistics required, and the less reliance there needs to be on the accuracy of the census base. This is because estimates of persons born after the relevant census date will be affected only by the reliability of data on births, deaths and migration. Internal migration data in Australia are not sufficiently reliable to enable the use of demographic methods for estimating census under-enumeration at sub-national levels. Use of demographic estimates for census evaluation is therefore limited to Australian totals.

Australian data on births and deaths are available as a time series going back to the 19th century and it is unlikely that there have been significant omissions. Successive reports by the Australian Commonwealth Statistician after each population census from 1911 to 1961 claimed that the registration of births and deaths in Australia was substantially complete although it was recognised that some omissions were possible and that there were time lags in registrations. The Statistician's Report was discontinued after 1961. However, there is no evidence that the level of coverage of birth and death registrations has deteriorated since then.

Australia has also maintained comprehensive and reliable statistics on overseas arrivals and departures over a long period of time. These statistics cover all movements including permanent, long term and short-term movements. However, there are several deficiencies in the statistics on overseas arrivals and departures which limit their usefulness for the evaluation of the census data. First, there have been periods in the past when arrivals and departures were suspected of being inaccurately recorded (e.g. during World War II and the period immediately following the war). Second, because of the increase in overseas short-term movements since the 1960 's only a sample (of about 1 in 20) of the arrivals and departure records has been processed for statistical purposes since 1971. Third, errors can occur in the classification of travellers into permanent, long-term and short-term categories. To avoid these errors of classifications the comparison of demographic estimates, census counts and PES estimates of the population at census date is made on the basis of actual location, which include all three categories of overseas movements.

For the assessment of under-enumeration at the 1986 Census, demographic estimates of the population as at census date 1986 by age and sex were made using births, deaths and overseas migration data going back to 1921 together with results of the 1921 Census. Demographic estimates of the population to age 65 years are therefore based solely on birth, deaths and migration data and would not be affected by the accuracy of the 1921 Census.

## 4. VALIDATION OF THE 1986 PES ESTIMATES

The following table shows the estimated population as at 30 June 1986 by age and sex based on demographic analysis and based on the 1986 PES. Medicare enrolments by age and sex are also shown.

There is a very high level of correspondence between PES and demographic estimates of the male population, particularly for those aged under 30 . However, there is a large discrepancy for males aged 30-34, the demographic estimates being 20,000 higher than PES estimates. This can be attributed to a large net gain in the number of males of these ages from short-term movements into and out of Australia in the period 1981-86. Net gains from short-term movements of this magnitude are not detectable in the adjacent age-groups and therefore may reflect some error in overseas arrivals and departures statistics. With the volume of overseas movements being very high (over 6 million in 1986), a small error in reporting of age or in processing can lead to a relatively large discrepancy in the demographic estimate in net absolute term. The possibility of error in demographic estimates is further illustrated by the very high implied under-enumeration rate of 5.3 per cent for this age group compared with much lower rates for the surrounding age group.

It is, of course, quite likely that under-enumeration of overseas visitors was not adequately measured by the PES. However, in either case, errors in estimating the visitor component of the population should not affect the accuracy of official population estimates because these are based on the concept of usual residence and do not include visitors.

Table 2
Estimates of 1986 Population by Age and Sex Based on the 1986 PES and Demographic Analysis, and Medicare Enrolment

| Age | Males ('000) |  |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Population |  |  |  | Difference from Census |  |  | Percent Underenumeration |  |  |
|  | Census <br> (a) | PES(a) | DE(b) | Medicare | PES | DE | Medicare | PES | DE | Medicare |
| 0-4 | 608.3 | 616.4 | 612.8 | 611.4 | 8.0 | 4.5 | 3.1 | 1.3 | 0.7 | 0.5 |
| 0-5 | 594.9 | 602.4 | 603.0 | 612.3 | 7.5 | 8.1 | 17.4 | 1.2 | 1.3 | 2.8 |
| 10-14 | 660.8 | 670.4 | 668.4 | 674.3 | 9.6 | 7.6 | 13.5 | 1.4 | 1.1 | 2.0 |
| 15-19 | 673.1 | 688.4 | 687.7 | 693.1 | 15.3 | 14.6 | 20.0 | 2.2 | 2.1 | 2.9 |
| 20-24 | 648.5 | 679.5 | 681.3 | 681.3 | 31.0 | 32.8 | 32.8 | 4.6 | 4.8 | 4.8 |
| 25-29 | 649.2 | 677.7 | 675.0 | 688.5 | 28.5 | 25.8 | 39.3 | 4.2 | 3.8 | 5.7 |
| 30-34 | 615.5 | 630.0 | 650.1 | 647.5 | 14.5 | 34.6 | 32.0 | 2.3 | 5.3 | 4.9 |
| 35-39 | 622.2 | 634.2 | 632.7 | 646.2 | 12.0 | 10.5 | 24.0 | 1.9 | 1.7 | 3.7 |
| 40-44 | 504.2 | 512.6 | 517.0 | 522.3 | 8.4 | 12.8 | 18.1 | 1.6 | 2.5 | 3.5 |
| 45-49 | 419.8 | 427.0 | 416.5 | 436.8 | 7.2 | -3.3 | 17.0 | 1.7 | -0.8 | 3.9 |
| 50-54 | 363.7 | 371.2 | 371.4 | 377.9 | 7.5 | 7.7 | 14.2 | 2.0 | 2.1 | 3.8 |
| 55-59 | 373.4 | 379.5 | 384.9 | 386.6 | 6.1 | 11.5 | 13.2 | 1.6 | 3.0 | 3.4 |
| 60-64 | 341.1 | 347.0 | 348.1 | 350.6 | 5.9 | 7.0 | 9.5 | 1.7 | 2.0 | 2.7 |
| 65-69 | 259.6 | 263.6 | 251.8 | 265.5 | 4.0 | -7.8 | 5.9 | 1.5 | -3.1 | 2.2 |
| 70-74 | 204.2 | 208.2 | 200.8 | 213.0 | 4.0 | -3.4 | 8.8 | 1.9 | -1.7 | 4.1 |
| $75+$ | 229.5 | 233.0 | 181.5 | 250.1 | 3.5 | -48.0 | 20.6 | 1.5 | -26.4 | 8.2 |
| Total | 7768.3 | 7941.0 | 7883.1 | 8057.3 | 172.7 | 114.8 | 289.0 | 2.2 | 1.5 | 3.6 |
| Females ('000) |  |  |  |  |  |  |  |  |  |  |


| Age | Population |  |  |  | Difference from Census |  |  | Percent Underenumeration |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Census <br> (a) | PES(a) | DE(b) | Medicare | PES | DE | Medicare | PES | DE | Medicare |
| 0-4 | 579.7 | 591.0 | 583.8 | 580.9 | 11.3 | 4.1 | 1.2 | 1.9 | 0.7 | 0.2 |
| 5-9 | 565.1 | 572.4 | 565.5 | 582.1 | 7.3 | 0.4 | 17.0 | 1.3 | 0.1 | 2.9 |
| 10-14 | 628.0 | 636.8 | 630.2 | 641.8 | 8.8 | 2.2 | 13.8 | 1.4 | 0.3 | 2.2 |
| 15-19 | 644.1 | 657.4 | 651.4 | 666.3 | 13.3 | 7.3 | 22.2 | 2.0 | 1.1 | 3.3 |
| 20-24 | 633.1 | 652.5 | 644.4 | 670.4 | 19.4 | 11.3 | 37.3 | 3.0 | 1.8 | 5.6 |
| 25-29 | 648.7 | 660.7 | 665.4 | 684.1 | 12.0 | 16.7 | 35.4 | 1.8 | 2.5 | 5.2 |
| 30-34 | 618.1 | 627.8 | 631.2 | 643.9 | 9.7 | 13.1 | 25.8 | 1.5 | 2.1 | 4.0 |
| 35-39 | 612.1 | 619.1 | 600.2 | 626.3 | 7.0 | -11.9 | 14.2 | 1.1 | -2.0 | 2.3 |
| 40-44 | 482.6 | 488.6 | 489.6 | 495.4 | 6.0 | 7.0 | 12.8 | 1.2 | 1.4 | 2.6 |
| 45-49 | 399.1 | 403.0 | 397.9 | 411.6 | 3.9 | -1.2 | 12.5 | 1.0 | -0.3 | 3.0 |
| 50-54 | 349.1 | 354.6 | 343.9 | 358.6 | 5.5 | -5.2 | 9.5 | 1.6 | -1.5 | 2.6 |
| 55-59 | 362.6 | 366.5 | 362.4 | 372.4 | 3.9 | -0.2 | 9.8 | 1.1 | -0.1 | 2.6 |
| 60-64 | 358.2 | 364.4 | 351.3 | 365.3 | 6.2 | -6.9 | 7.1 | 1.7 | -2.0 | 1.9 |
| 65-69 | 298.2 | 302.2 | 301.9 | 306.7 | 4.0 | 3.7 | 8.5 | 1.3 | 1.2 | 2.8 |
| 70-74 | 259.0 | 262.9 | 262.2 | 269.7 | 3.9 | 3.2 | 10.7 | 1.5 | 1.2 | 4.0 |
| $75+$ | 396.2 | 404.7 | 385.0 | 434.1 | 8.5 | -11.2 | 37.9 | 2.1 | -2.9 | 8.7 |
| Total | 7833.8 | 7964.6 | 7866.2 | 8109.6 | 130.8 | 32.4 | 275.8 | 1.6 | 0.4 | 3.4 |

[^6]

Figure 1. Percentage under-enumeration at the 1986 Census: Post-Enumeration Survey, Demographic Estimates and Medicare Enrolment-MALES.


Figure 2. Percentage under-enumeration at the 1986 Census: Post-Enumeration Survey, Demographic Estimates and Medicare Enrolment-FEMALES.

For females, the level of correspondence between PES results and demographic estimates for ages below 35 is satisfactory. However, the demographic estimates for some age groups are considerably lower than PES estimates, and for those aged 35 to 39 , and 45 to 64 , they are lower than the unadjusted census count. Demographic estimates for these groups appear to be too low. This supports the view that demographic estimates are not sufficiently accurate for the production of population estimates and should be used only to assess PES results.

PES under-enumeration rates by age show a pattern which is smooth and much less erratic than that shown by demographic estimates. The higher PES rates for young adults aged 20-29 compared with those for other ages are as expected, given the higher rates of mobility among young adults, particularly males.

Medicare registrations are considerably higher than PES estimates and demographic estimates, except for the 0-4 age-group. Studies of registration practice show that the lower number in the $0-4$ age group for medicare registration reflects the delays in births being registered with Medicare, and the higher numbers in other ages reflect delays in deleting from the Medicare register deaths and persons who have emigrated from the country.

Comparisons of PES estimates with estimates from family allowance registration and school enrolments for selected age-groups also show satisfactory correspondence. These results give some confirmation of the accuracy of the PES estimates in so far as the younger ages are concerned.

Although there is a satisfactory level of correspondence between PES estimates and other estimates of the population, there are two remaining problems which require consideration before the PES estimates can be accepted. The first emerges from an analysis of the PES estimates of census under-enumeration rates by age and sex. These rates are shown in Table 2.

Except for those aged 0-4 and $75+$, male under-enumeration rates are generally higher than female rates. While the rates for those aged $75+$ could be affected by small sample size, the rate for females aged $0-4$ appears too high, 1.9 per cent compared with 1.3 percent for males of the same age and for females aged 5-9. The number of females aged $0-4$ estimated by the PES to have been under-enumerated was 11,300 compared with about 7,000 for the age group 5-9. This large difference in under-enumeration between those aged $0-4$ and those aged 5-9 for females does not exist for males.

The PES sex ratio for persons aged $0-4$ is 104.3 males to 100 female, lower than the census count ratio of 104.9 and the ratio of 105.0 males to 100 females estimated from demographic data.

On the above evidence, it appears that the PES has over-estimated females aged 0-4, although it is difficult to see how the PES could have over-estimated this group more so than other groups.

The second problem relates to the very high PES under-enumeration rate estimated for the Northern Territory. As shown in Table 4 it is $9.97 \%$ on an actual location basis and $6.45 \%$ on a usual residence basis. Northern Territory is a sparsely populated area (the census count in 1986 was 154,800 in an area of 1.3 million square kilometers) with a highly mobile population. The PES estimate of the population of Northern Territory is considerably higher than that based on the 1981 Census. Comparisons of PES estimates for the Northern Territory with independent estimates such as the number of children on the family allowance register and the number of school enrolments, also show that PES estimates are high. While these independent estimates may very well contain errors, it appears very likely that the PES has overestimated the rate of under-enumeration for the NT.

The PES questionnaires were checked for the Northern Territory and were found to be satisfactory except for one collection district where problems with unreliable addresses and difficult terrain exposed inadequacies in field procedures and led to difficulties with matching.

Table 3
Comparison of 1986 PES Results with Independent Estimates

|  | PES <br> Estimates | Demographic <br> Estimates | Family <br> Allowance | School <br> Enrolment |
| :---: | :---: | :---: | :---: | ---: |
| Persons ('000) |  |  |  |  |
| $0-4$ | 1207.3 | 1196.6 |  | $1204.8(\mathrm{a})$ |
| $5-9$ | 1174.8 | 1168.5 | 1177.0 | - |
| $10-14$ | 1307.2 | 1298.6 | 1304.2 | 1289.6 |

(a) Family allowance registration for age 0 is understated because of the time lag in births being registered for family allowance. An adjustment was made by substituting the family allowance figure for age 0 by an estimate from the demographic analysis.
(b) School enrolment not compulsory for children aged 5 years.

Table 4
PES Under-Enumeration Rates (\%) by State

|  | Actual <br> location <br> basis | Usual <br> Residence <br> basis |
| :--- | :---: | :---: |
| New South Wales | 1.54 | 1.51 |
| Victoria | 1.59 | 1.77 |
| Queensland | 2.68 | 2.43 |
| South Australia | 1.54 | 1.59 |
| Western Australia | 2.32 | 2.26 |
| Tasmania | 1.32 | 1.16 |
| Northern Territory | 9.97 | 6.45 |
| Aust. Capital Territory | 1.95 | 1.61 |
| Australia | 1.91 | 1.84 |

A judgement was made that the PES over-estimation of females aged 0-4 and of the NT population should be corrected by adjusting the PES results. The adjustment to females aged $0-4$ was made by using the sex ratio from demographic estimates and applying this to the PES estimates of males aged 0-4. Essentially, this amounted to replacing the PES estimate of females aged $0-4$ by a better estimate using the PES estimate of males and the sex ratio. The result of this adjustment was to reduce the estimates of this group by 4,000 to 587,000 .

The problem with the NT estimates was handled by not using data from the problematic collection district. This reduced the Northern Territory under-enumeration rate to 9.1 per cent (on an actual location basis) and 5.5 per cent (on a usual residence basis).

The two adjustments to PES results reduced the overall national under-enumeration rate from 1.91 per cent to 1.87 per cent (on an actual location basis), or from 1.84 per cent to 1.81 per cent (on a usual residence basis). Table 5 shows PES estimates by age and sex after the above adjustments were made to the estimates for NT and for females aged 0-4.

Table 5
Census Count 1986 Adjusted for Under-enumeration by Age and Sex

| Age | On the basis of 'actual location' |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Males |  | Females |  | Persons |  |
|  | No. ('000) | $\%$ under enumeration | No. <br> ('000) | \% under enumeration | $\begin{aligned} & \text { No. } \\ & (’ 000) \end{aligned}$ | \% under enumeation |
| 0-4 | 616.3 | 1.30 | 586.6 | 1.17 | 1202.9 | 1.24 |
| 5-9 | 602.4 | 1.24 | 572.4 | 1.27 | 1174.8 | 1.26 |
| 10-14 | 670.1 | 1.39 | 636.8 | 1.38 | 1306.9 | 1.39 |
| 15-19 | 688.3 | 2.19 | 657.3 | 2.02 | 1345.6 | 2.11 |
| 20-24 | 679.4 | 4.54 | 652.4 | 2.95 | 1331.8 | 3.76 |
| 25-29 | 677.5 | 4.17 | 660.7 | 1.81 | 1338.2 | 3.00 |
| 30-34 | 629.9 | 2.29 | 627.8 | 1.55 | 1257.7 | 1.92 |
| 35-39 | 634.0 | 1.87 | 618.9 | 1.11 | 1252.9 | 1.49 |
| 40-44 | 512.6 | 1.64 | 488.5 | 1.21 | 1001.1 | 1.43 |
| 45-49 | 426.9 | 1.66 | 403.0 | 0.98 | 829.9 | 1.33 |
| 50-54 | 371.2 | 2.04 | 354.6 | 1.56 | 725.8 | 1.80 |
| 55-59 | 379.5 | 1.62 | 366.5 | 1.06 | 746.0 | 1.34 |
| 60-64 | 347.0 | 1.70 | 364.4 | 1.70 | 711.4 | 1.70 |
| 65-69 | 263.6 | 1.52 | 302.3 | 1.35 | 565.9 | 1.43 |
| 70-74 | 208.2 | 1.92 | 262.9 | 1.47 | 471.1 | 1.67 |
| $75+$ | 233.0 | 1.49 | 404.7 | 2.08 | 637.7 | 1.86 |
| All ages | 7940.1 | 2.16 | 7959.7 | 1.58 | 15899.8 | 1.87 |


| Age | Males |  | Females |  | Persons |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $\begin{aligned} & \text { No. } \\ & \left({ }^{\prime} 000\right) \end{aligned}$ | $\%$ under enumeration | No. ('000) | $\%$ under enumeration | $\begin{aligned} & \text { No. } \\ & (’ 000) \end{aligned}$ | $\%$ under enumeation |
| 0-4 | 615.3 | 1.29 | 585.9 | 1.22 | 1201.2 | 1.26 |
| 5-9 | 601.3 | 1.23 | 571.2 | 1.22 | 1172.5 | 1.23 |
| 10-14 | 668.5 | 1.29 | 635.7 | 1.36 | 1304.2 | 1.33 |
| 15-19 | 685.6 | 2.11 | 654.3 | 1.97 | 1339.9 | 2.04 |
| 20-24 | 673.1 | 4.33 | 646.9 | 2.83 | 1320.0 | 3.59 |
| 25-29 | 672.6 | 4.02 | 657.2 | 1.80 | 1329.8 | 2.92 |
| 30-34 | 626.6 | 2.21 | 625.6 | 1.53 | 1252.2 | 1.87 |
| 35-39 | 630.9 | 1.78 | 616.7 | 1.05 | 1247.6 | 1.41 |
| 40-44 | 510.3 | 1.59 | 487.0 | 1.19 | 997.3 | 1.39 |
| 45-49 | 424.7 | 1.52 | 401.7 | 0.98 | 826.4 | 1.26 |
| 50-54 | 369.6 | 1.97 | 353.0 | 1.52 | 722.6 | 1.75 |
| 55-59 | 377.7 | 1.52 | 364.0 | 0.92 | 741.8 | 1.22 |
| 60-64 | 345.6 | 1.74 | 361.6 | 1.61 | 707.3 | 1.67 |
| 65-69 | 262.1 | 1.47 | 300.2 | 1.31 | 562.3 | 1.38 |
| 70-74 | 207.2 | 1.89 | 261.3 | 1.46 | 468.5 | 1.65 |
| $75+$ | 232.4 | 1.52 | 403.3 | 2.01 | 635.7 | 1.83 |
| All ages | 7903.6 | 2.08 | 7925.5 | 1.54 | 15829.1 | 1.81 |

## 5. ESTIMATING SUB-NATIONAL POPULATIONS

Internal migration data are not sufficiently reliable for demographic estimates of the population at sub-national levels to be used to assess census under-enumeration. However, a comparison of the 1986 PES estimates of the number of children aged $1-15$ was made with the corresponding number receiving family allowance by State/Territory. This comparison shows a general agreement except for Northern Territory where the percentage difference was more than $2 \%$.

Given this general agreement between PES estimates and family allowance data, and in the absence of reliable independent data on higher ages for comparison with PES estimates, the PES estimates (after adjustments) of the State and Territory populations were accepted.

Population estimates at the State/Territory level by age and sex, and at the local government area level were not derived directly from the PES. The 1986 PES was a sample survey and the results are subject to sampling error. Sampling errors at the State/Territory level by age and sex and at the local government area level are high, many unacceptably high, relative to the amounts of adjustment for under-enumeration which need to be made. An alternative indirect method, using an iterative proportional fitting (IPF) procedure, was used to produce State/Territory estimates by age and sex from those higher level PES estimates with a low sampling error. For a description of the IPF procedure, see Purcell and Kish (1979). This procedure involved taking the national population estimates by age and sex and the State/ Territory estimates within each sex and adjusting the census age by State/Territory counts to these two margins.

The IPF procedures involves the following cycles $n=0,1, \ldots$.

$$
\begin{aligned}
X_{g a s}^{(2 n+1)} & =X_{g a s}^{(2 n)} \frac{X_{a s}}{X_{a s}}(2 n) \\
X_{g a s}^{(2 n+2)} & =X_{g a s}^{(2 n+1)} \frac{X_{g s}}{X_{g s}^{(2 n+1)}}
\end{aligned}
$$

and $X_{g a s}^{(o)}=Y_{g a s}$ the census count for state $g$, age category $a$ and sex $s$. The procedure converges to a unique solution. The use of IPF procedures, of course, assumes that the relationship between the variables within the assocation structure is valid and that this relationship is preserved.

For estimates for local government areas, the problem with high sampling error is more acute and results of the PES are not sufficiently reliable to make direct estimates of underenumeration for each local government area. Based on the premise that under-enumeration is age/sex and birthplace (Australian born/Overseas born) selective, and that it differs between States/Territories and between capital city and the rest of the State, adjustments for underenumeration at the local government area level were made to reflect under-enumeration differentials by age, sex, capital city/rest of State and Australian-born/overseas-born.

## 6. PROBLEMS WITH THE PES ESTIMATION

As pointed out by Bailar (1985), for example, the bias and consistency of the PES estimates is affected by errors in the matching process, any correlation between a person being missed in the census and in the PES, and erroneous inclusions in either the census or the PES. It is
because of the possible effects of these factors that the results of the PES are assessed using demographic and administrative data in the ways described above.

Errors in matching will bias the PES estimates. Failure to match records that in fact should match will lead to the creation of apparently under-enumerated persons and the PES estimate will be an over estimate. The effect of false matches will be the reverse.

Erroneous inclusions in either the census or PES will inflate the values of $Y$ or $x$ and hence the PES estimate. The US Bureau of the Census conducts a special "E-sample"' selected from the census to estimate the extent of erroneous inclusions in the census which can then be incorporated in the estimate by adjusting the census count $Y$. For a description of the $E$ sample, see Fay, Passel and Robinson (1988). The matching and estimation procedures used by the ABS attempt to adjust for some of the effect of erroneous inclusions by determining not only whether or not someone has been included but whether they should have been included and if they have been included more than once. For example in the 1986 PES, 250 people were determined to have been included twice and four persons had been included three times. Cases were also found where persons had been included but should not have been. In this way viewing the PES estimation as a ratio estimator rather than a dual system estimator enables the accounting for some erroneous inclusions.

The dual system estimation method makes the assumption that whether or not someone is missed in the PES is independent of whether or not that person is missed in the census. Whilst all practical steps have been taken in ensuring that the two field and processing systems involved in the collections are completely separate and independent it is still possible for correlation to exist. Positive correlation will mean that the PES estimate based on the assumption of independence will be an under-estimate, negative correlation leads the PES estimate to overestimate. Negative correlation would occur if being included in the census led people to be hard to enumerate in the PES but we have no clear evidence for this; the final response rate for the PES $(95 \%)$ is in line with other household surveys conducted by the ABS. Positive correlation seems more likely, and there appears to have been some evidence of this in the 1981 Census. If such positive correlation exists then the PES based adjustments will have not gone far enough but will have been in the right direction.

## 7. ALTERNATIVE METHODS OF ESTIMATION (WOLTER 1986)

The idea of combining PES data and demographically derived sex ratios or sex ratios obtained from other sources is the basis of methods suggested by Wolter (1986). Wolter suggests several models and associated methods which formally combine sex ratios and PES estimates. These methods are attempts to loosen the assumption of independence inherent in the PES estimation methods.

Wolter considers two models. In the first it is assumed that the degree of association in underenumeration between the PES and the census (as measured by the cross-product ratios in tables such as the diagram shown earlier in this paper) is the same for males and females within each age category. In the second model independence is assumed for females and an externally derived sex ratio is used to obtain the male figure. It is then possible to calculate the crossproduct ratios implied for males.

From an initial evaluation of these methods applied to Australian data, it was found that the first model produced very erratic estimates of the cross product ratios, with approximately $50 \%$ being negative. This was greatly reduced under the second model although some remained negative and were set to zero in a modified model. The problem with negative cross-product

Table 6
Sex Ratios: Males per 100 Females

| Age | Alternative | PES |
| :---: | :---: | ---: |
| $0-4$ | 105.0 | 104.3 |
| $5-9$ | 105.2 | 105.2 |
| $10-14$ | 105.2 | 105.3 |
| $15-19$ | 104.7 | 104.7 |
| $20-24$ | 104.1 | 104.1 |
| $25-29$ | 102.6 | 102.6 |
| $30-34$ | 100.3 | 100.3 |
| $35-39$ | 102.4 | 102.4 |
| $40-44$ | 104.5 | 104.9 |
| $45-49$ | 105.2 | 106.0 |
| $50-54$ | 104.2 | 104.7 |
| $55-59$ | 103.0 | 103.5 |
| $60-64$ | 95.2 | 95.2 |
| $65-69$ | 87.1 | 87.2 |
| $70-74$ | 78.8 | 79.2 |
| $75+$ | 57.9 | 57.6 |

ratios was also identified by Wolter (1986, p. 7). The second model, modified, was then applied to 1986 data. For age groups $5-9$ up to $35-39$, the sex ratio obtained from the PES were in line with expectations and those sex ratios were used giving exactly the PES estimate. For the 0-4 age group the sex ratio obtained from demographic estimates was used and for the 40-44 to $75+$ age groups, an alternative estimate of the sex ratios based on census counts was used. The sex ratios are given in Table 6.

The sex ratio used and the PES sex ratios are not greatly different so applying Wolter's second model leads to only small changes in the PES estimates. For the $0-4$ and $75+$ age groups the estimates of males are increased by $0.7 \%$ and $0.5 \%$ respectively. For the $45-49$ and $70-74$ age groups the estimates are reduced by between $0.7 \%$ and $0.5 \%$. This analysis suggests that the differences in biases between sexes in the PES estimation method due to the combined effect of the potential problems discussed above, are relatively small. It could be the case that any biases are affecting males and females to an approximately equal degree so that PES sex ratios are broadly acceptable.

Our experience in 1981 and 1986 demonstrated the need to use sex ratios in assessing measures of under-enumeration and we believe the Wolter method is a useful way of generating alternative estimates against which the Census count and direct PES estimates can be judged. The general acceptability of the PES sex ratios in 1986 has meant that using this method made little difference. The acceptability of the PES sex ratios in 1986 , except for the $0-4$ age group contrasts with the experience in 1981, where an adjustment to the PES estimates was considered necessary for a number of age groups based on alternative sex ratios. These differences in the 1981 and 1986 experience may reflect a reduction in correlation between under-enumeration in the census and the PES in 1986.

## 8. CONCLUSION

While the ABS has adjusted the past three censuses for under-enumeration, our confidence in the basic reliability of the PES stems from its general consistency with other data sources. No fundamental change in approach is anticipated for the next census to be conducted in 1991. However, we believe there is a need to investigate further potential causes of bias, in particular the adequacy of the clerical matching procedures, and methods to overcome correlation bias. It is also planned to investigate the possibility of creating a demographic data bank on a usual residence basis, so that the effects of the large volume of short-term movements can be eliminated or reduced.

## ACKNOWLEDGEMENT

We are grateful to the reviewers of this article for their comments and suggestions for further research.

## APPENDIX

## THE 1986 POST-ENUMERATION SURVEY

## General

The 1986 PES was conducted in the 4th and 5th weeks after census night. The survey involved interviews with a sample of the population from about 35,000 private dwellings ( $2 / 3$ of one percent of dwellings) across Australia involving about 100,000 persons. The sampling fraction varied between States and Territories, with the smaller States and Territories having higher sampling fractions. Personal data on name, age, sex, marital status and birthplace were obtained by interviewers for matching with information on the census form. For each person in the survey, information was sought on their place of usual residence, where they spent census night, their address before and after census night and any other address where they might have been included on a census form. At each given address, the personal information was matched to census forms to establish whether a person was missed, counted once or the number of times counted if counted more than once.

## Scope and Sample Structure of the PES

Except for the special cases mentioned below, the PES included in its scope all persons who should have been enumerated in the census, except those who had gone overseas or died between the census and PES dates. Diplomatic representatives and persons in diplomatic dwellings were not included in the census. These persons were excluded from the survey as were babies born after census night. Persons in the survey who were overseas on census night were matched to census forms to determine whether they were incorrectly included in the census.

For practical reasons, very sparsely settled areas were not included in the PES. In these areas, special census procedures were used to contact and enumerate Aboriginal groups, people in mining camps, cattle stations, etc. The PES in these areas would need to rely on the same contacts and procedures adopted for the census and therefore could not accurately and independently measure under-enumeration. Consequently, the scope of the PES excluded these areas.

Non-private or special dwellings such as hospitals, hotels, and motels also were not included in the PES. The vast majority of residents in non-private or special dwellings would have been short-term residents and, according to normal ABS survey rules short-term residents would have a chance of being included in the survey at their place of usual residence where information on such persons would be obtained. A relatively small number of long term residents of these dwellings were consequently not included in the PES. For estimation purposes, populations out-of-scope were assumed to have the average capital or non-capital city rate of underenumeration for each State as appropriate and the average Territory rate for each of the two Territories.

As non-private or special dwellings and sparsely settled areas contained less than $3 \%$ of the total population, any differences in under-enumeration of these areas compared with areas covered by the PES would be unlikely to have a significant effect on the overall estimated level of underenumeration at the State or National level.

Interaction Between the Census and the PES
It is important that the PES be conducted as independently of the census as possible. Otherwise, the factors that led to a person being missed or overcounted in the census may also be present in the PES, resulting in biased estimation of the under-enumeration. Furthermore, knowledge of the areas to be included in the PES might influence the performance of census collectors in these areas so that the PES sample would not be a representative sample of the under-enumeration. For these reasons the field and office staff used in the census and PES were totally separate. PES interviewers were not employed as census collectors or census group leaders, and census field staff were not told which areas were included in the PES.

Independence was further guaranteed in two ways - by ensuring the operational independence of the field systems, and by adopting special procedures for census forms received by mail after the PES field work commenced.

To ensure operational independence, PES field work commenced after all available census forms had been collected from the field. Thus census collectors were not in the field at the same time as PES interviewers and there was no possibility of interaction, even unintentional, between census and PES field staff.

Special procedures for census forms received after the PES commenced were required to overcome the effects PES fieldwork may have had on householders who were late returning their census forms. In some cases, PES interviewers discovered census forms still uncollected. This situation was possible because some people had preferred to post in their census forms and had not yet done so, or the census collector had been unable to make contact to collect them. Some of these people who were included in the PES may have been prompted to post their forms in, where they would not otherwise have done so. To overcome this potential bias, any census form returned by mail after Monday 20 July 1986 (the day PES interviewing commenced) was considered a late form. Special procedures for the treatment of late forms are described later in this Appendix.

## Matching procedures of the PES

Matching for the purpose of determining whether a person was missed, counted once or the number of times counted if counted more than once, was conducted in two stages. Both these stages were clerical processes undertaken by staff at the census Data Transcription Centre.

The first stage was the locating of census forms for the addresses of households selected in the PES. Processing of 1986 Census forms were centralized in Sydney. Staff at the Population Census Data Transcription Centre were requested to compare the address on the front of the PES interview form with all addresses given in the record book of the census collector
who was responsible for the collection district (CD) in which the PES household was located. The record book was used as a control in the delivery and collection of census forms, and contained information such as name, address and number of persons for all households in the CD.

To assist identification of households where addresses were sometimes vague, for example in rural areas, processing staff were asked to also use names of the householders, property names etc. In addition, staff were instructed to check through all addresses in the record book so that any duplicate census forms were identified. Addresses in record books of adjacent CDs were also checked if the address of the household selected by the PES was near the boundary of the $C D$.

The second stage was person-matching and this was based on the name and demographic details of the persons listed on the census and PES forms. In this matching process, a search form was generated for each address reported in the PES for any person in the household, other than the address of the PES selected dwelling. A search form was treated the same way as a PES interview form and an attempt was made to locate the census form which corresponded to the search form address.

In most cases, the person-matching procedure was straight forward. There were, however, cases of spelling errors and insufficient details on addresses to identify a clear match on name. In these cases, a judgement on whether or not a person was counted was made based on other information such as age, sex, marital status, birthplace and relationship to other members in the census household. For doubtful cases, processing staff were required to consult their supervisor.

The PES also asked the respondent whether each person was included on a census form. When matching failed because of lack of adequate information, the respondent's statement about whether or not the person was counted was accepted. There were a few cases where even this information was unavailable. These cases were considered not counted in the census.

After matching, the data was entered onto computer tapes, edited and reformatted to produce a clean unit record file giving the number of times person in the PES sample were counted in the census.
Treatment of Late Census Forms and 'Dummy' Census Forms
In forming the estimation equation:
$X=Y(x / y)$, where
$X=$ estimated census count adjusted for underenumeration
$Y=$ raw census count, unadjusted
$x=$ PES estimate of the number of persons who should have been included in the census and
$y=$ PES estimate of the number of persons who were included in the census,
two categories of census forms were treated as missed in the census. These are 'dummy' census forms and late census forms.

Dummy census forms were created during census fieldwork for dwellings at which households were known to be residing, did not return their census forms and could not be contacted. Census collectors were instructed to exercise extreme care in creating these dummy forms and they needed to be satisfied that there was concrete evidence that the dwellings were occupied on census night. The collectors were instructed also to obtain as much information as possible regarding the number and the demographic characteristics of these residents.

When a PES address was matched to a dummy census form, the lack of name and reliable personal characteristics on the census form made it impossible to perform the matching operation satisfactorily.

It is also necessary to handle late census forms differently from normal census forms. Because late census forms might have been prompted by a PES interviewer calling, their inclusion could lead to a bias in the estimation of under-enumeration.

In the 1986 Census, there were 115,000 persons recorded on dummy census forms or late census returns, or 0.7 per cent of the population. Both dummy and late census forms were excluded from the raw census count ( Y ) and the PES estimate of the number of persons who were counted in the census $(y)$, but were included in the PES estimate of the number of persons who should have been counted in the census $(x)$. In other words, persons on dummy and late forms were treated as missed and adjusted for by ( $x$ ). The adjustment factor $(x / y)$ is exaggerated because of the exclusion of dummy and late forms from ( $y$ ), but this exaggeration is compensated for by the exclusion of these forms from the raw census count ( $Y$ ).

## Estimation Procedure

The estimation procedure was applied at the age by sex by geographic area (capital city statistical division/rest of state) level. Adjustment factors were included in the estimation formulae to partly account for non-responding and non-contact households. These factors adjust both of the main estimates, $x$ and $y$, by effectively imputing, for each non-contact or refusing household, the average number of persons per household, and, for each person so imputed, the average rate of under-enumeration at the relevant age by sex by area level. To reduce the bias from the use of such adjustment factors, the factors were calculated for various subgroups of households by the status of enumeration at the census (such as occupied dwelling, late returned form). This enumeration status was considered to be related to what non-response was encountered in the PES.

## REFERENCES

BAILAR, B.A. (1985). Comments on "Estimating the population in a Census Year: 1980 and beyond" by E.P. Ericksen and J.B. Kadane, Journal of the American Statistical Association, 80, 109-114.
BISHOP, Y.M.M., FIENBERG, S.A., and HOLLAND, P.W. (1975). Discrete Multivariate Analysis: Theory and Practice. Cambridge: MIT Press.

FAY, R., PASSEL, J.S., and ROBINSON, J.G. (1988). The coverage of population in the 1980 Census. Evaluation and Research Report, PHC 80-E4, United States Bureau of the Census, Washington D.C.
PURCELL, N.J., and KISH, L. (1979). Estimation for small domains. Biometrics, 35, 365-384.
WOLTER, K.M. (1986). Capture-recapture estimation in the presence of known sex ratio. SRD Research Report, United States Bureau of the Census, Washington, D.C.

# When Are Census Counts Improved by Adjustment? 

NOEL CRESSIE ${ }^{1}$


#### Abstract

There are persuasive arguments for and against adjustment of the U.S. decennial census counts, although many of them are based on political rather than technical considerations. The decision whether or not to adjust depends crucially on the method of adjustment. Moreover, should adjustment take place using say a synthetic-based or a regression-based method, at which level should this occur and how should aggregation and disaggregation proceed? In order to answer these questions sensibly, a model of undercount errors is needed which is "level-consistent" in the sense that it is preserved for areas at the national, state, county, etc. level. Such a model is proposed in this article; like subareas are identified with strata such that within a stratum the subareas' adjustment factors have a common stratum mean and have variances inversely proportional to their census counts. By taking into account sampling of the areas (e.g., by dual-system estimation), empirical Bayes estimators that combine information from the stratum average and the sample value, can be constructed. These estimators are evaluated at the state level ( 51 states, including Washington, D.C.), and stratified on race/ethnicity ( 3 strata) using data from the 1980 postenumeration survey (PEP 3-8, for the noninstitutional population).


KEY WORDS: Emprical Bayes estimation; Loss functions; Measures of improvement; Quantile function; Spatial correlation; Synthetic estimation.

## 1. INTRODUCTION

This article is of a technical nature, but it is important to present a brief explanation of the political and social ramifications of the "undercount issue" in the United States of America. By December 31 of the year of the decennial census, the U.S. Census Bureau is specified by law to submit state population counts to Congress for the purpose of reapportionment of the House of Representatives, and by March 31, 1991, to submit small-area population counts for the purpose of redistricting. In recent decades, the number of uses to which census data are put have multiplied: revenue-sharing formulas use population and per capita income for each incorporated place, demographic and sociological research at regional, state, and national levels usually rely on census counts, etc.

Inaccurate census counts should be cause for concern to the whole nation. That certain groups of people (young black males, illegal aliens, etc.) are harder to count than others, is without question; see Ericksen and Kadane (1985), and Freedman and Navidi (1986), and the discussion following these articles. If the hard-to-count groups were distributed in equal proportions throughout the political and administrative regions of the USA there would be far less controversy over what to do about the uncounted people. As it is, many of the large American cities such as Chicago, Detroit, New York, and Los Angeles feel they are losing federal funds because their cities contain more of the types of people that tend to remain uncounted. And certain states such as New York and California feel they are under-represented in Congress, to the benefit of Midwestern states such as Indiana and Iowa.

[^7]Census undercount is defined simply as the difference between the true count and the census count, expressed as a percentage of the true count. My approach to its estimation is modelbased, relying on data obtained from the post-enumeration survey (PES). A number of technical aspects of a model-based approach to adjustment will be addressed in this article. Section 2 establishes the model, addresses the question of choice of measures of improvement, and presents results for aggregation and disaggregation based on Bayes and Synthetic estimators. Section 3 gives empirical Bayes versions of the results of Section 2. Section 4 summarizes what has been learned from this model-based approach; there is also discussion of the implications of the sufficient conditions that guarantee risks of adjusted counts to be smaller than risks of census counts.

## 2. THE MIXTURE MODEL AND ITS CONSEQUENCES

At the outset I would like to explain the source of random variation in my model, originally defined in Cressie (1986), and further developed in Cressie (1988). I consider the true population in any well-defined stratum of the USA, to be unknown. After observing the corresponding census population, the uncertainties about the true population are updated. In other words, all inference will be performed conditionally on the observed census counts.

### 2.1 The Model

The method of synthetic estimation constructs estimators of undercount at a particular level ( e.g., the state level) by summing undercounts of various strata (e.g., demographic strata) over the area being considered (e.g., California), where it is assumed that any stratum has a constant proportion of true counts to census count regardless of which area is being considered. For example, it would be assumed that the proportion for young black males is the same for California, Delaware and so on. Most often these strata are defined demographically according to the factors of age, race, and sex. However Tukey (1981) suggested that geographic and urban factors should be added. Two such stratifications of the USA are given in Isaki et al. (1986).

The mixture model I am proposing assumes a stratification has been defined already, although in Section 4 there is a suggestion how one might determine post hoc whether a chosen stratification is satisfactory.

Suppose there are $j=1, \ldots, J$ strata, and $i=1, \ldots, I$ areas (e.g., at the enumerationdistrict level, $I \simeq 300,000$, while at the state level, $I=51$, including the District of Columbia; for demographic stratification, $J=30$ say, while for the two stratifications in Isaki et al., 1986, $J=90$ and $J=96$. Think of stratum $j$ as fixed (for example, stratum $j$ might be the blacks in central cities in those SMSA's whose population's greater than or equal to 250,000 , in the New England Census Division). Then as $i$ ranges from 1, $\ldots, I$, a sequence of subareas is generated; the subarea indexed by " $j i$ " refers to that part of the $i$-th area that has stratum $j$ in it. Only subareas with nonzero census counts are considered.

Define

$$
\begin{align*}
Y_{j i} & \equiv \text { true count in the } j \text {-th stratum of area } i  \tag{2.1}\\
C_{j i} & \equiv \text { census counts in the } j \text {-th stratum of area } i  \tag{2.2}\\
F_{j i} & \equiv Y_{j i} / C_{j i} ; i=1, \ldots, I ; j=1, \ldots, J \tag{2.3}
\end{align*}
$$

Suppose for the moment that we know the ratios $\left\{F_{j i}: j=1, \ldots, J\right\}$ for the $i$-th area. Then from the census counts $C_{j i}$, the true count $Y_{i}$ can be calculated.

$$
\begin{equation*}
Y_{i}=\sum_{j=1}^{J} F_{j i} C_{j i} \tag{2.4}
\end{equation*}
$$

The $F_{j i}$ are often called adjustment factors. The strata are constructed so that these adjustment factors $\left\{F_{j i}: i=1, \ldots, I\right\}$ are as homogeneous as possible within the $j$-th stratum; $j=1, \ldots, J$ (Tukey 1981).

Realistically the adjustment factors are never known; synthetic estimators exploit the homogeneity and replace (2.4) with

$$
\begin{equation*}
Y_{i}^{\text {sya }}=\sum_{j=1}^{J} F_{j} C_{j i} . \tag{2.5}
\end{equation*}
$$

Now there are only $J$ synthetic adjusment factors $\left[F_{j}: j=1, \ldots, J\right.$ ) to estimate, which through (2.5) yields an estimate of $Y_{i}$. Synthetic estimators have the advantage that the adjustment factors are independent of $i$ and so can be applied to any level of aggregation.

The (estimated) adjustment factors could also be modeled by regression on independent variables that may or may not be census variables; for example, percent minority, crime rate, and percent conventionally counted in the census. Consider,

$$
\begin{equation*}
Y_{i}^{\mathrm{reg}}=\sum_{j=1}^{J}\left(\sum_{k=1}^{p} \beta_{k, j} z_{k, j i}\right) C_{j i} \tag{2.6}
\end{equation*}
$$

To fit the parameters $\beta_{1, j}, \ldots, \beta_{p, j}$ efficiently, various assumptions are made about the error components $\left\{F_{j i}-\sum_{k=1}^{p} \beta_{k, j} z_{k, j i}\right\}$, viz. independent and identically distributed with mean zero.

Ericksen and Kadane (1985) propose the fitting of a regression relation to $\sum_{j=1}^{J} F_{j i} C_{j i} /$ $\sum_{j=1}^{J} C_{j i} ; i=1, \ldots, I$. Freedman and Navidi (1986) criticize the approach and point out the consequences of failure of any of the error assumptions. A problem they did not perceive which I emphasize in (2.7) below, is the heteroskedasticity forced onto the problem by working with ratios; Section 2.2 justifies this model choice. Furthermore, in this latter regression approach undercounts across strata are combined, so that variation between strata is shared by both the regression relation and the error variance. More precise estimators can be obtained through (2.6) by allowing each stratum its own regression relation. Homoskedastic errors and a regression model based on the combination of heterogeneous strata, are also assumed by Ericksen and Kadane (1987) and Ericksen, Kadane and Tukey (1987). It seems that the combination of heterogeneous strata was made necessary by the lack of suitable data.

I do not assume $F_{j i}$ 's that depend only on $j$, nor a regression relation for the $F_{j i}$ 's, but instead reformulate the synthetic assumption $F_{j i} \equiv F_{j}$, into a (statistical) homogeneity assumption:

$$
\begin{equation*}
F_{j i} \sim N\left(F_{j}, \tau_{j}^{2} / C_{j i}\right) ; i=1, \ldots, I ; j=1, \ldots, J, \tag{2.7}
\end{equation*}
$$

where " $\sim$ " means "is distributed as," and $N\left(\mu, \sigma^{2}\right)$ is a normal distribution with mean $\mu$ and variance $\sigma^{2}$. Using a regression relation for the mean has the potential of explaining more of the variation of the $F_{j i}$ 's at the risk of introducing bias through misspecification. The strata chosen in Section 3 are based on race; it was decided not to cloud this sensitive issue with selection of controversial regression variables. I shall refer to the model (2.7) as a mixing
distribution. The normality assumption is made for convenience and will be relaxed later. Here $F_{j}$ is a fixed but unknown mean to be estimated, and $\tau_{j}^{2}=\operatorname{var}\left(\sqrt{C_{j i}} F_{j i}\right)$ is a parameter I shall call the (standardized) stratum variance. As a representation of reality, model (2.7) is better at higher levels of aggregation; see Section 3. All distributions in (2.7) are assumed independent.

There are good reasons for weighting the variance by $1 / C_{j i}$ (see Cressie 1987a, Appendix and 1988). The most attractive consequence of model (2.7), is that it is level-consistent; that is, it is preserved through different levels of aggregation. Specifically,

$$
\begin{equation*}
F_{j, i \& i i^{\prime}} \sim N\left(F_{j}, \frac{\tau_{j}^{2}}{C_{j, i \& i i^{\prime}}}\right) \tag{2.8}
\end{equation*}
$$

where

$$
\begin{equation*}
F_{j, i \& i^{\prime}} \equiv \frac{F_{j i} C_{j i}+F_{j i^{\prime}} C_{j i^{\prime}}}{C_{j, i \& i^{\prime}}}, \text { and } C_{j, i \& i^{\prime}} \equiv C_{j i}+C_{j i^{\prime}} \tag{2.9}
\end{equation*}
$$

This is a very important property that most of the currently proposed statistical models of undercount do not possess. It enables the modeler to escape from the geographical and historical accidents that divided up the country into the states, counties, etc., that we now see.

Of course the $\left\{F_{j i}: i=1, \ldots, I ; j=1, \ldots, J\right\}$ are not available as data; if they were, $\left\{Y_{i}: i=1, \ldots, I\right\}$ would be trivial to calculate. In reality, some sampling takes place so that $F_{j i}$ is observed imperfectly. The best way to think of it is that within stratum $j$ of the $i$-th area, a sample is taken for undercount. Let the outcome be $X_{j i}$ (e.g., $X_{j i}$ is the ratio of dual-system estimator to census count, for the $j$-th stratum in the $i$-th area), and model

$$
\begin{equation*}
X_{j i} \sim N\left(F_{j i}, \sigma_{j}^{2} / C_{j i}\right) ; i=1, \ldots, I ; j=1, \ldots, J \tag{2.10}
\end{equation*}
$$

where $F_{j i}$ is an unknown mean parameter to be estimated, and $\sigma_{j}^{2}=\operatorname{var}\left(\sqrt{C_{j i}} X_{j i}\right)$ is a parameter I shall call the (standardized) sampling variance. All distributions in (2.10) are assumed independent. When the number of strata is large, a large PES (say, 300,000 households) is needed to obtain data for each area-stratum combination.

Probability-proportional-to-size sampling was used by the U.S. Census Bureau in its 1980 post-enumeration program, which implies a sampling variance of the form given in (2.10). As a consequence of this weighting, (2.10) is also level-consistent.

### 2.2 Loss Functions (Measures of Improvement) and their Bayes Estimators

The term loss function is used in statistical decision theory (see, for example, Ferguson 1967) to quantify the loss incurred from using $\hat{\theta}$ as a parameter estimator when the true value is $\theta$. For example, a squared-error loss function is $(\hat{\theta}-\theta)^{2}$. Adopting a more optimistic terminology, the Census Bureau decided in 1986 to use "measure of improvement" instead of "loss function."

Think of (2.10) as a conditional distribution of $X_{j i}$ given $F_{j i}$, and (2.7) as the mixing (or "prior") distribution of $F_{j i}$. To predict $F_{j i}$ then, the "posterior" distribution of $F_{j i}$ given $X_{j i}$ is needed. Notice that a Bayesian terminology is being used since I am thinking of the $F_{j i}$ as random variables whose collection is modeled according to (2.7). But as well as these random parameters, there are fixed but unknown parameters $\left\{F_{j}\right\},\left\{\tau_{j}^{2}\right\},\left\{\sigma_{j}^{2}\right\}$ to be estimated. The posterior of $F_{j i} \mid X_{j i}$ is,

$$
\begin{equation*}
\frac{\text { (distribution of } \left.X_{j i} \mid F_{j i}\right) \cdot\left(\text { '"prior'' of } F_{j i}\right. \text { ) }}{\text { marginal of } X_{j i}} . \tag{2.11}
\end{equation*}
$$

For squared-error loss, the usual Bayes estimator of $F_{j i}$ is simply the expectation of $F_{j i}$ with respect to the posterior: $F_{j i}^{\mathrm{uba}}=E\left(F_{j i} \mid X_{j i}\right)$. Substituting the model (2.7), (2.10) into (2.11), the posterior distribution is easily obtained (see, for example, Lindley and Smith 1972):

$$
\begin{equation*}
F_{j i} \left\lvert\, X_{j i} \sim N\left(F_{j}+\frac{\tau_{j}^{2}}{\tau_{j}^{2}+\sigma_{j}^{2}}\left(X_{j i}-F_{j}\right), \frac{\sigma_{j}^{2} \tau_{j}^{2}}{\tau_{j}^{2}+\sigma_{j}^{2}} / C_{j i}\right)\right., \tag{2.12}
\end{equation*}
$$

for $i=1, \ldots, I ; j=1, \ldots, J$. Hence the posterior expectation is simply

$$
\begin{equation*}
F_{j i}^{\text {uba }}=F_{j}+D_{j}\left(X_{j i}-F_{j}\right), \tag{2.13}
\end{equation*}
$$

where $D_{j} \equiv \tau_{j}^{2} /\left(\tau_{j}^{2}+\sigma_{j}^{2}\right)$. To convert (2.13) into an empirical Bayes estimator, estimators have to be found for $F_{j}$ and $D_{j}$; see Section 3.1.

Although the normality assumptions in (2.7) and (2.10) were used to derive (2.13), more generally (2.13) can be shown to be Bayes for squared-error loss, when assuming simply the mean and variance structure of (2.7) and (2.10), and $E\left(F_{j i} \mid X_{j i}\right)=a_{j i}+b_{j i} X_{j i}$. Goldstein (1975) has an even more general result of which this is a special case. For ease of exposition I shall continue to assume normality but it should be remembered that there is a nonparametric optimality for all the estimators considered.

The estimator $F_{j i}^{\text {bba }}$ given by (2.13) is Bayes for squared-error loss, within the $j$-th stratum of the $i$-th area. Define the estimator of $Y_{i}$,

$$
\begin{equation*}
Y_{i}^{\mathrm{uba}} \equiv \sum_{j=1}^{J} F_{j i}^{\mathrm{uba}} C_{j i} ; i=1, \ldots, I \tag{2.14}
\end{equation*}
$$

and consider the following general loss function:

$$
\begin{equation*}
\sum_{i=1}^{I}\left(Y_{i}^{\text {est }}-Y_{i}\right)^{2} f\left(C_{i}\right) \tag{2.15}
\end{equation*}
$$

where $f\left(C_{i}\right)$ is any nonnegative function of the $i$-th area's census count. Minimizing (2.15) over all $Y_{i}^{\text {est }} \equiv \sum_{j=1}^{J} F_{j i}^{\text {est }} C_{j i}$ leads to choosing $F_{j i}^{\text {est }}$ 's such that $E\left[\sum_{i=1}^{l} \sum_{j=1}^{J} \lambda_{j i}^{\text {est }}\right.$ $\left.\left(F_{j i}^{\text {est }}-F_{j i}\right)^{2} \mid\left\{X_{j i}: i=1, \ldots, I ; j=1, \ldots, J\right\}\right]$ is minimized, where the $\lambda_{j i} \geq 0$ only depend on census counts $\left\{C_{j i}: i=1, \ldots, I ; j=1, \ldots, J\right\}$. This minimum is achieved by the estimator (2.14), which shows it to possess a certain robustness since it is optimal regardless of which $f(\cdot)$ is chosen.

In accordance with recommendation 7.2 in National Academy of Sciences (1985), choice of $f\left(C_{i}\right)=1 / C_{i}$ yields an area's contribution to the total loss that reflects the size of its population. Among the loss functions the Census Bureau has been using, the one most like (2.15) with $f\left(C_{i}\right)=1 / C_{i}$, is

$$
\begin{equation*}
\sum_{i=1}^{l}\left(Y_{i}^{\text {est }}-Y_{i}\right)^{2} / Y_{i} \tag{2.16}
\end{equation*}
$$

it is "most like" in the sense that it is also a weighted sum of squares where each summand yields an area's contribution to the total loss that reflects the size of its population. Here, undercount in more populous areas receive more weight, so that using such loss functions reflects an emphasis on national considerations. The loss function $\sum_{i=1}^{l}\left(Y_{i}^{\text {est }}-Y_{i}\right)^{2} / Y_{i}^{2}$, which guarantees undercount equity for the $I$ areas, will not be considered in this article.

It is easy to show that the Bayes estimator in the case of loss function (2.16) is given by,

$$
\begin{equation*}
Y_{i}^{\mathrm{est}}=\left[E\left(\left(\sum_{j=1}^{J} F_{j i} C_{j i}\right)^{-1} \mid\left\{X_{j i}: i=1, \ldots, I ; j=1, \ldots, J\right\}\right)\right]^{-1} \tag{2.17}
\end{equation*}
$$

which is not a linear combination of $\left\{F_{j i}^{\mathrm{uba}}: j=1, \ldots, j\right\}$. However to a first approximation, using the $\delta$-method, it can be shown that this $Y_{i}^{\text {est }} \simeq Y_{i}^{\text {uba }}$. This is in fact true for a much larger class of loss functions suggested by Cressie (1987b):

$$
\begin{equation*}
L^{\lambda} \equiv \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^{I}\left\{Y_{i}^{\text {est }}\left[\left(\frac{Y_{i}^{\text {est }}}{Y_{i}}\right)^{\lambda}-1\right]+\lambda\left[Y_{i}-Y_{i}^{\text {est }}\right]\right\} ; \lambda \neq 0,-1 \tag{2.18}
\end{equation*}
$$

the cases $\lambda=0,-1$ are defined as the respective limits of $L^{\lambda}$ as $\lambda \rightarrow 0,-1$. Read and Cressie (1988, Chapter 8) show that in this case the Bayes estimator is

$$
\begin{equation*}
Y_{i}^{\operatorname{est}(\lambda)}=\left[E\left(\left(\sum_{j=1}^{J} F_{j i} C_{j i}\right)^{-\lambda} \mid\left[X_{j i}: i=1, \ldots, I ; j=1, \ldots, J\right]\right)\right]-1 / \lambda \tag{2.19}
\end{equation*}
$$

which reduces to (2.14) when $\lambda=-1$, and to (2.17) when $\lambda=1$.
The curious fact is that most undercount estimators used are optimal (under various model assumptions) for $\lambda=-1$, but their performance is measured using $\lambda=1$; i.e., (2.16). The $\delta$-method argument gives $Y_{i}^{\text {est }(\lambda)} \simeq Y_{i}^{\text {uba }}$, and recall $Y_{i}^{\text {uba }}$ is optimal for (2.15); therefore squared-error loss estimators of undercount perform well according to a large class of loss functions. This was observed by Kadane (1984) in his heirarchical Bayesian analysis of 1980 census undercount data ( $\lambda=-1$ and $\lambda=-2$ were compared), and confirmed on the studies of artificial populations carried out by Cressie and Dajani (1988).

It has just been demonstrated that the estimators (2.13) and (2.14) are Bayes (or approximately so) for a large class of loss functions. However it is not likely that the ensemble properties of $\left(F_{j i}^{\text {uba }}: i=1, \ldots, I ; j=1, \ldots, j\right)$, estimate the corresponding ensemble properties of $\left\{F_{j i}: i=1, \ldots, I ; j=1, \ldots, J\right\}$, very well. This follows from the inequality $\operatorname{var}(\theta) \geq$ $\operatorname{var}(E(\theta \mid X))$; in other words the posterior mean of the parameter has a smaller variance than the parameter itself. For estimation of state population totals, this does not matter, but for estimation of the distribution of say $\left\{F_{j i} C_{j i}: i=1, \ldots, 51\right\} ; j=1, \ldots, J$, or $\left\{Y_{i}: i=1\right.$, $\ldots, 51\},(2.13)$ is ill-suited to the task. Such a distribution is needed in standards research (Mulry-Liggan and Hogan 1986) to determine the proportion of people in a stratum affected by an undercount more severe than $u \%$ (Cressie 1988, Section 4).

I shall constrain the estimator of $\left\{F_{j i}: i=1, \ldots, I\right.$ \} so that the posterior moments of its (weighted) empirical distribution function match the moments of the estimator's weighted empirical distribution function. This is achieved by modifying the usual Bayes estimator, yielding a constrained Bayes estimator with the right ensemble properties. Louis (1984) presents the details for an equal-variance version of the model (2.7), (2.10), but a straightforward modification of his approach is possible for weighted variances. Cressie (1986) shows that such a constrained Bayes estimator is

$$
\begin{gather*}
F_{j i}^{\mathrm{cba}}=\zeta_{j}+G_{j}\left(X_{j i}-\zeta_{j}\right)  \tag{2.20}\\
Y_{i}^{\mathrm{cba}}=\sum_{j=1}^{J} F_{j i}^{\mathrm{cba}} C_{j i} \tag{2.21}
\end{gather*}
$$

obtained by solving for $\zeta_{j}$ and $G_{j}$ in:

$$
\begin{gather*}
\zeta_{j}+G_{j}\left(X_{j} .-\zeta_{j}\right)=F_{j}+D_{j}\left(X_{j} .-F_{j}\right) \\
G_{j}^{2} \sum_{i}\left(C_{j i} / \sum_{h} C_{j h}\right)\left(X_{j i}-X_{j} .\right)^{2}= \\
(I-1) D_{j} \sigma_{j}^{2} / \sum_{h} C_{j h}+D_{j}^{2} \sum_{i}\left(C_{j i} / \sum_{h} C_{j h}\right)\left(X_{j i}-X_{j} .\right)^{2}, \tag{2.22}
\end{gather*}
$$

where

$$
\begin{equation*}
X_{j} .=\sum_{i=1}^{I} X_{j i} C_{j i} / \sum_{h=1}^{I} C_{j h} \tag{2.23}
\end{equation*}
$$

### 2.3 Risks of Adjustment; Model Parameters Assumed Known

The model-based approach described in the previous section specifies undercounts in various area-strata combinations, to be random variables. When it comes to comparing the value of one adjustment procedure against another, the expected loss (or the risk) is used. Statistical procedures with small risk are preferred.

In the absence of other considerations (e.g., political, practical, etc.), implementing the procedure with the smallest risk is the correct, impartial approach. The statistician knows that adherence to this modus operandi will yield better estimates on the average, where the average is taken over all problems considered by the statistician. However there is nothing to guarantee that for the particular problem being considered, here estimation of undercount in the 1990 census, a set of area-strata estimates derived from the criterion of minimum risk will actually have smaller loss than another set of estimates. To put it more succinctly, the inequality $E\left(V^{2}\right)<E\left(W^{2}\right)$ does not guarantee that $V^{2}<W^{2}$ for a particular realization. If, in the light of the data collected, a minimum risk prediction did not prove to be the most accurate, the statistical procedure should still be seen as optimal.

In the rest of this section, various results about Bayes estimators will be stated (proofs are given in Cressie 1988). Needless to say, these results rely on the correctness of the assumed model. In practice, the more relevant results are for empirical Bayes estimators, which are given (with proofs) in Section 3.

The first thing to recall (from Section 2.2) about the usual Bayes estimators (2.13), (2.14) is that they are optimal or near optimal for a large class of loss functions. Moreover the estimators are level-free; i.e., they are not only optimal at the level at which they are constructed, but after aggregation they are also optimal at the higher level. From (2.14),

$$
\begin{equation*}
Y_{i}^{\mathrm{uba}}+Y_{i^{\prime}}^{\mathrm{uba}}=Y_{i \& i^{\prime}}^{\mathrm{uba}}, \tag{2.24}
\end{equation*}
$$

where $i \& i^{\prime}$ denotes the area obtained by combining the two disjoint areas $i$ and $i^{\prime}$.
Therefore, one should aim to construct a Bayes estimator at the very lowest level (census blocks) and aggregate up to whatever level is desired, thus ensuring consistency of counts at all levels. In practice this is out of the question, simply because the post-enumeration survey would never be large enough to give dual-system estimated undercount data for all the blocks. The same is true at the enumeration-district level and the county level. Moreover, at these lower levels the model (2.7) and (2.10) does not fit as well (Cressie and Dajani 1988); an adequate fit at the state level is shown in Section 3.1.

It is certain that the post-enumeration survey will gather data from each of the 51 states, allowing construction of (empirical) Bayes estimators at the state level. Politically, the state level is the most sensitive; reapportionment of the 50 states' representation (Washington, D.C. is excluded) in the House is the first use made of decennial census counts (mandated to reach Congress by December 31 in the year of the census). Thus at this level, the Bayes estimators (2.13) and (2.14) offer a compromise between a state's observed adjustment factors \{ $X_{j i}$ : $j=1, \ldots, J\}$; and the (synthetic) adjustment factors $\left\{F_{j}: j=1, \ldots, J\right\}$. For example, Mississippi's black undercount is recognized as being potentially different from New York's black undercount, when using the Bayes estimators.

I shall now explore the consequences of synthetic estimation at lower levels, after Bayes estimation is carried out at a given level. For consistency of counts at all levels, it is desired to estimate undercount at the block level and aggregate up to whatever level is desired. Suppose an adjustment factor $F_{j i}^{\text {est }}$ is estimated for the $j$-th stratum in the $i$-th area. Now suppose $i=i_{1} \& i_{2} ; i . e$, the $i$-th area is split up into two disjoint subareas $i_{1}$ and $i_{2}$. Then the synthetic method at the lower level posits,

$$
\begin{equation*}
F_{i_{1}}^{\text {sye }}=F_{j_{2}}^{\text {sye }}=F_{j i}^{\text {est }}, \tag{2.25}
\end{equation*}
$$

so that estimators of the true population are given by,

$$
\begin{equation*}
Y_{i_{1}}^{\text {sye }}=\sum_{j=1}^{J} F_{j i_{1}}^{\text {sye }} C_{j i_{1}} ; Y_{i_{2}}^{\text {sye }}=\sum_{j=1}^{J} F_{j i_{2}}^{\text {sye }} C_{j i_{2}} . \tag{2.26}
\end{equation*}
$$

Notice that from (2.25) and (2.26).

$$
\begin{equation*}
Y_{i_{1}}^{\text {sye }}+Y_{i_{2}}^{\text {sye }}=Y_{i}^{\text {est }} \equiv \sum_{j=1}^{J} F_{j i}^{\text {est }} C_{j i} \tag{2.27}
\end{equation*}
$$

which is the desired disaggregation-aggregation property.
Compare the risk of using $Y_{i}^{\text {uba }}, Y_{i}^{\text {sya }}$, and $Y_{i}^{\text {cba }}$ (given by (2.14), (2.5), and (2.21) respectively) to the risk of using $C_{i}$, the census count of the $i$-th area. Using the loss function (2.15), the risks are:

$$
\begin{align*}
\text { uba-risk }_{i} & \equiv E\left[\left(Y_{i}^{\mathrm{uba}}-Y_{i}\right)^{2} f\left(C_{i}\right)\right]  \tag{2.28}\\
\text { cen-risk }_{i} & \equiv E\left[\left(C_{i}-Y_{i}\right)^{2} f\left(C_{i}\right)\right]  \tag{2.29}\\
\text { sya-risk }_{i} & \equiv E\left[\left(Y_{i}^{\text {sya }}-Y_{i}\right)^{2} f\left(C_{i}\right)\right]  \tag{2.30}\\
\text { cba-risk }_{i} & \equiv E\left[\left(Y_{i}^{\text {cba }}-Y_{i}\right)^{2} f\left(C_{i}\right)\right] \tag{2.31}
\end{align*}
$$

The following sequence of inequalities can be proved (Cressie 1988):

$$
\begin{equation*}
\text { uba-risk }_{i} \leq \text { cba-risk }_{i} \leq \text { sya-risk }_{i} \leq \text { cen-risk }_{i}, \tag{2.32}
\end{equation*}
$$

where the middle inequality requires $\sigma_{j}^{2} / \tau_{j}^{2} \leq 3 ; j=1, \ldots, J$.
Now compare the risk of using $Y_{i_{1}}^{\text {sye }}$ and $Y_{i_{2}}^{\text {sye }}$ (estimators of $Y_{i_{1}}$ and $Y_{i_{2}}$ respectively) based on $F_{j i}^{\mathrm{uba}}$ in (2.25), with the risk of using $C_{i_{1}}$ and $C_{i_{2}}$, where area $i=i_{1} \& i_{2}$, the union of disjoint areas $i_{1}$ and $i_{2}$. It can be shown (Cressie 1988) that the synthetic estimation based on the usual Bayes estimator defined at a particular level but applied at a lower level, always has smaller risk than the census counts.

It is also of interest to determine the behaviour of the census-based risk minus the Bayes-then-synthetic-based risk as a function of the level; the larger this difference, the more advantageous it is to adjust the census counts. Here use $f\left(C_{i}\right)=1 / C_{i}$ in loss function (2.15). It is possible to show (Cressie 1988) that as disaggregation proceeds to a lower level, the "risk gap" between Bayes-then-synthetic estimation and census counts widens in absolute terms. Although this is proved there for the uba-then-synthetic- based estimator, the same is true for cba-then-synthetic-based and sya-then- synthetic-based estimators, and the ordering of risks (2.32) is preserved at any level of disaggregation. This conclusion depends on the model (2.7) and (2.10) holding at all levels. Unfortunately at the lower levels there is some evidence that biases can be substantial. That is, $E\left(F_{j i}\right)=F_{j}+b_{j i} ; E\left(X_{j i} \mid F_{j i}\right)=F_{j i}+d_{j i}$. Realistically $b_{j i}$ 's and $d_{j i}$ 's are never zero, but at sufficiently high levels of aggregation they are unimportant. At the block and enumeration-district level they can be substantial (Cressie and Dajani 1988) and could invalidate the risk inequalities proved so far. Moreover, at lower levels, the data $\left\{X_{j i}\right\}$ are more variable leading to less precise estimates of $D_{j}=\tau_{j}^{2} /\left(\tau_{j}^{2}+\sigma_{j}^{2}\right)$ in the empirical Bayes version (see Section 3) of the Bayes estimator (2.14). These observations, as well as a recognition of the difference between risk and loss, help to explain the deterioration of the performance of the adjusted counts at lower levels, observed in artificial populations (Schultz et al. 1986).

## 3. EMPIRICAL BAYES ADJUSTMENT OF CENSUS COUNTS

Obtain from (2.14), (2.21), and (2.5), the estimated (or adjusted) true area counts $Y_{i}^{\text {uba }}$, $Y_{i}^{\text {cba }}$, and $Y_{i}^{\text {sya }}$, respectively. In order to make these functions only of the data, estimators are needed for the unknown parameters $F_{j}, \tau_{j}^{2}$, and $\sigma_{j}^{2}$; Fay and Herriot (1979) give empirical Bayes estimators in a regression setting, of which the model (2.7), (2.10) is a special case. For reasons of statistical consistency (see Cressie 1986, Section 3.3), choose,

$$
\begin{gather*}
\hat{F}_{j}=X_{j}  \tag{3.1}\\
\hat{\tau}_{j}^{2}=\max \left\{\left[\sum_{i} C_{j i} I\left(C_{j i}>0\right)\left(X_{j i}-X_{j} .\right)^{2} /\left(\sum_{i} I\left(C_{j i}>0\right)-1\right)\right]-\hat{\sigma}_{j}^{2}, 0\right\} \tag{3.2}
\end{gather*}
$$

$\hat{\sigma}_{j}^{2}$ is obtained from sampling considerations: it is known for dual-system estimation, and Schultz et al. (1986) determine it for their artificial populations by replicating probability-proportional-to-size sampling of 1,440 enumeration districts from the approximately 300,000 total number.

Statistical stability (i.e., small sampling variance) for sample means is easier to achieve than for sample variances. The coefficient of variation of the sample variance is approximately $\sqrt{2} / \sqrt{n}$; therefore to achieve a relative confidence region ( $0.5,1.5$ ) for the population variance, a value of $n=32$ is needed; and to achieve a region ( $0.95,1.05$ ) a value of $n=3,200$ is needed. Thus the estimator, $\sum_{i=1}^{l} C_{j i} I\left(C_{j i}>0\right)\left(X_{j i}-X_{j} .\right)^{2} /\left(\sum_{i=1}^{I} I\left(C_{j i}>0\right)-1\right)$ of $\tau_{j}^{2}+\sigma_{j}^{2}$ is very unstable, particularly when there are a large number of strata and hence $\sum_{i=1}^{l} I\left(C_{j i}>0\right)$ is small (smaller than 30 ).

One way around this is to introduce a further mixing distribution into the problem, namely, model the $\left\{\tau_{j}^{2}: \mathrm{j}=1, \ldots, J\right\}$ as being generated by the reciprocal of a gamma distribution for example. Thus instead of estimating $J$ parameters $\left\{\tau_{j}^{2}: \mathrm{j}=1, \ldots, J\right\}$, the problem can be reduced to estimating just two gamma parameters (see e.g., Hui and Berger 1983). Another possibility is to aggregate temporarily some of the strata for the purpose of estimating the stratum variance. In other words, define disjoint groups of strata indices, $A_{1}, \ldots, A_{K}$, such that $\cup\left\{A_{k}: k=1, \ldots, K\right\}=\{1,2, \ldots, J\}$, and $\tau_{j}^{2}=\tau_{j^{\prime}}^{2}=T_{k}^{2}$, whenever $j$ and $j^{\prime}$ belong to the same $A_{k}$. In this way, Cressie and Dajani (1988) reduce the number of stratum variance parameters from $J=96$ down to $K=4$. For the data analyzed below, since $\sum_{i=1}^{l} I\left(C_{j i}>0\right)=51$ for each of the three race strata, it was not necessary to "borrow strength" in the ways just described.

### 3.1 Emprical Bayes Estimators

The usual (see, for example, Morris 1983) and constrained (Louis 1984) empirical Bayes estimators can now be constructed:

$$
\begin{align*}
& F_{j i}^{\mathrm{ueb}}=X_{j}+\left\{\hat{\tau}_{j}^{2} /\left(\hat{\tau}_{j}^{2}+\hat{\sigma}_{j}^{2}\right)\right\}\left(X_{j i}-X_{j}\right),  \tag{3.3}\\
& Y_{i}^{\mathrm{ueb}}=\sum_{j=1}^{J} F_{j i}^{\mathrm{ueb}} C_{j i} ; i=1, \ldots, I ;  \tag{3.4}\\
& F_{j i}^{\mathrm{ceb}}=X_{j} .+\left\{\hat{\tau}_{j}^{2} /\left(\hat{\tau}_{j}^{2}+\hat{\sigma}_{j}^{2}\right)\right\}^{1 / 2}\left(X_{j i}-X_{j} .\right),  \tag{3.5}\\
& Y_{i}^{\mathrm{ceb}}=\sum_{j=1}^{J} F_{j i}^{\mathrm{ceb}} C_{j i} ; i=1, \ldots, I . \tag{3.6}
\end{align*}
$$

The usual empirical Bayes estimator (3.3) can also be obtained from standard theory for linear models with random effects (Henderson 1976).

Notice that when $\hat{\tau}_{j}^{2}=0$, the empirical Bayes estimators of the $j$-th stratum adjustment factors all reduce to the synthetic estimator $X_{j}$. The presence of the weight $\left\{\hat{\tau}_{j}^{2} /\left(\hat{\tau}_{j}^{2}+\hat{\sigma}_{j}^{2}\right)\right\}^{1 / 2}$ in the constrained empirical Bayes estimator (3.5) may look a little strange at first, but it is seen in Cressie (1987a) to yield an unbiased estimator of the stratum error $C_{j i}^{1 / 2}\left(F_{j i}-F_{j}\right)$.

An earlier suggestion for empirical Bayes modeling of undercount came from Dempster and Tomberlin (1980), who proposed that the number of undercounted people in a subarea might be a binomial random variable. They defined a heirarchical Bayes model but did not take into account the heteroskedastic variation. Stroud (1987) introduces a covariate into a two-stage Bayesian model, but his assumptions of homoskedastic variation and equal sample sizes in each subarea, are too restrictive for the problem considered in this article.

Formulas for the bias and mean-squared error of the usual empirical Bayes (ueb) estimators (3.3), (3.4), the constrained empirical Bayes (ceb) estimators (3.5), (3.6), and the synthetic estimators

$$
\begin{gather*}
F_{j i}^{\mathrm{syn}}=X_{j}  \tag{3.7}\\
Y_{i}^{\mathrm{syn}}=\sum_{j=1}^{J} F_{j i}^{\mathrm{syn}} C_{j i} ; i=1, \ldots, I, \tag{3.8}
\end{gather*}
$$

are given in Cressie (1987a, Section 4). Since undercount is a nonlinear function of the true population, its estimators based on $\left\{F_{j i}^{\text {est }}: i=1, \ldots, I ; j=1, \ldots, J\right\}$, viz.

$$
\begin{align*}
& u_{j i}^{\text {est }} \equiv 1-\frac{1}{F_{j i}^{\text {est }}} ; i=1, \ldots, I ; j=1, \ldots, J,  \tag{3.9}\\
& u_{i}^{\text {est }} \equiv 1-\frac{\mathrm{C}_{\mathrm{i}}}{Y_{i}^{\text {est }}} ; i=1, \ldots, I, \tag{3.10}
\end{align*}
$$

are biased; estimated biases and mean-squared errors can be obtained by the $\delta$-method (Cressie 1987a, Section 4). All of these bias and mean-squared error calculations do not take into account variation due to the (nonlinear) estimation of $\tau_{j}^{2} /\left(\tau_{j}^{2}+\sigma_{j}^{2}\right)$.

Suppose that the following three U.S. strata (based on race/ethnicity) are chosen: blacks, nonblack hispanics, and others. Data from the post-enumeration survey following the 1980 U.S. Census are given in Cressie (1987a, Table 1). These are from the noninstitutional population (Cowan and Bettin 1982) and have been labeled "PEP 3-8" by the U.S. Census Bureau the " 3 " refers to census omissions being obtained from an April survey and to imputing missing data, and the " 8 " refers to erroneous enumerations being obtained from a separate survey that imputed missing data with the help of U.S. Post Office information.

From these data and (3.1), (3.2), Cressie (1987a) estimated the mean of the mixture distribution, and standardized stratum and sampling variances defined in (2.7) and (2.10):

$$
\begin{array}{llll}
\text { blacks: } & \hat{F}_{1}=1.06076 & \hat{\tau}_{1}^{2}=673.982 & \hat{\sigma}_{1}^{2}=522.183, \\
\begin{array}{llll}
\text { nonblack } \\
\text { hispanics: }
\end{array} & \hat{F}_{2}=1.04667 & \hat{\tau}_{2}^{2}=308.990 & \hat{\sigma}_{2}^{2}=246.585, \\
& & &  \tag{3.13}\\
\text { Others: } & \hat{F}_{3}=0.99981 & \hat{\tau}_{3}^{2}=242.134 & \hat{\sigma}_{3}^{2}=242.152 .
\end{array}
$$

Based on these parameter estimators and the PEP 3-8 data $\left\{X_{j i}: j=1,2,3 ; I=1, \ldots, 51\right.$ \}, Cressie (1987a) gave undercount estimates $\left\{u_{j i}^{\text {est }}\right\},\left\{u_{i}^{\text {est }}\right\}$ for ueb-based and syn-based estimators defined by (3.3) and (3.7) respectively.

To check the fit of the model, the residuals $\left\{C_{j i}^{1 / 2}\left(F_{j i}^{\text {ceb }}-F_{j}^{\text {ceb }}\right): i=1, \ldots, I\right\}$ were computed for each of the three strata. Table 1 shows the results, presented as stem-and-leaf plots for the three race strata; a bell-shaped plot for each is the ideal. The model appears to fit the data, except for the nonblack-hispanic stratum in the state of New York. In light of the lawsuit, Cuomo vs. Baldridge, heard by the Southern District Court of New York in 1983, this new way of looking at the data tells an interesting story. The nonblack hispanics in New York State
Blacks $(j=1)$
Stem and leaf plots of residuals based on "ceb" estimator
Nonblack ( $j=2$ )
$(\mathbf{I}=!)$ s_3 110 MULTIPLY STEM.LEAF BY $10^{* *+01}$

were grossly undercounted, even in relation to their undercounted fellow nonblack hispanics in other states. Incidentally, the judge decided in favour of the U.S. Department of Commerce (in December 1987) on the grounds that the statistical and demographic professions had not developed adequate methods of adjustment for the whole country by 1980.

When are census counts improved by replacing $\left\{C_{i}: i=1, \ldots, I\right\}$ with $\left\{Y_{1}^{\text {est }}: i=1\right.$, $\ldots, I]$ ? The next section gives conditions under which an analogous ordering to (2.32) still holds in the empirical Bayes setting.

### 3.2 Adjustment at Different Levels; Model Parameters Estimated

The same comments at the beginning of Section 2.3 apply; in a model-based approach a small risk does not guarantee a small loss in every problem but only on the average. Also the analogous aggregation property to (2.24) holds for ueb-based, ceb-based, and syn-based estimators, namely

$$
\begin{equation*}
Y_{i}^{\text {est }}+Y_{i}^{\text {est }}=Y_{i d i}^{\text {est }} \tag{3.14}
\end{equation*}
$$

for "est" = "ueb," "ceb," and "syn," given by (3.4), (3.6), and (3.8) respectively. Moreover the disaggregation-aggregation property (2.27), namely

$$
\begin{equation*}
Y_{i_{1}}^{\text {sye }}+Y_{i_{2}}^{\text {sye }}=Y_{i}^{\text {est }} \tag{3.15}
\end{equation*}
$$

where $i=i_{1} \& i_{2}$ and $F_{i_{1}}^{\text {sye }}=F_{j_{2}}^{\text {sye }}=F_{j i}^{\text {est }}$, holds for any estimator of $F_{j i}$, including those based on ueb, ceb, and syn.

Write the risk of estimating $Y_{i}$ by $Y_{i}^{\text {est }}\left(=\sum_{j=1}^{J} F_{j i}^{\text {est }} C_{j i}\right)$ as

$$
\begin{equation*}
\text { est-risk }_{i} \equiv E\left[\left(Y_{i}^{\text {est }}-Y_{i}\right)^{2} f\left(C_{i}\right)\right] \tag{3.16}
\end{equation*}
$$

The estimators given by "est" = "ueb," "ceb," and "syn," will be compared to "cen" ( $F_{j i}^{\text {cen }} \equiv 1$ ) via (3.16). For the rest of this section consider the estimator,

$$
\begin{equation*}
F_{j i}^{\text {est }}=r_{j} X_{j i}+\left(1-r_{j}\right) X_{j} ; 0 \leq r_{j} \leq 1, \tag{3.17}
\end{equation*}
$$

a convex combination of the data $X_{j i}$ and the synthetic estimator $X_{j}$. Then

$$
\begin{equation*}
\text { est-risk }_{i}=\sum_{j=1}^{J} \tau_{j}^{2}\left(1-r_{j}\right)^{2}\left\{C_{j i}-\frac{C_{j i}^{2}}{\sum_{h} C_{j h}}\right\}+\sigma_{j}^{2}\left\{r_{j}^{2} C_{j i}+\frac{\left(1-r_{j}^{2}\right) C_{j i}^{2}}{\sum_{h} C_{j h}}\right\} \tag{3.18}
\end{equation*}
$$

It is easy to see that the value of $r_{j}$ that minimizes (3.18) is $r_{j}=D_{j}=\tau_{j}^{2} /\left(\tau_{j}^{2}+\sigma_{j}^{2}\right) ;$ i.e., neglecting the effect of estimating $\tau_{j}^{2}$ and $\sigma_{j}^{2}$, I obtain

$$
\begin{equation*}
\text { ueb-risk }_{i} \leq \text { est-risk }_{i} ; 0 \leq r_{j} \leq 1 \tag{3.19}
\end{equation*}
$$

Now compare ueb-risk (put $^{r_{j}}=D_{j}$ in (3.17)) with cen-risk ${ }_{i}$; recall from (2.29)

$$
\begin{equation*}
\text { cen-risk }_{i}=\sum_{j=1}^{J} \tau_{j}^{2} C_{j i} f\left(C_{i}\right)+\left[\sum_{j=1}^{J}\left(F_{j}-1\right) C_{j i}\right]^{2} f\left(C_{i}\right) \tag{3.20}
\end{equation*}
$$

Also, by putting $\tau_{j}^{2}=k_{j} \sigma_{j}^{2} ; j=1, \ldots, J$,

$$
\begin{equation*}
\text { ueb-risk }_{i}=\sum_{j=1}^{J} \sigma_{j}^{2}\left\{\frac{k_{j}}{1+k_{j}}+\frac{C_{j i}}{\sum_{h} C_{j h}} \cdot \frac{1}{1+k_{j}}\right\} C_{j i} f\left(C_{i}\right) \tag{3.21}
\end{equation*}
$$

A sufficient condition for ueb-risk ${ }_{i} \leq$ cen-risk $_{i}$ is,

$$
\left\{\frac{k_{j}}{1+k_{j}}+\frac{C_{j i}}{\sum_{h} C_{j h}} \cdot \frac{1}{1+k_{j}}\right\} \leq k_{j}
$$

that is, if

$$
\begin{equation*}
\sigma_{j}^{2} / \tau_{j}^{2} \leq\left\{\sum_{h} C_{j h} / C_{j i}\right\}^{1 / 2} ; j=1, \ldots, J \tag{3.22}
\end{equation*}
$$

then

$$
\begin{equation*}
\text { ueb-risk }_{i} \leq \text { cen-risk }_{i} \tag{3.23}
\end{equation*}
$$

Similarly, it can be shown that if

$$
\begin{equation*}
\sigma_{j}^{2} / \tau_{j}^{2} \leq 1 ; j=1, \ldots, J \tag{3.24}
\end{equation*}
$$

then

$$
\begin{equation*}
\text { syn-risk }_{i} \leq \text { cen-risk }_{i} . \tag{3.25}
\end{equation*}
$$

Finally, if $\left(\sigma_{j}^{2} / \tau_{j}^{2}\right) \leq 1$, and
$4\left(\sigma_{j}^{2} / \tau_{j}^{2}\right)^{2}\left(\frac{C_{j i}}{\sum_{h} C_{j h}}\right)^{2}-\left(\sigma_{j}^{2} / \tau_{j}^{2}\right)\left(1+\frac{2 C_{j i}}{\sum_{h} C_{j h}}\right)+3 \geq 0 ; j=1, \ldots, J$,
then

$$
\begin{equation*}
\text { ceb-risk }_{i} \leq \text { syn-risk }_{i} \tag{3.27}
\end{equation*}
$$

once again (from (3.26)), if $\sigma_{j}^{2} / \tau_{j}^{2}$ is small, risks can be bounded.

Therefore an analogous sequence of inequalities to (2.32) is possible:

$$
\begin{equation*}
\text { ueb-risk }_{i} \leq \text { ceb-risk }_{i} \leq \text { syn-risk }_{i} \leq \text { cen-risk }_{i} \tag{3.28}
\end{equation*}
$$

where the middle inequality requires the condition (3.26) and the last inequality requires the condition (3.24). If either of these two inequalities do not hold, at least the ueb-based estimator is an improvement over the census counts if condition (3.22) is satisfied. For the PEP 3-8 data from the 1980 U.S. Census,

$$
\begin{equation*}
\hat{\sigma}_{1}^{2} / \hat{\tau}_{1}^{2}=0.77, \hat{\sigma}_{2}^{2} / \hat{\tau}_{2}^{2}=0.80, \hat{\sigma}_{3}^{2} / \hat{\tau}_{3}^{2}=1.00 ; \tag{3.29}
\end{equation*}
$$

that is, for the 1980 U.S. decennial census the census risk is larger than the synthetic risk and the usual-empirical-Bayes risk is smallest of all.

Now compare the risk of using $Y_{i_{1}}^{\text {sye }}$ and $Y_{i_{2}}^{\text {sye }}$ (estimators of $Y_{i_{1}}$ and $Y_{i_{2}}$ respectively, based on $F_{j i}^{\text {est }}$ given by (3.17)), with the risk of using $C_{i_{1}}$ and $C_{i_{2}}$, where area $i=i_{1} \& i_{2}$ is disaggregated into two disjoint areas $i_{1}$ and $i_{2}$.

$$
\begin{align*}
& \sum_{\ell=1}^{2} E\left[\left(Y_{i_{\ell}}^{\text {sye }}-Y_{i_{\ell}}\right)^{2} f\left(C_{i_{\ell}}\right)\right] \\
& =\sum_{\ell=1}^{2} \sum_{j=1}^{J}\left[\tau_{j}^{2}\left\{\left(1-r_{j}\right)^{2}\left(\frac{1}{C_{j i}}-\frac{1}{\sum_{h} C_{j h}}\right)+\left(\frac{1}{C_{j i \ell}}-\frac{1}{C_{j i}}\right)\right\}\right. \\
& \left.+\sigma_{j}^{2}\left\{\frac{1-r_{j}^{2}}{\sum_{h} C_{j h}}+\frac{r_{j}^{2}}{C_{j i}}\right\}\right] C_{j i \ell}^{2} f\left(C_{i_{\ell}}\right) . \tag{3.30}
\end{align*}
$$

It is easy to see that under precisely the same conditions (3.22), (3.24), (3.26), the same sequence of inequalities (3.28) holds; interpret est-risk ${ }_{i}$ in (3.28) as being equal to (3.30) with $r_{j}=D_{j}$ for "est" = "ueb," with $r_{j}=D_{j}^{1 / 2}$ for "est" = "ceb," and with $r_{j}=0$ for "est" = "syn". Moreover for the loss function (2.15) with $f\left(C_{i}\right)=1 / C_{i}$, risk gaps widen as lower levels of aggregation are attained.

## 4. DISCUSSION

Various assumptions are made in deriving the risk inequalities (3.28), all of which deserve further investigation. The model (2.7) and (2.10) is assumed to fit, and in particular the independence of distributions between subareas is assumed. Moreover, the effect of estimating $D_{j}$ in the empirical Bayes estimators of $F_{j i}$ is assumed negligible. Notice however that synthetically estimated $F_{j i}$ 's do not use an estimate of $D_{j}$ and so those risk inequalities only rely on the appropriateness of the model (2.7), (2.10).

The conditions which order the various risks and bound them below the census risk in (3.28), all depend on $\sigma_{j}^{2} / \tau_{j}^{2}$ being "small." The practical implication is that a large number of households need to be chosen in the post-enumeration survey (PES) or there can be no guarantee that census counts can be improved by adjustment. With prior knowledge of stratum variation (e.g., from a previous census), the PES could be designed so that the conditions are satisfied.

After the survey has been conducted and the data $\left\{X_{j i}: i=1, \ldots, I ; j=1, \ldots, J\right\}$ are available, the various conditions (3.22), (3.24), and (3.26) can all be checked by using the estimators $\hat{\tau}_{j}^{2}$ and $\hat{\sigma}_{j}^{2}$ given by (3.2).

Concentrate on the best convex combination of $X_{j i}$ and $X_{j}$, namely $F_{j i}^{\text {ueb }}$ given by (3.3). Then, ueb-risk ${ }_{i} \leq$ cen-risk $_{i}$, if (3.22) holds; i.e., if

$$
\begin{equation*}
\sigma_{j}^{2} / \tau_{j}^{2} \leq\left\{\sum_{h} C_{j h} / C_{j i}\right\}^{1 / 2} ; j=1, \ldots, J . \tag{4.1}
\end{equation*}
$$

Notice that the condition is less stringent when the $i$-th area has a small census population; conversely, areas of large census population may have a ueb-based estimated population further from the truth than census. A sufficient condition for (4.1) to hold is, $\sigma_{j}^{2} / \tau_{j}^{2} \leq 1$; $j=1, \ldots, J$, which is also the condition that guarantees the syn-based estimated population improves over census. This condition was satisfied for the 1980 PEP 3-8 data (see Section 3.2).

Finally, the condition (4.1) becomes less stringent at lower levels, and indeed the results of Section 3.2 show that the risk gap between the adjusted population and the census population widens. This deserves comment. The results are true provided the model holds at lower levels, but this is probably not the case at the block and the enumeration-district level. Presence of bias in (2.7) and (2.10); namely

$$
\begin{equation*}
E\left(F_{j i}\right)=F_{j}+b_{j i} ; E\left(X_{j i} \mid F_{j i}\right)=F_{j i}+d_{j i} \tag{4.2}
\end{equation*}
$$

could cause a reversal in some of the risk inequalities. At the state level however, Table 1 and Cressie (1988) show through an examination of residuals, that (2.7) and (2.10) does fit for the 1980 PEP 3-8 data. And since (3.29) implies that condition (4.1) is satisfied, one can be confident that ueb-based adjusted state totals are closer to the truth than census state totals. That may not be true at the block level; clearly a decision regarding the level at which it is most important to have accurate census counts, needs to be made. The first use of U.S. Census data is the reporting of state totals to Congress for the purpose of redistricting House seats. One might include a number of large cities in with the states, and create e.g., the "states" New York City, and New York State Except New York City. It seems to me that this "state" level is the most sensitive politically and that accurate totals at this level should receive the highest priority.

## ACKNOWLEDGEMENT

This research has benefited from the input of members of the Undercount Research Staff at the U.S. Bureau of the Census, and from the perceptive comments of J.B. Kadane and J.W. Tukey. Support from Joint Statistical Agreements JSA 86-5, JSA 87-4, JSA 87-10, and JSA 88-13, between Iowa State University and the Census Bureau is gratefully acknowledged. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau.

## REFERENCES

COWAN, C.D., and BETTIN, P.J. (1982). Estimates and missing data problems in the post enumeration survey. Internal Report. Statistical Methods Division, Bureau of the Census, Washington, D.C.
CRESSIE, N. (1986). Empirical Bayes estimation of undercount in the decennial census. Statistical Laboratory Preprint 86-58, Iowa State University, Ames, IA.

CRESSIE, N. (1987a). Empirical Bayes estimation of undercount in the decennial census. Manuscript submitted to Journal of the American Statistical Association.
CRESSIE, N. (1987b). Comment on "Census undercount adjustment and the quality of geographic population distributions," by A.L. Schirm and S.H. Preston. Journal of the American Statistical Association, 82, 980-983.

CRESSIE, N. (1988). Estimating census undercount at national and subnational levels. Proceedings of Bureau of the Census Fourth Annual Resarch Conference. Bureau of the Census, Washington, D.C., 123-150.

CRESSIE, N., and DAJANI, A. (1988). Empirical Bayes estimation of U.S. undercount based on artificial populations. Statistical Laboratory Preprint 88-17, Iowa State University, Ames, IA.
DEMPSTER, A.P., and TOMBERLIN, T.J. (1980). The analysis of census undercount from a postenumeration survey, in Proceedings of the 1980 Conference on Census Undercount. Bureau of the Census, Washington, D.C. 88-94.
ERICKSEN, E.P., and KADANE, J.B. (1985). Estimating the population in a census year: 1980 and beyond. Journal of the American Statistical Association, 80, 98-131.
ERICKSEN, E.P., and KADANE, J.B. (1987). Sensitivity analysis of local estimates of undercount in the 1980 U.S. Census, in Small Area Statistics, (Eds. R. Platek, J.N.K. Rao, C.E. Särndal and M.P. Singh.) New York: Wiley, 23-45.

ERICKSEN, E.P., KADANE, J.B., and TUKEY, J.W. (1987). Adjusting the 1980 census of housing and population. Technical Report No. 401, Department of Statistics, Carnegie-Mellon University, Pittsburgh, PA.
FAY, R.E. III, and HERRIOT, R.A. (1979). Estimates of income for small places: An application of James-Stein to census data. Journal of the American Statistical Association, 74, 269-277.
FERGUSON, T.S. (1967). Mathematical Statistics: A Decision Theoretic Approach. New York: Academic Press.

FREEDMAN, D.A., and NAVIDI, W.C. (1986). Regression models for adjusting the 1980 census. Statistical Science, 1, 3-39.
GOLDSTEIN, M. (1975). Approximate Bayes solutions to some nonparametric problems. Annals of Statistics, 3, 512-517.
HENDERSON, C.R. (1976). A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics, 32, 69-83.
HUI, S.L., and BERGER, J.O. (1983). Empirical Bayes estimation of rates in longitudinal studies. Journal of the American Statistical Association, 78, 753-760.
ISAKI, C.T., DIFFENDAL, G.J., and SCHULTZ, L.K. (1986). Statistical synthetic estimates of undercount for small areas. Proceedings of Bureau of the Census Second Annual Research Conference. Bureau of the Census, Washington, D.C., 557-569.

KADANE, J.B. (1984). Allocating Congressional seats among the states when state populations are uncertain. Technical Report No. 309, Department of Statistics, Carnegie-Mellon University, Pittsburgh, PA.

LINDLEY, D.V., and SMITH, A.F.M. (1972). Bayes estimates for the linear model. Journal of the Royal Statistical Society, Series B, 34, 1-41.
LOUIS, T.A. (1984). Estimating a population of parameter values using Bayes and empirical Bayes methods. Journal of the American Statistical Association, 79, 393-398.

MORRIS, C.N. (1983). Parametric empirical Bayes inference: theory and applications. Journal of the American Statistical Association, 78, 47-55.

MULRY-LIGGAN, M., and HOGAN, H. (1986). Research plan on census adjustment standards. Proceedings of Bureau of the Census Second Annual Research Conference. Bureau of the Census, Washington, D.C., 381-392.
NATIONAL ACADEMY OF SCIENCES (1985). The Bicentennial Census: New Directions for Methodology in 1990, (Eds. C.F. Citro and M.L. Cohen.) Washington: National Academy Press.

READ, T.R.C., and CRESSIE, N.A.C. (1988). Goodness-of-fit Statistics for Discrete Multivariate Data. New York: Springer-Verlag.

SCHULTZ, L.K., HUANG, E.T., DIFFENDAL, G.J., and ISAKI, C.T. (1986). Some effects of statistical synthetic estimation on census undercount of small areas. Proceedings of the Section on Survey Research Methods, American Statistical Association, 321-325.
STROUD, T.W.F. (1987). Bayes and empirical Bayes approaches to small area estimation, in Small Area Statistics, (Eds. R. Platek, J.N.K. Rao, C.E. Särndal and M.P. Singh.) New York: Wiley, 124-137.

TUKEY, J.W. (1981). Discussion of "Issues in adjusting for the 1980 census undercount," by Barbara Bailar and Nathan Keyfitz, presented at the Annual Meeting of the American Statistical Association, Detroit, MI.

# Imputation Strategies for Missing Values in Post-Enumeration Surveys 

DONALD B. RUBIN, JOSEPH L. SCHAFER, AND NATHANIEL SCHENKER ${ }^{1}$


#### Abstract

To estimate census undercount, a post-enumeration survey (PES) is taken, and an attempt is made to find a matching census record for each individual in the PES; the rate of successful matching provides an estimate of census coverage. Undercount estimation is performed within poststrata defined by geographic, demographic, and housing characteristics, $X$. Portions of $X$ are missing for some individuals due to survey nonresponse; moreover, a match status $Y$ cannot be determined for all individuals. A procedure is needed for imputing the missing values of $X$ and $Y$. This paper reviews the imputation methods used in the 1986 Test of Adjustment Related Operations (Schenker 1988) and proposes two alternative model-based methods: (1) a maximum-likelihood contingency-table estimation procedure that ignores the missing-data mechanism; and (2) a new Bayesian contingency table estimation procedure that does not ignore the missing-data mechanism. The first method is computationally simpler, but the second is preferred on conceptual and scientific grounds.


KEY WORDS: Bayesian methods; Categorical data; Coverage error; EM algorithm; Multiple imputation; Nonignorable nonresponse; Undercount.

## 1. INTRODUCTION

The U.S. Bureau of the Census has used a post-enumeration survey (PES) to evaluate coverage error in several past censuses, and it plans to conduct a PES after the 1990 Decennial Census as well. For each individual in the PES, an attempt is made to find a census record (i.e., a match) to determine whether the person was enumerated in the census. The proportion of PES persons who were missed in the census is used as an estimate of the proportion of persons in the population who were missed. A similar matching operation is performed to match a sample of individuals from the census to the PES; this provides an estimate of the census overcount resulting from erroneous (e.g., duplicate or fictitious) enumerations.

The data on matches and erroneous enumerations obtained from the PES are combined to estimate the population size via the dual-system estimator; this capture-recapture type of estimator is discussed in Marks, Seltzer and Krotki (1974), Krotki (1978), Wolter (1986), Diffendal (1988), and Fay, Passell and Robinson. (1988, Chapter 5). Dual-system estimates of population size are computed within poststrata defined by geographic, demographic (age, sex, race), and housing (owner/renter, type of housing structure) characteristics.

Two problems of missing data occur in the PES and complicate the estimation process:

1. Geographic, demographic, or housing characteristics may be missing for a person, so it is not known to which poststratum that person belongs.
2. After the processing of the PES, there are some individuals with match status (dichotomous variable indicating matched/not matched to census) or erroneous enumeration status missing. This can occur, for instance, when an incomplete name is obtained in the PES, or when there is difficulty in specifying a Census Day address for someone who moved between Census Day and the PES.
[^8]Missing data were a major source of uncertainty in undercount estimation for the 1980 Decennial Census (Freedman and Navidi 1986; Fay, Passell and Robinson 1988, Chapter 6). Improvements in the PES design should reduce the amount of missing data in 1990 (Hogan and Wolter 1988), but a method for dealing with missing data will still be necessary.

The 1986 Test of Adjustment Related Operations (TARO), a recent test of undercount estimation and adjustment (Diffendal 1988; Schenker 1988), used a PES that was similar in design to that planned for 1990 . This paper reviews the methods used to handle missing data in TARO (Schenker 1988), identifies potential weaknesses of these methods, and discusses potential alternatives.

Our goal is to indicate issues and problems, and to suggest methods for their solution. The long range plan for research is to carefully evaluate these methods. Although we only discuss imputation for missing PES data when estimating undercount, missing data also occur in the census sample used to estimate overcount. The missing-data problems in estimating overcount, however, are analogous to those in estimating undercount (Schenker 1988), and so our discussion applies to both problems.

In our discussion of alternatives to the TARO procedures, we propose a new method based on a Bayesian model that does not ignore the missing-data mechanism, and thus does not assume that the missing data are missing at random. Nonignorable models for incomplete categorical data are a recent development in the theory of handling missing data; see Fay (1986), Little and Rubin (1987, Section 11.6), and Baker and Laird (1988) for discussions and reviews of the literature. Moreover, the types of missing data that we discuss occur not only in undercount estimation, but in many other situations as well; thus our discussion is relevant to the general problem of handling missing categorical data.

Section 2 discusses the imputation methods used in TARO. In Section 3, alternative methods are described and illustrated using a simple example. Section 4 presents a concluding discussion.

## 2. IMPUTATION METHODS USED IN TARO

### 2.1 Description of Methods

For each individual in the PES, let $X$ denote categorical variables for age, sex, race, owner/renter status, and type of housing structure; let $Y$ denote match status ( $1=$ match, $0=$ nonmatch); and let $Z$ denote variables indicating whether the PES interview was with a household member or a proxy, and whether the PES person moved between Census Day and the PES. In TARO, the $X$ variables (except type of housing structure) were used in forming poststrata (Diffendal 1988); $Z$ was observed for all PES individuals, but $Y$ and components of $X$ were sometimes missing (Schenker 1988).

Missing values of $X$ and $Y$ were imputed in two stages. (Our description is simplified for ease of presentation; see Schenker (1988) for the precise procedure). First, all missing $X$ values were imputed using a "hot deck" scheme based on observed $X$ variables; that is, imputed values were drawn from the observed distributions of $X$ values. Second, after the missing values of $X$ were filled in, a logistic regression model predicting $Y$ from $X$ and $Z$ was fitted to the cases with $Y$ observed. This logistic regression model was then used to impute probabilities of match for all missing $Y$ values. Probabilities rather than zeros and ones were imputed to (a) increase the precision of estimation, and (b) allow the assessment of variability due to imputation (Schenker 1989).

### 2.2 Critique of Methods

The TARO imputation methods have many positive features. They are easily understood and use explicit modeling for the imputation of $Y$. They also condition on much of the observed data, rather than imputing from marginal distributions. Finally, in principle they allow the assessment of uncertainty in undercount estimates due to the missing $Y$ values. The methods have some potential weaknesses, however, which we now describe.

The TARO imputation procedure is an "ignorable" procedure, because it ignores the missing-data mechanism. Ignorable procedures assume that the missing data are missing at random (MAR) (Rubin 1976); that is, they assume that given the observed data, the missingness is independent of the values of the missing items. For example, if $X$ and $Z$ are observed for all people, MAR implies that $Y$ can be imputed using the conditional distribution of $Y$ given $X$ and $Z$ for those individuals having $X, Y$, and $Z$ observed.

The TARO procedure is actually a special case of an ignorable procedure, because it makes assumptions that are stronger than the general MAR assumption. The TARO procedure treated $X$ and $Y$ asymmetrically; that is, it imputed missing values of $Y$ conditional on all observed data, but it imputed missing $X$ values conditional only on the observed $X$ 's, rather than on the observed values of $X, Y$, and $Z$. Hence, in addition to the general MAR assumption, the TARO procedure also effectively assumed that, given the observed components of $X$, the missing components of $X$ are conditionally independent of both $Y$ and $Z$.

This additional independence assumption may not be realistic; it may be that given the observed $X$ data, there is a residual dependence of values of missing components of $X$ on $Y$ and/or $Z$. If this is the case, then observed values of $Y$ and $Z$ should be used in the imputation of $X$. For instance, suppose a PES individual has sex missing, but is found not to match any census record ( $Y=0$ ) on the basis of observed age, race, and address; and suppose males tend to be undercounted in the census more than females with identical other characteristics. Then knowing that $Y=0$ provides some evidence that the person in question is more likely to be male than if $Y$ were 1. The most general ignorable imputation procedure would use information provided by $Y$ and $Z$ in imputing missing $X$ values; this is one of the alternative imputation methods, which we discuss in Section 3.4.1.

Another feature of the TARO procedure that may be unrealistic is the ignorability assumption itself. It may be that the missing data are not MAR - i.e., given the observed data, the missingness is not independent of the values of the missing items; if so, then it would be more appropriate to use a nonignorable model for the missing-data mechanism. For instance, consider a group of people with identical values of all variables except race; it may be more difficult to obtain information on race for minorities than nonminorities, and consequently the distribution of race will be different among those missing race and those with race observed. Similarly, even after all $X$ and $Z$ variables are controlled for, it may be that people who were not enumerated in the census are more likely to be missing $Y$ than those who were enumerated in the census. An alternative imputation method based on a general class of nonignorable models is presented in Section 3.4.2.

## 3. ALTERNATIVE METHODS OF IMPUTATION IN THE PES

### 3.1 Introduction

Let $X=\left(X_{1}, X_{2}, X_{3}\right)$ denote three individual characteristics recorded by the PES (e.g., age, sex, and race). The variables $X_{1}, X_{2}$, and $X_{3}$ are assumed to be categorical, taking $I, J$, and $K$ possible values respectively. We have chosen three variables merely for illustrative purposes
and notational simplicity; all ideas developed here will extend immediately to any number of categorical variables. In practice, these $X$ variables will probably include the demographic, geographic, and housing characteristics used to define poststrata for undercount estimation; they may also include additional PES variables, such as mover status and household member/proxy status, which are not of intrinsic interest but which may be useful for imputation purposes.

We will form $I J K$ different classes of individuals by cross-classifying them according to $X_{1}$, $X_{2}$, and $X_{3}$. These classes may or may not be the same as the poststrata for undercount estimation; in practice the poststrata will probably be coarser than these classes. It is convenient, but not necessary, for these classes to be defined as cross-classifications of all possible values of $X_{1}, X_{2}$, and $X_{3}$; more complicated patterns (such as nested ones) are also possible. We will be constructing loglinear models for cross-classified contingency tables, but loglinear models may be based on other patterns as well.

Let $Y$ be the dichotomous variable denoting match status, taking values 1 (matched to census) or 0 (not matched). If there were no missing data, the results of the PES could be summarized in a single four-dimensional contingency table with $I \times J \times K \times 2$ cells, since each individual could be fully classified according to $X_{1}, X_{2}, X_{3}$, and $Y$. But those individuals missing one or more variables can be only partially classified according to those variables that are observed. Those having $X_{1}, X_{2}, X_{3}$, and $Y$ all observed will constitute a four-dimensional table, which we will call the table of complete cases ( $C C$ ), or the data table for missingness pattern 1 (no variables missing). Those having $X_{1}, X_{2}$, and $X_{3}$ observed but $Y$ missing will constitute a threedimensional supplementary table with IJK cells, which we will call the data table for missingness pattern 2 . In general, there will be $2^{4}$ such tables corresponding to all possible missingness patterns, one CC table and $2^{4}-1$ supplementary tables.

### 3.2 Imputation from Reference Tables

In our model-based approach to imputation, we will model the data tables for different missingness patterns as multinomial observations. Corresponding to each missingness pattern, we will define a set of cell probabilities $\theta^{t}=\left\{\Theta_{i j k l}^{\prime}\right\}$, where the superscript $t$ indexes the missingness pattern, $t=1, \ldots, 2^{4}$, and the subscripts $i, j, k$, and $l$ indicate the levels of $X_{1}, X_{2}$, $X_{3}$, and $Y$ respectively. Because we will refer to $\theta^{t}$ when imputing missing values for the $t$-th data table, we will call $\Theta^{t}$ the reference table for the $t$-th data table, and $\left\{\Theta^{t}: t=1, \ldots, 2^{4}\right\}$ the set of reference tables.

Imputation of missing values corresponds to expanding each supplementary data table to make it fully four-dimensional, according to its corresponding reference table. For example, consider the imputation of $Y$ for those individuals missing only $Y$. This is equivalent to expanding the supplementary data table for missingness pattern 2 , by dividing each cell count in this table into two parts, a count of those having $Y=1$ and a count of those having $Y=0$, split according to the reference table $\Theta^{2}$. With known $\Theta^{2}$ this procedure is straightforward: we first obtain from $\Theta^{2}$ the conditional distribution of $Y$ given $X$ for this missingness pattern, i.e.,

$$
\begin{equation*}
P\left(Y=1 \mid X_{1}, X_{2}, X_{3}, t=2\right)=\frac{\theta_{i j k 1}^{2}}{\theta_{i j k 0}^{2}+\theta_{i j k 1}^{2}} \tag{1}
\end{equation*}
$$

for $i=1, \ldots, I, j=1, \ldots, J$, and $k=1, \ldots, K$. Then, we impute $Y=1$ for each observation in cell $i j k$ of this table with probability given by the right-hand side of (1); alternatively, we could impute the mean of this distribution, which is just the probability of a match (1). The relative merits of random draw versus mean imputation for the PES will be discussed in Section 3.3.

Note that in the example above, the only information from $\theta^{2}$ needed for the imputation is the conditional distribution of $Y$ given $X$; hence, any value of $\Theta^{2}$ yielding the same values for (1) leads to the same imputation procedure. For an imputation procedure to be accurate, then, our estimate of $\theta^{t}$ need not correspond to the joint distribution of $Y$ and $X$ for the $t$-th missingness pattern; the only requirement is that the conditional distribution of the missing variables given the observed ones derived from our estimate of $\theta^{\prime}$ be close to the correct one.

In particular, if the missing-data mechanism is ignorable, one common reference table $\Theta^{t}=\Theta, t=1, \ldots, 2^{4}$, provides valid imputations for all missingness patterns, even though the joint distribution of $X$ and $Y$ might vary across missingness patterns. The fact that only one reference table is needed follows from the definition of ignorability, which implies that the conditional distribution of missing values given observed values does not depend on the missingness pattern. The value $\theta$ that provides valid imputations is not $\theta_{C C}$, the cell probabilities for the joint distribution of $X_{1}, X_{2}, X_{3}$, and $Y$ underlying the CC table; rather, it is the the joint distribution of $X_{1}, X_{2}, X_{3}$, and $Y$ marginalized across missingness patterns. Generally, if the missing-data mechanism is nonignorable, we will need to specify a different reference table for each missingness pattern.

In our model-based approach, the two crucial issues to be addressed are: (1) how to estimate the set of reference tables using well-established principles of efficient estimation; and (2) how to perform the imputation once these estimates are obtained. Two methods of estimation will be compared in Section 3.4; in Section 3.3 we briefly discuss various alternatives for imputation.

### 3.3 Single, Multiple, and Mean Imputation

Once the reference tables have been estimated, distributions for each individual's missing variables given the observed ones have been completely specified. In theory, these distributions could be used to analytically calculate correct point and interval estimates for any quantities of interest. In practice, however, these calculations are usually intractable; some other procedure is needed. Filling in the missing values by imputation is an attractive alternative, because it creates a completed dataset, which can be analyzed by complete-data methods. Little (1986) summarizes the strengths and weaknesses of various imputation methods; we shall only comment on aspects relevant to the PES.

In current practice, each missing value is typically filled in by taking a single random draw from a distribution, thereby producing a simulated complete dataset, which is analyzed in the usual complete-data fashion. Interval estimates derived from this method will be artificially too precise, because they do not reflect the uncertainties of the imputation. One remedy for this, which is coming into use, is multiple imputation (Rubin 1987), in which each missing value is replaced by $m$ random draws from the distribution. With moderate amounts of missing information, $m=5$ draws are enough to produce efficient point estimates and adequate interval estimates. With rates of missing information that appear likely in the PES (typically 5-10 percent or less, judging from TARO), $m=2$ draws will be perfectly adequate for essentially all purposes. In a large-scale survey like the PES, however, even a small number of multiple imputations may be computationally difficult to handle.

Since the estimates of interest in the PES are the match rates within poststrata, it is probably more important to accurately reflect the variability of imputation for $Y$ than for $X$; that is, it is probably more important to reflect uncertainty in overall undercount rates than uncertainty in the allocation of undercount to poststrata. Thus it may be possible to obtain adequate results by imputing a single set of $X$ values, and then multiply imputing $Y$ given $X$. Yet another possibility is to impute a single set of $X$ values, and then impute the probability of match given $X$. This approach was used in TARO (Schenker 1988); it allows the imputed $X$ 's and fractional $Y$ 's to be treated like single imputations when estimating undercount rates.

Choosing an acceptable imputation procedure given a set of reference tables is the subject of ongoing research. It is hoped that the TARO approach of imputing a single value of $X$ and then imputing $P(Y=1 \mid X)$ will prove to be a useful compromise between the accuracy of multiple imputation and the computational ease of single imputation.

### 3.4 Models and Methods of Estimation

In this section, we present two alternative procedures for modeling the missing data and estimating the reference tables for imputation. The two procedures are the Ignorable MaximumLikelihood (IML) method and a new Nonignorable Bayesian (NB) method that should be an improvement over IML if the missing data are not MAR.

### 3.4.1 The Ignorable Maximum-Likelihood Method

As mentioned previously, an ignorable imputation procedure needs to specify only a single reference table and apply it to all missingness patterns. One naive approach is to estimate this common reference table $\Theta$ by the cell proportions observed in the CC table. The resulting estimate $\hat{\theta}_{C C}$ is asymptotically unbiased for $\theta$ if the missing data are missing completely at random (MCAR), that is, if the probability of missingness for each item is completely independent of the data values, observed or missing. If the missing data are merely MAR, and not MCAR, then using $\hat{\theta}_{C C}$ for imputation introduces biases into the data. Moreover, even when the data are MCAR, $\hat{\theta}_{C C}$ is not efficient because it does not make use of all of the observed data to estimate $\theta$.

The IML method makes use of all the data, both in the CC table and in the suplementary tables, to estimate $\theta$. The estimated value $\hat{\theta}_{I M L}$ is chosen to maximize the likelihood ignoring the missing-data mechanism (Little and Rubin 1987, Section 5.3). In general, there is no closed form expression for $\hat{\theta}_{I M L}$; it must be obtained iteratively, for instance via the EM algorithm (Dempster, Laird and Rubin 1977; Little and Rubin 1987, Section 9.3).

The EM algorithm for contingency tables is easy to implement, and the resulting maximum likelihood estimate $\hat{\Theta}_{I M L}$ is both efficient and consistent under the assumption of ignorability; thus this EM procedure for IML is attractive from both computational and theoretical perspectives. When the missing data are not MAR, however, the IML method will generally introduce biases. Since there are good reasons to believe that the missing data in the PES are not missing at random, we propose a new method of estimation that makes a different assumption.

### 3.4.2 Nonignorable Modeling and Nonuniqueness of the MLE

When the missing data are not MAR, it is no longer valid to ignore the missing-data mechanism; the fact that a data value is missing conveys information about its value. Hence, a model that reflects this dependence must include indicator variables for response, indicating whether data values were observed or missing. Consequently, a nonignorable model will generally estimate a separate reference table for each missingness pattern, or equivalently, an expanded reference table $\theta$ with twice as many dimensions (i.e., with an additional dimension for each missingness indicator).

Let $R=\left(R_{1}, R_{2}, R_{3}, R_{Y}\right)$ be indicator variables for whether $X_{1}, X_{2}, X_{3}$, and $Y$ are observed, respectively; for example, $R_{1}=1$ if $X_{1}$ is observed and $R_{1}=0$ if $X_{1}$ is missing. Consider the eight-dimensional contingency table formed by cross-classifying individuals by $X, Y$, and $R$, and now let $\theta$ be the eight-dimensional table of cell probabilities for this expanded table.

Each individual in the survey belongs to a cell of the expanded table, but because some data are missing, we only observe certain margins of this table. Because $R$ is fully observed, any margin involving only missingness indicators is fully observed, but a margin involving $Y$ or one of the $X^{\prime}$ s might not be observed. For example, in the cross-section of the table with $R_{1}=R_{2}=R_{3}=1$ and $R_{y}=0$, we can classify individuals by $X_{1}, X_{2}$, and $X_{3}$, but not by $Y$; therefore we observe only the marginal totals obtained by summing across $Y$.

The number of parameters in the fully saturated model for this table is $2^{5} I J K-1$, which is larger than the number of observed sufficient statistics; hence the maximum-likelihood estimate (MLE) for $\theta$ is not uniquely determined. In order to obtain a unique estimate for $\theta$, one must impose additional structure.

One possible way to obtain a unique MLE is to build a log-linear model for the expanded contingency table, with some of the higher-order interactions set equal to zero (Little 1985; Fay 1986; Little and Rubin 1987, Section 11.6). We might try to set to zero those interactions that are not estimable from the data, but the formalization of this does not always work well in practice. For example, it may at first appear that the $R_{1}$ by $X_{1}$ interaction is not estimable, because the value of $X_{1}$ is never observed when $R_{1}=0$; however, the data may contain information about the $R_{1}$ by $X_{1}$ interaction indirectly through another variable, one that is observed for some individuals having $R_{1}=1$ and some having $R_{1}=0$. An example of a quantity that is truly inestimable from the data is $P\left(Y=1 \mid X_{1}=i\right.$, $X_{2}=j, X_{3}=k, R_{1}=R_{2}=R_{3}=1, R_{y}=0$ ), but this does not correspond to any single interaction term in the log-linear model parameterization. (By "truly inestimable" we mean in Rubin's (1974) sense that the parameter's posterior distribution equals its prior distribution for all priors).

In a dataset with a complicated pattern of missingness, it is not easy to find a set of loglinear terms that, if set to zero, will yield a unique MLE for $\theta$. The minimum number of terms that must be set to zero to produce uniqueness is $2^{5} I J K-1$, the dimension of $\theta$, minus the number of observed sufficient statistics. Even if such a minimal set can be found, it is usually not unique, and one is faced with the task of deciding which set of terms should be excluded from the model. Rather than attempting to obtain a unique MLE by placing these kinds of prior restrictions on the log-linear model, we will instead use a Bayesian approach involving the use of a prior distribution.

### 3.4.3 A Nonignorable Bayesian Method

In the Bayesian paradigm, one expresses prior assumptions about the parameters formally through a prior distribution. For our situation, a proper unimodal prior, when combined with the observed-data likelihood, produces a posterior distribution for $\theta$ that can yield a unique estimate; for example, we may take the posterior mode, $\hat{\theta}_{N B}$, as our estimate of $\theta$. This method is attractive because it automatically allows precise estimation of those functions of $\Theta$ about which the data contain much information, while using the prior to select appropriate values for those quantities that are strictly inestimable from the data. If applied properly, this method will produce a nonignorable model that fits the data as well as any other model - it essentially maximizes the likelihood function, and yet is as consistent as possible with our beliefs about the nature of the missing-data mechanism as expressed in the prior distribution.

Sound scientific practice suggests that we should choose a prior distribution that favors simple structure (i.e., small higher-order interactions) over complicated structure (i.e., large higher-order interactions). If we choose a prior that assigns a low (but nonzero) a priori probability to the presence of higher-order interactions in the log-linear model, then we will be making assumptions that are similar in nature to the assumptions of the IML method - that
missing values are not radically different from their observed counterparts in their relationships with other observed variables - although in a smoother, more systematic fashion than the IML method does.

Following the notation of Bishop, Fienberg, and Holland (1975), consider the saturated loglinear model for the eight-way contingency table for $R, X$, and $Y$,

$$
\begin{align*}
\log \theta_{i j k \ldots p}= & \mu+\mu_{1(i)}+\mu_{2(j)}+\ldots+\mu_{8(p)} \\
& +\mu_{12(i j)}+\mu_{13(i k)}+\ldots \\
& +\mu_{123 \ldots 8(j j \ldots p)} \tag{2}
\end{align*}
$$

where $\theta_{i j k \ldots p}$ is the probability that an observation falls in cell $i j k \ldots p$, and the $\mu$ 's are the oneway, two-way, three-way, and higher-order interactions. We propose the simple family of independent normal prior distributions

$$
\begin{align*}
\mu_{i} & \sim N\left(0, \sigma^{2}\right) \\
\mu_{i j} & \sim N\left(0, \sigma^{2} / \tau\right) \\
\mu_{i j k} & \sim N\left(0, \sigma^{2} / \tau^{2}\right) \\
& \vdots  \tag{3}\\
\mu_{i j k \ldots p} & \sim N\left(0, \sigma^{2} / \tau^{7}\right),
\end{align*}
$$

for some choice of $\sigma^{2}>0$ and $\tau>1$. This prior distribution pulls the higher-order interactions toward zero, and hence pulls the estimate of $\Theta$ toward a more parsimonious or simpler model. We believe that this approach will produce estimates of $\theta$ that are not too different from $\hat{\theta}_{I M L}$ when the missing data are truly MAR, but will be more robust than the IML method under departures from MAR. The only cases when IML will be superior occur when the missing data are MAR and strong higher-order interactions exist among the $X$ 's and $Y$.

Leonard (1975) and Laird (1978) examined log-linear models with normal prior distributions on the $\mu$ terms for complete data; our situation is complicated by the fact that only certain margins of the eight-way table are observed. Finding the posterior mode $\hat{\Theta}_{N B}$ under this model is conceptually straightforward; the EM algorithm can be applied to the posterior distribution of $\Theta$, just as to the likelihood function. The E-step remains the same; the M-step, however, poses some computational difficulties. The posterior distribution is nearly a ridge in high-dimensional space; it is very steep in certain directions, but nearly flat in others. The second-derivitive matrix is nearly singular along this ridge; hence Newton-Raphson and other gradient methods for maximization will not work well. Difficulty arises as $\sigma^{2}$ becomes large, because the ridge becomes flat as $\sigma^{2} \rightarrow \infty$ and a unique mode no longer exists. Difficulty also arises as the number of observations grows, because the posterior becomes very steep in certain directions and thus portions of the second-derivitive matrix become very large. More work is needed to develop effective methods for finding or approximating $\hat{\theta}_{N B}$.

### 3.4.4 A Numerical Example

We now present a simple numerical example and compare the results obtained from the IML and NB methods. For simplicity, we will only use a single dichotomous $X$ variable (taking values 0 or 1 ) and match status $Y$.

If there were no missingness, the data could be fully cross-classified by $X$ and $Y$ and hence summarized in a single $2 \times 2$ contingency table. With four patterns of missingness, however, the data are summarized in a CC table and three supplementary tables (Figure 1).

The CC estimate $\hat{\boldsymbol{\theta}}_{C C}$ is simply the observed proportions in Table A. The IML estimate $\hat{\theta}_{I M L}$ is found iteratively via the EM algorithm; using $\hat{\theta}_{C C}$ as the starting value, the algorithm converges in approximately four cycles. The NB estimate $\hat{\Theta}_{N B}$ was found using a prior distribution with $\sigma^{2}=10$ and $\tau=3$. This means that the one-way terms are a priori normally distributed about zero with variance 10 , so there is a 95 percent probability that the log-odds for each main effect lies inside the interval ( $-4 \sqrt{ } 10,+4 \sqrt{ } 10$ ). The two-way terms have variance $10 / 3$, the three-ways have variance $10 / 9$, and the four-ways have variance $10 / 27$; this represents a moderate pulling of the higher-order terms toward the origin. (Finding $\hat{\Theta}_{N B}$ for varying values of $\sigma^{2}$ and $\tau$ proved difficult, because of the numerical instability of the particular maximization routine applied at each M-step.) The values of $\hat{\theta}_{I M L}$ and $\hat{\theta}_{N B}$ are given in Figure 2. The expected imputations under these models are given in Figure 3, along with the expected imputations under $\hat{\theta}_{C C}$ for comparison.

The differences between the imputation methods can be seen most clearly by comparing the expected imputations for Table D. Imputation using $\hat{\Theta}_{C C}$ simply reproduces the proportions observed in Table A. Imputation using $\hat{\boldsymbol{\theta}}_{I M L}$ differs from imputation using $\hat{\theta}_{C C}$ because Tables B and C, as well as Table A, contribute to the estimation of $\theta$ and hence to the imputation for Table D.

Imputation using $\hat{\Theta}_{N B}$ is fundamentally different from imputation using $\hat{\theta}_{C C}$ or $\hat{\theta}_{I M L}$ in that it assumes missingness is informative. From Table $\mathbf{B}$, it surmises that missingness of $Y$ is associated with $X=0$. From Table C, it surmises that missingness of $X$ is associated with $Y=0$. It then combines this information in a smooth fashion to conclude that a larger proportion of the individuals who have both $X$ and $Y$ missing fall into the ( $X=0, Y=0$ ) category.

## 4. DISCUSSION

Our work is clearly at an early stage of development. Nevertheless, we feel that it has important potential applications, both specifically to the estimation of undercount using a PES, and generally to contingency table modeling when some data are missing. We conclude with two brief comments: first, on the need for continuing research on these procedures; and second, on the need to judge the relative propriety of models when devising an imputation procedure.


Table A Complete Cases


Table B $Y$ missing
$Y=1 Y=0$

| 28 | 60 |
| :--- | :--- |

Table C $X$ missing


Table D Both missing

Figure 1. Observed Data
$\hat{\Theta}_{I M L}$

| .279 | .174 |
| :--- | :--- |
| .239 | .308 |

All missingness patterns


Pattern A

| .193 | .140 |
| :--- | :--- |
| .305 | .361 |

Pattern B

| .165 | .262 |
| :--- | :--- |
| .153 | .419 |

Pattern C

| .211 | .224 |
| :--- | :--- |
| .222 | .333 |
| Pattern D |  |

Pattern D

Figure 2. Reference Tables for Imputation

Observed Data
$Y=1 \quad Y=0$

|  |  |  |
| :--- | ---: | ---: |
|  | 100 | 50 |
|  |  |  |
|  | 75 |  |
|  |  | 75 |

Table A

| 20 | 10 |
| :---: | :---: |
| 30 | 30 |


| 18.5 | 11.5 |
| :--- | :---: |
| 26.2 | 33.8 |


| 17.4 | 12.6 |
| :--- | :--- |
| 27.5 | 32.5 |

Table B

$$
Y=1 \quad Y=0
$$

| 28 | 60 |
| :--- | :--- |


| 16 | 24 |
| :---: | :---: |
| 12 | 36 |


|  |  |
| :---: | :---: |
| 15.1 | 21.7 |
| 12.9 | 38.3 |


| 14.5 | 23.1 |
| :---: | :---: |
| 13.5 | 36.9 |

Table C
12

| 4 | 2 |
| :--- | :--- |
| 3 | 3 |


| 3.35 | 2.09 |
| :--- | :--- |
| 2.86 | 3.69 |


| 2.54 | 2.70 |
| :--- | :--- |
| 2.68 | 4.09 |

Table D
Figure 3. Expected Imputations Under $\hat{\theta}_{C C}, \hat{\theta}_{I M L}, \hat{\theta}_{N B}$

### 4.1 Continuing Research

Two kinds of research efforts are needed before our NB method can become broadly applicable. First, computationally-oriented research is needed to address the ridge-like posterior distribution. Alternatives to the mode, such as the posterior mean, are worth considering. Furthermore, measures of uncertainty should also be calculated, and considering the odd nonnormal shape of the posterior, these may not be simple to summarize or compute. One strategy focuses directly on drawing multiple values of $\Theta$ from this posterior distribution without explicitly finding the posterior mode or the mean; these draws of $\Theta$ may be used to multiply impute the missing data.

Related to the issue of measuring uncertainty is the issue of performance in repeated sampling experiments. Although we believe our Bayesian approach is fully appropriate, it is important for broad application to evaluate the operating characteristics of this procedure in the wide range of circumstances to which it might be routinely applied. For example, how well does it work in realistic cases when, unknown to the data analyst, the missing data are MAR?

These topics will be the focus of a major continuing research effort.

### 4.2 The Need to Judge the Relative Propriety of Models

Considering the fully saturated model for ( $X, Y, R$ ) with parameter $\theta$, any method of imputation, no matter how illogical, can be viewed as the correct procedure under some model. For example, consider imputation using $\hat{\boldsymbol{\theta}}_{C C}$ as the reference table for all missingness patterns. This posits conditional distributions for the missing data, given the observed data and $R$, about which there is no information in the observed values. Hence, coupling these distributions with the estimable distributions (the distributions of $R$ and the observed data) implies an estimate for $\theta$, which maximizes the likelihood under the saturated model! It is not a very sensible answer, since it corresponds to the unique MLE under a model in which all sorts of conditional distributions given various missingness patterns $R$ are equal to the conditional distributions given $R=(1,1, \ldots, 1)$; however, if we consider the likelihood function only, there is no reason to prefer any other maximum-likelihood estimate to this one.

Even stranger methods of imputation, such as "impute all missing values as zero," correspond to particular models with estimated $\theta$ 's that are MLE's under the saturated model, but they violate good sense. Any sensible attempt to impute missing data values is based on the belief that two individuals with similar values of observed characteristics, and similar missingness patterns, are not radically different in those characteristics that are observed for one and missing for the other. Our NB method formalizes this notion of smoothness by specifying a contingency table model with small higher-order interactions.

Choosing one imputation procedure over another, then, cannot be done on maximum-likelihood-type principles alone, but must involve consideration of the propriety of the underlying prior specifications. This is not really a serious problem; sound statistical practice has always advocated the use of smooth or parsimonious models when less smooth models fit the data equally well. Consider fitting straight lines or polynomial curves through a collection of data points; simpler models are preferable to complicated ones on scientific grounds - the same issues arise in imputation. We believe that the model, given by (2) and (3), underlying our NB method, will be reasonable in many problems, just as linear regression is a reasonable tool in many problems.

## ACKNOWLEDGEMENTS

This paper reports research undertaken primarily while Nathaniel Schenker was employed by the Statistical Research Division, Bureau of the Census, Washington, DC 20233, USA. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau. This research was supported in part by Joint Statistical Agreements 87-07 and 88-02 between the U.S. Bureau of the Census and Harvard University, and in part by the U.S. National Science Foundation under grant SES-88-05433, and represents a clarification and revision of Rubin, Schafer, and Schenker (1988). The authors wish to thank the two referees for their very helpful comments.

## REFERENCES

BAKER, S.G., and LAIRD, N.M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. Journal of the American Statistical Association, 83, 62-69.
BISHOP, Y.M.M., FIENBERG, S.E., and HOLLAND, P.W. (1975), Discrete Multivariate Analysis, Cambridge: MIT Press.
DEMPSTER, A.P., LAIRD, N.M., and RUBIN, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 39, 1-38.
DIFFENDAL, G. (1988). The 1986 test of adjustment related operations in Central Los Angeles County. Survey Methodology, 14, 71-86.
FAY, R.E. (1986). Causal models for patterns of nonresponse. Journal of the American Statistical Association, 81, 354-365.
FAY, R.E., PASSEL, J.S., and ROBINSON, J.G. (1988). The Coverage of Population in the 1980 Census. 1980 Census of Population and Housing Evaluation and Research Report PHC80-E4, Washington: U.S. Government Printing Office.
FREEDMAN, D.A., and NAVIDI, W.C. (1986). Regression models for adjusting the 1980 Census. Statistical Science, 1, 3-39.
HOGAN, H., and WOLTER, K. (1988). Measuring accuracy in a Post Enumeration Survey. Survey Methodology, 14, 99-116.
Krotki, K.J. (1978). Developments in Dual System Estimation of Population Size and Growth, Edmonton: The University of Alberta Press.
LAIRD, N.M. (1978). Empirical Bayes methods for two-way contingency tables. Biometrika, 65, 1, 581-590.
LEONARD, T. (1975). Bayesian estimation methods for two-way contingency tables. Journal of the Royal Statistical Society, Series B, 37, 23-37.
LITTLE, R.J.A. (1985). Nonresponse adjustments in longitudinal surveys: models for categorical data. Bulletin of the International Statistical Institute, 15, 1-15.
LITTLE, R.J.A. (1986). Missing data in Census Bureau surveys. Proceedings of the Second Annual Research Conference, United States Bureau of the Census, 442-454.
LITTLE, R.J.A., and RUBIN, D.B. (1987). Statistical Analysis with Missing Data, New York: Wiley.
MARKS, E.S., SELTZER, W., and KROTKI, K.J. (1974). Population Growth Estimation. New York: The Population Council.
RUBIN, D.B. (1974). Characterizing the estimation of parameters in incomplete-data problems. Journal of the American Statistical Association, 69, 467-474.
RUBIN, D.B. (1976). Inference and missing data. Biometrika, 3, 581-592.
RUBIN, D.B. (1987). Multiple Imputation for Nonresponse in Surveys, New York: Wiley.

RUBIN, D.B., SCHAFER, J.L., and SCHENKER, N. (1988). Imputation strategies for estimating the undercount. Proceedings of the Fourth Annual Research Conference, United States Bureau of the Census, 151-159.
SCHENKER, N. (1988). Handling missing data in coverage estimation, with application to the 1986 Test of Adjustment Related Operations. Survey Methodology, 14, 87-98.

SCHENKER, N. (1989). The use of imputed probabilities for missing binary data. Proceedings of the Fifth Annual Research Conference, United States Bureau of the Census (forthcoming).

WOLTER, K.M. (1986). Some coverage error models for census data. Journal of the American Statistical Association, 81, 338-346.

# The Sources of Census Undercount: Findings from the 1986 Los Angeles Test Census 

DAVID J. FEIN and KIRSTEN K. WEST ${ }^{1}$


#### Abstract

This paper presents results from a study of the causes of census undercount for a hard-to-enumerate, largely Hispanic urban area. A framework for organizing the causes of undercount is offered, and various hypotheses about these causes are tested. The approach is distinctive for its attempt to quantify the sources of undercount and isolate problems of unique importance by controlling for other problems statistically.


KEY WORDS: Census; Undercount; Coverage improvement; Post enumeration survey.

## 1. INTRODUCTION

In the last decade or two the need to better understand the causes of undercount in the U.S. census has become pressing. As the census has become an increasingly important tool in governing the nation, conducting business, and monitoring social change (Citro and Cohen 1985; Clogg et al. 1986), public concern about the quality of census data has intensified. Much of this concern has arisen because it is perceived, with good foundation, that net census undercount disproportionately affects the economically disadvantaged members of society (Citro and Cohen 1985, ch. 5; Ericksen 1983). Representatives of the disadvantaged believe that as a result their constituents are being denied a fair share of public funds and political representation (Choldin 1987).

Assuming that an acceptable method could be found, one solution to the problem would be to correct the census for the bias due to differential undercount. In the fall of 1987, however, the Department of Commerce decided not to adjust the 1990 census but instead to concentrate on achieving a more complete enumeration (Ortner 1987).

Improving census coverage implies a need to understand the causes of census undercount better than ever before. Many special coverage improvement programs were implemented in the 1980 census, and these may have contributed to the achievement of historically low levels of overall net coverage error. In spite of such efforts, wide socioeconomic coverage differentials have persisted. In response, the Census Bureau has embarked on a broad research program to identify the causes of undercount, concentrating on population subgroups that are especially difficult to enumerate.

This paper presents results from a study of the causes of census undercount in a hard-toenumerate, largely Hispanic area in Los Angeles. The approach is distinctive for its attempt to quantify the sources of undercount and isolate problems of unique importance by controlling for other problems statistically.

Though the putative inequities mentioned above result from net census coverage error (omissions less erroneous enumerations), to keep the analysis manageable only census omissions are investigated here. Omissions in the U.S. census deserve a higher position on the research agenda

[^9]because they are more numerous, vary more systematically with socioeconomic characteristics, and have been more politically controversial than erroneous inclusions.

The paper begins by describing a system for classifying the causes of undercount. Methods and results are presented next. A concluding discussion summarizes the implications for coverage improvement.

## 2. RESEARCH MODEL

The research model is presented in Figure 1. It represents undercount as a problem that occurs primarily at the household, rather than the individual, level. This specification is consistent with the basic sources of undercount in a census based on contacting each household rather than every individual in the population.

Three different household-level undercount problems are distinguished in the top margin of Figure 1: the omission of an entire household due to failure to enumerate a physical housing unit, the omission of an entire household in an enumerated housing unit, and the omission of only some members in a household where others are enumerated. Each of the three undercount problems can originate in census operations, in the society being enumerated, or in an interaction between operational and social system features. The following discussion is restricted to errors associated with the mailout/mailback methods used in the 1986 Los Angeles test census for a largely low income, Hispanic population.

### 2.1 Implementation of Census Operations

Operational difficulties during the census can cause the omission of housing units, of households in enumerated units, and of individuals in enumerated units. Occupied housing units can be missed because they are never added to the address lists or because they are on the lists but are erroneously deleted (U.S. General Accounting Office 1980). Given that a housing unit is correctly listed, all of the persons living in that unit may still be missed by the census due to misclassification of occupied units as vacant during nonresponse followup (U.S. Bureau of the Census 1987b; Ericksen 1983).

For questionnaires which households complete and mail back there are relatively few procedures for detecting missing persons. Procedures aimed at improving within household coverage include a question asking respondents if they were uncertain about including anyone and a clerical consistency check between a roster of household members requested at the beginning of the questionnaire and the number of persons for whom data are provided later on in the form (U.S. Bureau of the Census 1987b; Edson 1987). These procedures "cause"' within household omission if they do not operate as intended due to errors in the administration of edit followup. Similarly, errors by enumerators during mail nonresponse followup may result in failure to add persons who should have been added.

Another important census operation is public information. Census publicity programs are designed to motivate mail response and reduce deliberate concealment by educating people about the uses of census data, the importance of complete reporting, and the confidentiality of census records. The extent to which such programs can reduce within household omission is unknown.

### 2.2 The Social System

At each stage of the census, data collection procedures come into contact with a social system which has many attributes that can impede enumeration. These attributes include unwillingness



Figure 1. Research Model
to report some or all household members, inability to report in a manner consistent with census definitions, and low "social visibility" of household members or the housing units in which they live. (Social visibility is the degree to which household members and housing units possess characteristics which make them perceptible to outsiders.)

The most important social system factors causing housing unit omission are those affecting the social visibility of units. Some kinds of units are easier to find and more likely to appear on commercial address lists that others. Social system sources of omission for households in enumerated units include factors depressing the visibility of household members and refusal to report.

All three broad sets of social system causes are implicated in within household omission: unwillingness to report, definitional problems, and the differential social visibility of household members. Willingness to report can be approached by considering the perceived costs and benefits of reporting for respondents (Dillman 1978). There has been much discussion of the perceived costs of census reporting. People may fear that disclosure of adult males will jeopardize welfare eligibility, that persons illegally in the country will be deported, that reporting more persons than allowed by a lease will prompt landlord troubles, and that police will be informed of the whereabouts of lawbreakers (Bailar and Martin 1987). Such fears may cause noncompliance when there is disbelief in the Census Bureau's promise of confidentiality.

The sources of definitional error are quite different from those of concealment. Definitional errors arise in the complexities of household living arrangements, as conditioned by respondents' abilities to understand and apply census enumeration and residence rules (Hainer et al. 1988).

Having mentioned some of the major sources of undercount, we will now examine the extent to which they occurred during the 1986 Los Angeles test census.

## 3. METHODS

### 3.1 Data Sources

This study takes an intensive look at undercount in a March 1986 test census conducted in the northern half of Los Angeles County. The population was low income and largely Hispanic. Nearly two-thirds ( $65 \%$ ) of the heads of households enumerated in the census were of Spanish origin and $13 \%$ were Asian. Residences in this part of Los Angeles were largely single family dwellings ( $73 \%$ ) and small apartment buildings ( $15 \%$ ). Owners lived in half ( $51 \%$ ) of the occupied units, in contrast with nearly two thirds ( $65 \%$ ) of all occupied units nationwide (U.S. Bureau of the Census 1987a: 106, table 18; U.S. Bureau of the Census 1987c: 712, table 1285).

The data analyzed are from the 1986 Los Angeles test census itself; the Post Enumeration Survey, or PES, conducted to measure test census coverage; and a special followup to the PESthe Causes of Undercount Survey. The census enumerated 109,900 housing units and was intended primarily as a test of planned 1990 census operations.

The Post Enumeration Survey (PES) was one of these operations. The purpose of the PES, conducted in July 1986, was to identify census omissions and erroneous enumerations (Diffendal 1988). It did this by attempting to match PES to census records. When a PES person's record was found in the census it was termed 'matched"; otherwise the person was considered "nonmatched".

Three kinds of PES households are distinguished here, depending on whether all, some, or none of their members were matched to the census. "Complete match" households contain only persons in the PES who were matched to persons in the census. "Partial nonmatch"
households contain at least one person who could not be matched and at least one person who was matched to the census. "Total nonmatch" households include only persons who could not be matched to the census.

These three household types are distinguished to allow examination of problems associated with housing unit omission, omission of entire households in enumerated units, and omission of persons from households that were partially enumerated. Completely matched households are included for reference purposes, to represent households correctly enumerated in the census.

A special followup survey - the Causes of Undercount Survey - was conducted in November 1987 to obtain additional information needed to compare these household types. The survey obtained information on census characteristics for nonmatched persons, as well as some new household and housing unit data not available on the census or PES files.

The entire partial nonmatch stratum and nearly all households in the total nonmatch stratum were selected for reinterview. Eight total nonmatch households had to be omitted because several items needed to reinterview them were missing. Households in the complete match stratum were subsampled to reduce survey costs.

The distribution of the 966 completed Causes of Undercount Survey interviews by household type is shown in the right-most column of Table 1. This table also gives the unweighted numbers for all 5814 PES households and the 1420 cases in the Causes of Undercount Survey sample. The overall response rate for the survey was $68 \%$, reflecting considerable success in locating households in a transitory urban area despite the 16 months intervening between the survey and the PES.

### 3.2 Analysis Plan

There are several parts to the analysis. PES total nonmatch households are examined first. Two sets of comparisons are made: 1) of missed housing units with enumerated housing units and 2) of missed households in enumerated units with enumerated households. Missed housing units were expected to contain a higher percentage of clustered housing units and unusual unit types and locations than enumerated units. Missed households in enumerated housing units were expected to be smaller, contain adults who were less frequently at home, and move more often than enumerated households. Most of the explanatory variables for housing unit and household omission were obtained either from the census Address Control File or from the PES matched file, and thus are available for all 193 total nonmatch households in the sample.

## Table 1

Numbers of Households in the PES and Causes of Undercount Survey Sample, and Numbers of Completed Interviews, by Household Type.

| Household Type | Post <br> Enumeration <br> Survey | Sample | Causes of Undercount Survey <br> Completed <br> Interviews |
| :--- | :---: | :---: | :---: |
| Complete Match | 4,871 | 489 | 382 |
| Partial Nonmatch | 738 | 738 | 484 |
| Total Nonmatch | 205 | 193 | 100 |
| All Types | 5,814 | 1,420 | 966 |

The second part of the analysis compares partial nonmatch with complete match households to identify factors responsible for within-household omission. Two sets of explanatory factors are distinguished, those indicating inadvertent or "definitional" errors and those representing reasons for deliberate concealment. Indicators for definitional errors include large size and complex composition of households, poorly-spoken English and educational deficits. Concealment indicators include presence of recent immigrants, welfare recipiency, crowded housing, and disbelief in census confidentiality. It was hypothesized that partial nonmatch households would score higher on the definitional and concealment indicators than would complete match households.

The analysis begins with bivariate relationships between each of the explanatory factors and partial omission and then considers multivariate relationships. The source for many of these indicators was the Causes of Undercount Survey; hence, only data from interviewed households are used.

In the final part of the analysis, characteristics of four types of individuals are compared: persons matched in complete match and partial nonmatch households, and those nonmatched in partial and total nonmatch households. Characteristics compared include age, sex, education, relationship to the household head, and citizenship status.

Bivariate percentages are based on weighted data to compensate for the PES and Causes of Undercount Survey sampling designs, though tests for differences between these percentages used unweighted numbers. Unweighted data were used to estimate parameters of loglinear models. The effects of the PES sampling design on estimates for the final models were evaluated by adding in all two-way interactions which included the PES stratification variable. This adjustment did not greatly change the results; thus, the estimates presented here do not include the stratification variable. Because the second stage of PES sampling entailed cluster sampling of households in census blocks, the standard errors calculated are likely to underestimate the true sampling errors: they are presented only as rough guides to the significance of parameters.

## 4. FINDINGS

### 4.1 Total Nonmatch Households

Table 2 shows the final status assigned in the census to PES total nonmatch households for cases sent and not sent to nonresponse followup. Of the 193 total nonmatch cases 97 , or $50 \%$, never appeared on the census address lists. Thus, housing unit omission appears to explain why the PES could not find anyone in these households in the census.

The remaining 96 cases did appear on the census address lists. What caused these households to be missed? The explanation is probably that most of these units were census closeout interviews, where a landlord or neighbor provided only an estimate of the total number of persons in the household and not detailed information for individuals. This hunch is supported by the finding that of the 44 cases the census classified as occupied, population counts for 37 were "goldplated". This means that the final count accepted for these households was not obtained in the usual manner by allowing the FOSDIC (Film Optical Device for Input to Computers) machines to count persons. Instead, goldplating involved accepting a total count for the household entered on the questionnaire in the field. This is likely an indication that the household was a closeout case.

Thus, the census really did not miss most of these 44 households entirely, though when it came time for PES matching, there were no individual census person records to be matched.

Table 2
Final Status Assigned in the Census to PES Total Nonmatch Households By Nonresponse Followup Status: Numbers of Units ${ }^{\text {a }}$

| Final Status of Unit in Census | Sent to Nonresponse Followup? |  |  |
| :--- | ---: | ---: | ---: |
|  | No | Yes | Total |
| Omitted from the Census Address Lists | 97 | 0 | 97 |
| Included in the Census Address Lists | 4 | 92 | 96 |
| Occupied, Direct Accept ${ }^{\text {b }}$ | 1 | 6 | 7 |
| Occupied, Gold-plated | 2 | 35 | 37 |
| Vacant, Direct Accept | 1 | 34 | 35 |
| Vacant, Gold-plated | 0 | 17 | 17 |
| All Units | 101 | 92 | 193 |

Notes: a N's are unweighted.
${ }^{b}$ Direct Accept: FOSDIC person count accepted.
Gold-plated: Field counts accepted instead of FOSDIC.
An allowance is made for these cases in the dual system estimation method. Nevertheless, it still is true that these households were not directly enumerated.

To summarize, $50 \%$ of the PES total nonmatch households were in units which appeared to have been entirely omitted. Of the households living in units which were enumerated, $54 \%$ had been classified as vacant, possibly erroneously, and $46 \%$ had been found to be occupied. Of the total nonmatch households classified as occupied in the census, up to $84 \%$ may have been enumerated in closeout interviews.

Figure 2 compares some physical characteristics of units left off the census address lists (light bars) with units that were not left off the lists (dark bars). The top set of bars represents the basic types of housing units. Attached single family homes, such as duplexes, appear to have been a major problem in the L.A. test census. Thirty-four percent (34\%) of the missed units fell into this category, in contrast to only $8 \%$ of enumerated units. Missed units were less likely than enumerated units to be detached single family homes or apartments in large buildings, suggesting that the census was more successful at finding such units.

Whether or not an interview was completed, Causes of Undercount Survey interviewers were asked to record when units they visited fit any of several " unusual unit'" categories listed on the front of their questionnaires. The bottom half of Figure 2 shows that the interviewers identified a higher percentage of unusual units among units that were missing from the census address lists, $28 \%$, than among units that were included, $7 \%$. Unit types found to be particular problems were abandoned-looking buildings and secondary units on a lot.

Physical characteristics of units thus do appear to affect their visibility during census address list development. What might cause households to be missed in units that were enumerated?

Households may be more easily missed if they are small and mobile. Figure 3 compares characteristics of total nonmatch households in enumerated units with a combined group of complete match and partial nonmatch households - that is, households which were enumerated. Households missed in the test census (light bars) were on average considerably smaller than those where some or all members were counted (dark bars). Whereas $53 \%$ of the total nonmatch households in enumerated units had one or two members, only $35 \%$ of the enumerated households were this small.


Figure 2. Physical Characteristics of Enumerated and Missed Housing Únits (Weighted Percentages)


Figure 3. Characteristics of Enumerated Households and Total Nonmatch Households in Enumerated Units (Weighted Percentages)

Indicators of the propensity to move include home ownership and actual household mobility in the four months between the census and the PES. Households missed in the census were more likely to be renters and movers ( $61 \%$ and $8 \%$, respectively) than were enumerated households ( $46 \%$ and $0 \%$, respectively). The percentage of households in which all adults were employed full-time in March 1986 was greater by $12 \%$ for omitted households than for enumerated households, though the number of interviews for omitted households was too small for this difference to be statistically significant.

These results support the hypothesis that missed housing units and households missed in enumerated units possess attributes which reduce their visibility during a census.

### 4.2 Partial Nonmatch Households

From total nonmatch households, the focus shifts to the factors associated with partial household omission. In this phase of the analysis, 484 partial nonmatch households were compared with 331 complete match households. Single person households were excluded from the 382 complete match households in the Causes of Undercount Survey sample, since they were not at risk of partial omission.

Two different sets of explanatory factors were considered. The first represents household characteristics thought to be associated with definitional errors, described earlier as errors resulting from inconsistencies between household membership as understood by the Census Bureau and by census respondents. The second set of indicators represents factors thought to be associated with the deliberate concealment of household members.


Figure 4. Definitional Error Indicators for Partial Household Omission: Households with $2+$ Persons (Weighted Percentages)

## Definitional Errors

Indicators for definitional errors include household size and composition, English language ability, census respondent's education, and edit followup status. Larger households, those containing more distant relatives and persons unrelated to the household head, those speaking a language other than English at home, those where the census respondent's education was low, and households not sent to edit followup were all expected to be at greater risk of definitional errors.

Figure 4 supports these hypotheses. It shows that partial nonmatch households (light bars) were considerably larger than complete match households (dark bars): $45 \%$ of the partial nonmatch households but only $19 \%$ of the complete match households contained six or more members. Whereas $40 \%$ of the partial nonmatch households contained only nuclear relatives of the household head, fully $72 \%$ of the complete match households were nuclear. Partial nonmatch households were less likely to have been sent to edit follow-up by a slight, but statistically significant, amount. Partial nonmatch households were more likely to speak a language other than English at home ( $83 \%$ ) than were complete match households ( $64 \%$ ). Finally, census respondents from partial nonmatch households had less formal education than those from complete match households: $36 \%$ of the census respondents from partial nonmatch households had not attended high school, in contrast with $24 \%$ of the respondents from complete match households.

Log-linear models were fitted to see whether these differences persisted at the multivariate level. The dependent variable in these models was partial household omission, with complete match households coded as 0 and partial nonmatch households coded as 1. Interactions between partial omission and each of the independent variables in Figure 4 were tested in a series of nested models. All two-way interactions among independent variables were included in each model as controls.

In the multivariate analysis, significant interactions with partial omission were found for all definitional error indicators except census respondent's education. Table 3 presents the chi square (Wald) statistics associated with the final definitional model, which excludes census respondent's education. Significant interactions of household size with composition and language other than English were also detected. Parameter estimates in Table 4 show the effects to be in the directions expected. Estimates for standardized parameters, obtained by dividing

Table 3
Chi Square Statistics For Testing Two-Way Interactions in the Final Definitional Error Modela ${ }^{\text {a }}$

|  | Interactions with . . |  |  |  |
| :--- | :---: | :---: | :---: | :---: |
| Variables | Size | Composition | Edit <br> Followup | Language <br> at Home |
| Partial Omission | $38.1^{* *}$ | $42.3^{* *}$ | $6.3^{*}$ | $5.2^{*}$ |
| Size | - | $112.0^{* *}$ | .9 | $50.0^{* *}$ |
| Composition | - | - | 1.6 | $1.3^{3}$ |
| Edit Followup | - | - | - | 1.0 |

[^10]Table 4
Parameter Estimates for Interactions Between Definitional Error Indicators and Partial Household Omission in the Final Model

| Marginals with Partial | Parameter <br> Estimate | Standard <br> Error | Standardized <br> Poramatch Household and . . |
| :--- | :---: | :---: | :---: |
| Household Size: | -.34 | .06 | -5.7 |
| $\quad$ 2-3 Persons |  |  |  |
| 4-5 Persons |  |  |  |
| Composition: | -.02 | .05 | -.4 |
| $\quad$ All nuclear |  |  |  |
| $\quad$ All non-nuclear |  |  |  |
| Edit Followup Status |  |  |  |
| $\quad$Not sent | -.36 | .06 | -6.0 |
| Other Language at Home? <br> Yes | .22 | .09 | 2.4 |

parameter estimates by their standard errors, indicate that the effects of size and composition are about the same in magnitude and that both are larger than the effects of edit followup and language spoken at home.

## Concealment Indicators

Factors hypothesized to cause concealment of household members by census respondents include: fear that persons illegally in the country would be deported, fear that disclosure of adult males would jeopardize welfare aid, and concern that reporting more persons than allowed by a lease would bring landlord troubles. Indicators for these factors were, respectively, whether the household contained recent immigrants, defined as persons entering the country in or after 1980; whether anyone in the household was receiving welfare during the census month; and the average number of persons per room in the household. Nonresponse to the census mailout was also included as a general indicator of failure to perceive positive benefits from responding to the census. Finally, belief in census confidentiality was included to see whether it helped to reduce fears resulting in concealment.

Figure 5 shows that all of these indicators were related to partial omission at the bivariate level. For example, recent immigrants were present in $26 \%$ of the partial nonmatch households (light bars), but only $12 \%$ of the complete match households (dark bars). Whereas $24 \%$ of the partial nonmatch households reported receiving welfare, only $15 \%$ of the complete match households did so. Partial nonmatch households were considerably more likely to exhibit crowding: $63 \%$ contained more than one person per room, in contrast to only $34 \%$ of the complete match households. Partial nonmatch households were also somewhat less likely than complete match households to have returned their census questionnaires by mail or to believe in census confidentiality.

Again, loglinear models were fitted, with partial omission as the dependent variable and the concealment indicators as independent variables. All two-way interactions with household size were included as controls, since other things being equal, larger households would be more likely to exhibit crowding and contain recent immigrants than small ones.

This time, two variables did not survive preliminary testing: mail nonresponse and belief in census confidentiality. Before completely dropping the confidentiality variable, tests were performed to see if interactions of partial omission with presence of immigrants, welfare recipiency, and crowding depended on belief or disbelief in confidentiality. Belief in confidentiality was not found to affect these relationships.


Figure 5. Concealment Indicators for Partial Household Omission: Households with $2+$ Persons (Weighted Percentages)

Table 5
Chi Square Statistics For Testing Two-Way Interactions in the Final Concealment Modela

|  | Interactions with . . |  |  |  |
| :--- | :---: | :---: | :---: | :---: |
| Variables | Size | Immigrants | Welfare <br> Assistance | Crowding |
| Partial Omission | 2.9 | $11.3^{* *}$ | $10.1^{* *}$ | $16.7^{* *}$ |
| Size | - | .2 | $7.5^{*}$ | $221.7^{* *}$ |
| Recent Immigrants | - | - | 1.6 | $30.0^{* *}$ |
| Welfare Assistance | - | - | - | 5.4 |

[^11]Table 5 shows that three of the remaining concealment variables immigrants, welfare, and crowding interacted significantly with partial household omission in a model which included all two-way interactions with size and all two-way interactions among independent variables. Standardized parameter estimates (see Table 6) suggest effects of roughly equal magnitude for the three indicators.

It is noteworthy that the relationship between partial omission and size vanished when crowding was included (see Table 5), suggesting that the effects of size were due to its association with crowding rather than scale alone. Crowding was also strongly associated with the presence of recent immigrants.

### 4.2 Person Characteristics

For the final part of the analysis of individual-level characteristics associated with undercount, four kinds of persons were compared: persons the census counted in complete match and partial nonmatch households, and persons the census missed in partial and total nonmatch households.

Figure 6 shows differences between the percentages in 10 year age groups for persons in complete match households and each of the three other groups. It shows an excess in the 20-29 year old group for persons missed in partial and total nonmatch households relative to persons in complete match households. There is also evidence of an excess in the 20-29 year age groups for persons who were enumerated in partial nonmatch households.


Figure 6. Excess Weighted Percentage in Age Group Relative to Persons in Complete Match Households

Table 6
Parameter Estimates for Interactions Between Concealment Indicators and Partial Household Omission in the Final Concealment Model

| Marginals with Partial <br> Nonmatch Household and . . . | Parameter <br> Estimate | Standard <br> Error | Standardized <br> Parameter Estimate |
| :--- | :---: | :---: | :---: |
| Recent Immigrants: <br> Immigrants Present | .19 | .06 | 3.2 |
| Welfare Recipiency: <br> Receiving Aid | .17 | .05 |  |
| Crowding: | -.49 |  | 3.4 |
| < Persons/Room | -.01 | .13 | -3.8 |
| $.5-1.0$ Persons/Room | .08 | .08 | -.1 |
| $1.0-1.5$ Persons/Room |  | 1.0 |  |

Table 7
Percentage Distributions for Characteristics of Individuals by PES Match Status and Household Type

| Characteristic | PES Match Status |  |  |  |
| :---: | :---: | :---: | :---: | :---: |
|  | Matched in |  | Nonmatched in |  |
|  | Complete Match HHs | Partial Nonmatch HHs | Partial Nonmatch HHs | Total Nonmatch HHs |
| Sex |  |  |  |  |
| Male | 46.2\% | 50.6\% | 54.2\% | 48.2\% |
| Female | 53.8 | 49.4 | 45.9 | 51.8 |
| Unweighted $n$ | 1667 | 2564 | 1324 | 582 |
| Education |  |  |  |  |
| No Formal Education | 10.2 | 10.9 | 17.0 | 14.3 |
| Less than High School | 30.7 | 34.4 | 27.2 | 37.5 |
| Some High School | 20.5 | 20.6 | 19.5 | 19.5 |
| High School Graduate | 38.6 | 34.1 | 36.4 | 28.8 |
| Unweighted $n$ | 1197 | 1560 | 599 | 315 |
| Relationship to Head |  |  |  |  |
| Nuclear Relative | 86.1 | 83.2 | 63.6 | 85.9 |
| Non-nuclear Relative | 11.3 | 12.6 | 25.4 | 7.9 |
| Non-relative | 2.6 | 4.2 | 11.0 | 7.0 |
| Unweighted $n$ | 1659 | 2560 | 1359 | 590 |
| Citizenship |  |  |  |  |
| Citizen Since Birth | 66.2 | 53.5 | 52.6 | 50.4 |
| Naturalized Citizen | 9.2 | 9.5 | 6.4 | 6.4 |
| Noncitizen | 24.6 | 37.0 | 41.0 | 43.2 |
| Unweighted $n$ | 1223 | 1567 | 612 | 316 |

Persons missed by the census in partial nonmatch households were slightly more likely than persons in complete match households to be males and have no formal schooling, and less likely to be citizens or close relatives of the household head (Table 7). Persons missed by the census in total nonmatch households were also slightly more likely to be noncitizens and lower in education than persons in complete match households, but displayed no differences in sex and relationship to household head. Thus, on the whole, persons missed in partial nonmatch households differed from those in complete match households in more ways than did persons missed in total nonmatch households.

In addition to biasing more census characteristics, partial household omission caused the omission of many more persons than did total household omission. Two thirds (67\%) of all PES nonmatch cases were in partial nonmatch households and only one third were in total nonmatch households. Fully $82 \%$ of all PES omissions were found in housing units the census enumerated and only $18 \%$ were in missed units.

## 5. DISCUSSION

The findings reported here support evidence from more qualitative studies that partial household omission is the most serious undercount problem in hard-to-enumerate urban areas of the United States today. As compared with total household omission, partial omission in the Los Angeles test census accounted for twice as many missing persons, reflected more intractable sources of error, and biased more individual-level census characteristics.

The chief problems identified for total household omission were failure to include certain types of housing units in the census address lists and misclassifying occupied units as vacant. Housing units especially at risk of misclassification as vacant were those with households which were small and mobile and those in which all adults were working full-time. Experience with coverage improvement programs at the Census Bureau suggests that further reductions in housing unit omission may be possible. Such programs were responsible for adding about $10 \%$ of the units enumerated in Los Angeles. The Bureau adopted special precanvassing procedures in the test census to find units in large multi-unit structures. Considerable success in reducing this source of error in the test census is evident in Figure 2: none of the apartment units missed were in large buildings.

The misclassification of occupied units as vacant will be more difficult to remedy. Allowing nonresponse enumerators more time per unit and improved training for certain kinds of problem households may help somewhat. Coupling these efforts with special callback procedures for smaller and more transient households and those whose members are rarely at home would also help.

It is clear that improvements at the margin of what is already a largely successful census operation will be expensive. Keyfitz (1979) and others have observed that the incremental costs from adding persons to the count soar as coverage approaches $100 \%$. Programmatic innovations to reduce the errors observed in the 1986 test census would add to the $\$ 2.6$ billion cost projected for the 1990 census, since the methodology to be used in urban areas will be very similar to the L.A. test census.

Within-household errors will be even more difficult to address than total household omissions. The Bureau must redouble its efforts to understand the complex living arrangements and cognitive and/or cultural factors that condition how people perceive household membership. The findings reported here suggest that further efforts targeted to respondents for whom English is not a native tongue, and households containing persons only distantly related to each other may help to reduce definitional errors.

However, in light of the considerable research already performed to improve the design of the census questionnaire and the complex enumeration and residence rules to which the Bureau is bound by statute and tradition, further reductions in definitional error will require extraordinary efforts. Definitional errors are deeply embedded in cultural differences and educational deficits among hard-to-enumerate groups.

Within-household omission also was found to be strongly related to the presence of immigrants, welfare recipiency, and crowding. That a PES-based study could detect such effects suggests that the PES succeeded in counting many persons whose presence had been concealed in the census. Some of the effects of the so-called concealment variables may be due to uncontrolled factors other than concealment, but the persistence of relationships even after household composition was added in a final log-linear model (not shown) suggests that the PES really did detect some persons who were concealed in the census. Thus, there appears to be a continuum from households that are highly resistant to enumeration to those which are less resistant, and for the latter more intensive methods like those used in the PES may be effective.

The social conditions underlying the most resistant forms of concealment present the most difficult problems for the Census Bureau. Public information programs attempting to convince people that the census is important and that census data will be kept confidential were not very effective for the hard- to-enumerate population in the Los Angeles test census, as reported by Moore and McDonald (1987), though these programs may work better under real decennial census conditions. The minimal role found for belief in census confidentiality, either in its own right or in mediating between household circumstances and concealment, suggests that the relationship between attitudes and census response behavior is not a simple one.

The findings reported here should not be generalized uncritically to the sources of undercount expected to affect urban areas in the 1990 Census. Because the data are based on a test census, errors may reflect inexperience with experimental procedures or failure to convince respondents (and census workers) that the project was as serious as the decennial census. Further, to the degree that Los Angeles is unlike other major urban areas, it may experience unique census-taking problems. For example, Los Angeles is thought to be home to more i!legal aliens than any other major city (Heer and Passel 1987).

On the other hand, the net undercount rate for Los Angeles in 1980 was quite similar to the rates for other major cities, as measured in the 1980 Post Enumeration Program (Fay et al. 1988). Thus, what they lack in illegal aliens, these cities may make up in other hard-toenumerate groups. Further research is needed to assess the degree to which causes of undercount differ by race, ethnicity, and other social characteristics.

It is encouraging that the causes of undercount identified in this Post Enumeration Surveybased study were reasonably consistent with more qualitative reports by ethnographers and focus groups. Also, the PES estimates for undercount from the Los Angeles test census are believed to be of high quality (Hogan and Wolter 1988). For these reasons, extension of the PES-based methodology developed in this paper to other urban (and nonurban) areas is recommended.

On the social system side, further research on how rationally people weigh the costs and benefits of responding to censuses and surveys would help to weigh the potential for improving census coverage through the Census Bureau's public information and community action programs. Better indicators for household-level reasons for concealment are also needed. Examining specific assistance programs would help to confirm the effects of welfare participation on census coverage, since not all aid would be imperiled by revealing true householdcomposition.

Improved measurement of the sources of undercount arising in census operations is also needed. If data from census quality control programs were combined with PES matching results, error sources could be identified with greater precision.

## ACKNOWLEDGEMENTS

We would like to thank the following persons for their assistance in this research: Irwin Anolik, Miriam Balutis, Gregg Diffendal, Chris Dyke, Sue Finnegan, Howard Hogan, Jan Jaworski, Pete Long, and Lynn Weidman. Betsy Martin, Jim O'Brien, and two anonymous reviewers provided valuable comments on earlier versions of this paper.

## REFERENCES

BAILAR, B., and MARTIN, E. (1987). Report on Meetings in Los Angeles, Chicago and Denver. Unpublished Census Bureau memorandum.

CHOLDIN, H. (1987). Science and Scientists in the 1980 Census Lawsuits. Paper presented at the May 1987 meeting of the Population Association of America, Chicago.
CLOGG, C.C., MASSAGLI, M.P., and ELIASON, S.R. (1986). Population undercount as an issue in social research, Proceedings of the Second Annual Research Conference. United States Bureau of the Census, Washington, D.C., 335-343.
CITRO, C.F., and COHEN, M.L. (eds.) (1985). The Bicentennial Census: New Directions for Methodology in 1990, Panel on Decennial Census Methodology, National Research Council, Washington, D.C.: National Academy Press.

DIFFENDAL, G. (1988). The 1986 test of adjustment related operations in Central Los Angeles County. Survey Methodology 14, 71-86.
DILLMAN, D.A. (1978). Mail and Telephone Surveys: The Total Design Method. New York: John Wiley and Sons.

EDSON, R.G. (1987). Preliminary coverage improvement results from tests for the 1990 Census. Paper presented at the August 1987 meeting of the American Statistical Association, San Francisco.
ERICKSEN, E.P. (1983). Affidavit, Mario Cuomo, et al. vs. Malcolm Baldridge et al., U.S. District Court, Southern District of New York, 80 Civ. 4550 (JES).

FAY, R.E., PASSEL, J.S., and ROBINSON, J.G. (1988). The coverage of population in the 1980 Census. Evaluation and Research Reports. 1980 Census of Population and Housing PHC80E4, Washington, D.C.

HAINER, P., HINES, C., MARTIN, E., and SHAPIRO G.M. (1988). Research on improving coverage in household surveys. Proceedings of the Fourth Annual Research Conference. United States Bureau of the Census, Washington, D.C., 513-539.
HEER, D.M., and PASSEL, J.S. (1987). Comparison of two methods for estimating the number of undocumented Mexican adults in Los Angeles County. International Migration Review, 21(4), 1446-1473.

HOGAN, H., and WOLTER, K. (1988). Measuring accuracy in a Post-Enumeration Survey. Survey Methodology 14, 99-116.
KEYFITZ, N. (1979). Information and allocation: Two uses of the 1980 Census. The American Statistician, 33(2), 45-56.
MOORE, J.C., and McDONALD, S.-K. (1987). The Census community awareness program: an evaluation of the potential and actual effectiveness of CCAP based on evidence from the 1986 Los Angeles Census Test. Unpublished Census Bureau report.

ORTNER, R. (1987). Statement. United States Department of Commerce News, October 30, 1987.
U.S. BUREAU OF THE CENSUS (1960). The Post-Enumeration Survey: 1950. Technical Paper No. 4, Washington, D.C.
U.S. BUREAU OF THE CENSUS (1987a). 1986 Test Census, Central Los Angeles County, California. General Population and Housing Statistics, TC86-1, Washington, D.C.
U.S. BUREAU OF THE CENSUS (1987b). Programs to improve coverage in the 1980 Census. Evaluation and Research Reports. 1980 Census of Population and Housing, PHC80-E3, Washington, D.C.
U.S. BUREAU OF THE CENSUS (1987c). Statistical Abstract of the United States: 1987, (106th edition), Washington, D.C.
U.S. GENERAL ACCOUNTING OFFICE (1980). Problems in Developing the 1980 Census Mail List. Washington, D.C.: General Accounting Office.

# Total Error in the Dual System Estimator: The 1986 Census of Central Los Angeles County 

MARY H. MULRY and BRUCE D. SPENCER ${ }^{1}$


#### Abstract

The U.S. Bureau of the Census uses dual system estimates (DSEs) for measuring census coverage error. The dual system estimate uses data from the original enumeration and a Post Enumeration Survey. In measuring the accuracy of the DSE, it is important to know that the DSE is subject to several components of nonsampling error, as well as sampling error. This paper gives models of the total error and the components of error in the dual system estimates. The models relate observed indicators of data quality, such as a matching error rate, to the first two moments of the components of error. The propagation of error in the DSE is studied and its bias and variance are assessed. The methodology is applied to the 1986 Census of Central Los Angeles County in the Census Bureau's Test of Adjustment Related Operations. The methodology also will be useful to assess error in the DSE for the 1990 census as well as other applications.


KEY WORDS: Nonsampling error; Post enumeration survey; Coverage evaluation, Undercount; Capture-Recapture.

## 1. INTRODUCTION

The dual system estimator (DSE) is used in several contexts for estimating the size of a population. Its applications range from wildlife populations to human populations. DSEs of births are used at the U.S. Bureau of the Census in the formation of the demographic analysis estimates of the national population. Currently, the Census Bureau intends to use DSEs for measuring coverage error in the 1990 Decennial Census. This paper focuses on the application of the DSE in the census context where the two systems are the original enumeration and a Post Enumeration Survey (PES).

The obvious estimator based on the DSE of census undercoverage is $\widehat{U C}$, given by $\widehat{U C}=$ DSE - CEN, with CEN referring to the size of the original census enumeration. Since $\mathrm{DSE}=\mathrm{CEN}+\widehat{U C}$, the DSEs also provide alternative estimates of population. A more general class of alternative estimates based on the DSE (Spencer 1980; 1986) is $(1-f) \times \mathrm{CEN}+f \times \mathrm{DSE}$, or equivalently

$$
\mathrm{CEN}+f \times \widehat{U C}
$$

with $0 \leq f \leq 1$.

Estimates of total error of the DSE are essential for determining what value of $f$ leads to the most accurate estimator of population size. Since the range of values for $f$ include 0 and 1 , the selection of either CEN or DSE is possible. The criteria for improvement of one set of population estimates over another may be based on measures of the quality of the distribution of the population (Hogan and Mulry 1987; Spencer 1986). Estimates of total error in the

[^12]DSE are also important for statistical planning purposes, e.g., how much money should be spent and how big a sample should be fielded in the PES.

DSEs are subject to several components of nonsampling error, in addition to sampling error. We present models of the total error and the components of error in the DSE. The models relate observed indicators of data quality to the first two moments of the components of the error. We then use techniques of propagation of error to estimate the bias and variance of the DSE. In doing so, we assess the total error, or the joint effect of the errors. Previous work on error models for the DSE includes Seltzer and Adlakha (1974).

The methodology is applied to the 1986 Census of Central Los Angeles County, also known as the 1986 Test of Adjustment Related Operations (TARO) conducted in Los Angeles (Diffendal 1988). The PES in TARO comprised about 6,000 housing units and over 19,000 people. A sensitivity analysis shows how the component errors interact, which ones cancel, and which ones compound each other. The methods described here to estimate the error in the TARO DSE can be extended to estimate the error in the 1990 DSEs.

We have tried to organize this paper to facilitate incomplete reading of the paper. Section 2 introduces and presents the rationale for the TARO DSE and its major components. Our strategy for assessing the component errors and combining them to estimate the total error in the DSE is described next (Section 3). A detailed description of the DSE, with notation, is necessary for precise description of the component errors (Section 4). Following that description is an assessment of the component errors (Section 5). A synthesis of the component errors leads to estimates of the total error of the DSE (Section 6). Our major conclusions are then presented (Section 7).

## 2. DUAL SYSTEM ESTIMATOR

The application of the dual system estimator requires assuming that there are two lists of the population. The first list is the original census enumeration, and the second is an implicit list of those covered by the sampling frame for the $P$ sample of the PES, whom we will call the P -sample population. The sampling frame itself is not a list of people, but of census blocks.

The $\mathbf{P}$ sample is one of the two samples that comprise the PES. The PES is composed of the $E$ sample, which is a sample of census enumerations, and the $P$ sample, which is a sample of the population. The $E$ sample is selected to estimate the number of enumerations that are erroneous. The $P$ sample is selected to estimate, through dual system estimation, the number of people missed by the original enumeration.

Table 1
Probabilities of Inclusion in a Cell

|  | Original Enumeration |  |  |
| :---: | :---: | :---: | :---: |
|  | In | Out | Total |
|  |  |  |  |
| P sample In | $p_{i 11}$ | $p_{i 22}$ | $p_{i 1+}$ |
| Out | $p_{i 21}$ | $p_{i 22}$ | $p_{i 2+}$ |
| Total | $p_{i+1}$ | $p_{i+2}$ | $p_{i++}$ |

Table 2
True Population Size in Each Cell

|  | Original Enumeration |  |  |
| :---: | :---: | :---: | :---: |
|  | In | Out | Total |
| P sample In | $N_{11}$ | $N_{12}$ | $N_{1+}$ |
| Out | $N_{21}$ | $\left(N_{22}\right)$ | $\left(N_{2+}\right)$ |
| Total | $N_{+1}$ | $\left(N_{+2}\right)$ | $\left(N_{++}\right)$ |

The dual system estimator is based on a model that the probabilities that the $i$-th individual in the population of size $N$ is in the census or not and in the $\mathbf{P}$ sample or not are as shown in Table 1 (Wolter 1986a); see Wolter (1986a) for discussion and references to earlier work. The true population size in each category is defined in Table 2.

In Table 2, $N_{++}=N$, the total population size. Even if we could observe the $N_{i j}$ 's in the first row and first column, the $N_{i j}$ 's in parentheses would not be observed directly, but would have to be estimated from the model. The DSE of $N$ then would have the form $N_{1+} N_{+1} / N_{11}$, which we will refer to as the ideal DSE.

In estimating population size for measuring census coverage error, the $N$ 's are replaced by estimates from the original enumeration and two sample surveys, the P sample and the E sample. The survey data are weighted by the reciprocals of the selection probabilities. In the following definitions, the estimates with " "" reflect the possible presence of nonsampling error:
$N_{p}=$ the weighted number of P-sample selections
$\hat{N}_{p}=$ the estimate of the total population from the P sample.
CEN $=$ the size of the original enumeration
$I_{1} \quad=$ the number of persons imputed
$I_{2}=$ the weighted number of census enumerations with insufficient information for matching
$E E=$ the weighted number of erroneous enumerations in the original enumeration, based on the E sample
$\widehat{E E}=$ the estimate of the number of erroneous enumerations in the original enumeration
$C=\mathrm{CEN}-I I_{1}-I I_{2}-E E=$ the weighted number of distinct people in the original enumeration from the $E$ sample,
$\hat{C}=\mathrm{CEN}-I_{1}-I_{2}-\widehat{E E}=$ the estimate of the number of distinct people in the original enumeration from the E sample,
$M=$ the weighted number of people in the census and the $P$ sample
$\hat{M}=$ the estimate of the number of people in the census and the P sample.
With this notation, $\hat{N}_{p}$ estimates $N_{p}$, which unbiasedly estimates $N_{1+}$. The ratio $\hat{C} / \hat{M}$ is used to estimate the ratio $N_{+1} / N_{11}$. (By themselves, $\hat{C}$ and $\hat{M}$ are not good estimators of $N_{+1}$ and $N_{11}$.) Thus, the estimator has the form $\hat{N}_{++}=\hat{N_{p}} \hat{C} / \hat{M}$. The ratio $\hat{C} / \hat{M}$ contains a correction for erroneous enumerations and for cases with insufficient information for matching, $I_{1}$ and $I_{2}$, so that cases with no chance of being included in the denominator are also excluded from the numerator.

The DSE is used to estimate the percent net undercount, or the net undercount rate, in the original enumeration,

$$
\hat{U}=100\left(\mathrm{CEN}-\hat{N}_{++}\right) / \hat{N}_{++} .
$$

For the TARO site (i.e. Central Los Angeles County) as a whole, CEN $=355,352$, $\hat{N}_{p}=336,707, \hat{C}=343,567, \hat{M}=298,204$, and $\hat{N}_{++}=388,040$. Using these numbers, the estimate of the net undercount rate is 8.42 .

## 3. STRATEGY FOR ASSESSING TOTAL ERROR

The DSE is subject to various sources of error, including error due to incorrect addresses from the $P$ sample, error due to missing data (unit and item nonresponse), response errors, interviewer errors, correlation bias, sampling error, etc. We wish to estimate the effects of these diverse sources of error on the DSE.

The first step in our strategy is to express the DSE as a function of components. We have constructed the components so that, for the most part, the different sources of error act either independently or perfectly dependently on different components. By isolating the effects of the various errors, we are better able to identify the major distinct sources of error.

Next, we estimate the first two moments of the component errors, one component at a time. In doing so we draw upon the results of various TARO evaluations and quality control programs. The way we constructed the components implies that correlation between component errors typically equals either 0 or 1 .

To study the propagation of errors we have used computer simulation methods. A multivariate distribution of the error components, say $F$, was assumed. The specification of $F$ was consistent with the first two moments as estimated in Section 5 . Realizations of the component errors were simulated by pseudo-random draws from F and then the DSE was calculated; this procedure was repeated 10,000 times and the resulting empirical distribution of the DSE was used as an estimate of its actual distribution. The first two moments of the latter distribution provide numerical estimates of the total error of the DSE.

Sensitivity analysis was performed to discover the importance of using one distributional form for $F$ rather than another. The results suggest that the exact distributional form (beyond the first two moments) is relatively unimportant (see Section 6).

We adopted a Bayesian approach in investigating of the error in the DSE. We estimated the first two moments of the distributions for the error components, then we derived the posterior distribution of the undercount rate conditional on the observed values of $\hat{C}, \hat{N}, \hat{M}$, etc.

## 4. COMPONENTS OF THE DSE

The DSE is subject to sampling errors and nonsampling errors, including failure of assumptions underlying the DSE model. The DSE does have a bias, but the bias in the census context is negligible (Wolter 1986a). Nonsampling errors may affect the accuracy of estimation of $N_{+1}, N_{1+}$, and $N_{11}$. Descriptions of the nonsampling error follow.

The error in the estimation of $N_{+1}$ is defined by $\hat{C}-N_{+1}=(\hat{C}-C)+\left(C-N_{+1}\right)$. The first term $(\hat{C}-C)$ is the net nonsampling error, which contributes to both bias and variance, and the second term ( $C-N_{+1}$ ) is the sampling error, which contributes only to the variance. Define the net nonsampling error as $c=\hat{C}-C$.

The net error $c$ arises during the processing of the $E$ sample when respondents are misclassified as to whether they are correctly or erroneously enumerated in the original enumeration. Therefore, $c$ has three components: $c_{e}$, which occurs during the data collection and processing; $c_{b}$, caused by a PES design that fails to balance estimates of the gross overcount and gross undercount; and $c_{i}$, caused by missing data, $c=c_{e}+c_{b}+c_{i}$. Sections 5.5, 5.6 and 5.7 cover $c_{e}, c_{b}$ and $c_{i}$, respectively.

The error in the estimation of $N_{1+}$ is defined by $\hat{N}_{p}-N_{1+}=\left(\hat{N}_{p}-N_{p}\right)+\left(N_{p}-\right.$ $N_{1+}$ ). The first term ( $\hat{N}_{p}-N_{p}$ ) is the nonsampling error, which contributes to both bias and variance and the second term ( $N_{p}-N_{11}$ ) is the sampling error, which contributes only to the variance. The net nonsampling error is defined by $n_{p}=\hat{N}_{p}-N_{p}$.

The net error $n_{p}$ arises during the interviewing for the P sample when the P -sample selections are not interviewed. This situation occurs when household members are fabricated or when there is missing data. Therefore, $n_{p}$ has two components: $n_{p f}$, the error due to fabrication and $n_{p i}$, the error due to missing data, $n_{p}=n_{p f}+n_{p i}$. Section 5.3 discusses $n_{p f}$, and Section 5.7 covers $n_{p i}$.

The error in the estimation of $N_{11}$ is defined by $\hat{M}-N_{11}=(\hat{M}-M)+\left(M-N_{11}\right)$. The first term ( $\hat{M}-M$ ) is the net nonsampling error, which contributes to both bias and variance, and the second term ( $M-N_{11}$ ) is the sampling error, which contributes only to the variance.

To facilitate the description of the nonsampling error in the estimation of $N_{11}$, consider the following tables of P -sample selections and respondents. Entries in Table 3 are the weighted number of $P$-sample selections in each category. Entries in Table 4 are the weighted number of P-sample responses in each category. Entries in Table 5 are estimates of the number of people in each category based on the P -sample interviewing, responses, and matching operation.

Table 3
P-sample Selections

|  | Census Enumeration Status |  |
| :--- | :---: | :---: |
| P-sample Selections | Enumerated | Not <br> Enumerated |
| Not reported | $D_{11}$ | $D_{12}$ |
| Reported | $D_{21}$ | $D_{22}$ |
| Correct Census Day Address | $D_{31}$ | $D_{32}$ |
| Wrong Census Day Address |  |  |

Table 4
Enumeration Status of P-sample Respondents

|  | Census Enumeration Status |  |
| :--- | :---: | :---: |
| P-sample Status | Enumerated | Not <br> Eumerated |
| Fabricated | $A_{11}$ | $A_{12}$ |
| Not Fabricated | $A_{21}$ | $A_{22}$ |
| Correct Census Day Address | $A_{31}$ | $A_{32}$ |
| Wrong Census Day Address |  |  |

Table 5
Match Status of P-sample Respondents

|  | Match Status |  |
| :--- | :---: | :---: |
| P-sample Status | Matched | Not <br> Matched |
| Fabricated | $B_{11}$ | $B_{12}$ |
| Not Fabricated | $B_{21}$ | $B_{22}$ |
| Correct Census Day Address | $B_{31}$ | $B_{32}$ |
| Wrong Census Day Address |  |  |

Since the P-sample selections who appear as reported in Table 3 are the respondents who are not fabricated in Table $4, D_{21}=A_{21}$ and $D_{31}=A_{31}$. Also, $A_{11}=0$ since a case fabricated during the PES cannot be enumerated in the census. Therefore,

$$
M=D_{11}+D_{21}+D_{31}=D_{11}+A_{21}+A_{31} .
$$

Since a case fabricated during the PES would not have a corresponding census enumeration, we assume $B_{11}=0$. Therefore, $\hat{M}=B_{11}+B_{21}+B_{31}=B_{21}+B_{31}$.

Then the nonsampling error in the estimation of $N_{11}$, called $m$, may be defined as follows:

$$
\begin{aligned}
m & =\hat{M}-M \\
& =\left(B_{11}+B_{21}+B_{31}\right)-\left(D_{11}+D_{21}+D_{31}\right) \\
& =-D_{11}+\left(B_{21}-A_{21}\right)+\left(B_{31}-A_{31}\right) .
\end{aligned}
$$

The error $m$ has three components: ( $B_{21}-A_{21}$ ), which is the error introduced in the matching operation (Section 5.2); ( $B_{31}-A_{31}$ ), which is the error introduced by respondents giving the wrong Census Day address (Section 5.3); and $-D_{11} . D_{11}$ has two components: missing match status $m_{i}$ and fabrication $m_{f}$. Section 5.7 covers missing match status, and Section 5.4 covers fabrication.

The ideal DSE can be written as follows:

$$
N_{1+} N_{+1} / N_{11}=(\hat{C}-c)\left(\hat{N}_{p}-n_{p}\right) /(\hat{M}-m)
$$

## 5. COMPONENTS OF PES ERROR

Estimates of the first two moments of the posterior distribution of the undercount rate derive from estimates of the first two moments of the components of PES error. The components are correlation bias, matching error, accuracy of the reported Census Day address, fabrication in the $P$ sample, measurement of erroneous enumerations, balancing the estimates of the gross overcount and the gross undercount, missing data, and sampling error. We next describe the source of each component of PES error and give models for each component. We model the component errors in terms of observable indicators of data quality. We estimate the first two moments of the distributions of the errors for use in the total error model in Section 6.

### 5.1 Correlation Bias

### 5.1.1 Source of Error

An important concern for dual system estimation is that the estimate of the proportion of the population enumerated in the census, based on the $\mathbf{P}$ sample, is accurate. The violation of one of the independence assumptions underlying dual system estimation may cause the estimate of the proportion of the population enumerated in the census, and thereby the estimate of the population, to be biased.

Three independence assumptions are made for dual system estimator:
Causality. The event of being included in the census is independent of the event of being included in the PES. That is, the cross-product ratio satisfies

$$
\theta_{i}=p_{i 11} p_{i 22} / p_{i 12} p_{i 21}=1, \text { for } i=1, \ldots, N
$$

Homogeneity. The capture probabilities satisfy $p_{i I+}=p_{1+}$ or $p_{i+1}=p_{+1}$ for $i=1, \ldots$, $N$, within each of the post-strata.
Autonomy. The census and the PES are created as a result of $N$ mutually independent trials.
The homogeneity assumption follows combination model $M_{t h}$ in Wolter (1986a). All the development for the Peterson model $M_{t}$ in Wolter (1986a) also applies to model $M_{t h}$ when enough information is available to form post-strata where $M_{t}$ holds.

To control heterogenity in the population the Census Bureau post-stratifies the data based on demographic and geographic variables, a technique originally recommended by Sekar and Deming (1949). An estimate of the population in each post-stratum is calculated and then all the estimates are summed to give an estimate of the total population. Unless the failure of the homogeneity assumption is severe, the estimate lies between the census and the truth.

Research by Wolter (1986b) and Cowan and Malec (1986) has demonstrated that the failure of the autonomy assumption has a negligible effect on the bias of the DSE but causes an increase in its variance. Wolter's formulation allows household members to act individually (autonomy) or together (failure of autonomy). Cowan and Malec present a model that permits clustering of the census misses (failure of autonomy). Next, we model the combined effect of the sources of correlation bias on the DSE.

### 5.1.2 Definition

For insight into the effect of correlation bias, assume all $\theta_{i}=\theta$ and write the true population size as

$$
N=N_{11}+N_{12}+N_{21}+\theta\left(N_{12} N_{21} / N_{11}\right)
$$

where $\theta_{i}$ is the cross-product ratio defined in Section 5.1.1.
The correlation bias affects only the last term because the other three may be estimated directly. The parameter $\theta$ represents the effect of the failure of the independence assumptions. When the independence assumptions hold, $\theta=1$.

The correlation bias, arising when $\theta$ does not equal 1 , is the only contributor to $t$, the error due to failure of the model. The population size can be written as follows:

$$
\begin{aligned}
N & =N_{1+} N_{+1} / N_{11}+t \\
& =N_{1+} N_{+1} / N_{11}+(\theta-1)\left(N_{12} N_{12} / N_{11}\right) .
\end{aligned}
$$

Therefore, the correlation bias, $t=(\theta-1)\left(N_{12} N_{21} / N_{11}\right)$.

### 5.1.3 Measurement

The parameter $\theta$ may be estimated at the national level for racial and ethnic subgroups using demographic analysis estimates of the population size. Note, however, that this technique presumes that the demographic analysis estimates are accurate. Even so, this formulation also permits varying $\theta$ to assess the sensitivity of the DSE to the estimate of the effect of the violation of the independence assumptions.

### 5.1.4 Estimation

Estimates for $\theta$ were not made for the 1986 TARO because an alternate source for population estimates did not exist, e.g., no demographic analysis estimates were feasible. However, Ericksen and Kadane (1985) made three estimates of $\theta$ for blacks for the 1980 census: 2.1, 2.7, and 3.7. Since the population in the 1986 TARO was predominantly minority ( 73 percent Hispanic, 12 percent Asian, and 15 percent non-Asian and non-Hispanic), the Ericksen and Kadane estimates for 1980 will be used in this paper: $\mathrm{E}(\theta)=2.1,2.7$, or $3.7, \operatorname{Var}(\theta)=0$. We are treating $\theta$ as fixed, but unknown. A sensitivity analysis is conducted in Section 6 to demonstrate the effect of alternative values of $\theta$.

These estimates of $\theta$ are consistent with the reports of the participant observers in the Los Angeles test site (Childers et al. 1987). Our professional judgment is that correlation bias is higher for urban areas than for the country as a whole. This implies that these estimates may be conservative for the Los Angeles test site because it was urban.

### 5.1.5 Summary

In the total error model the first two moments of the posterior distribution of the correction factor for correlation bias are assumed to be $\mathrm{E}(\theta)=2.1,2.7$, or 3.7 , and $\operatorname{Var}(\theta)=0$.

### 5.2 Matching Error

### 5.2.1 Source of Error

Matching error in this discussion refers to errors that occur in the operation where the $P$ sample is matched to the original enumeration. Therefore, matching error does not encompass response errors that arise in the data collection. Although other types of errors may result in an inaccurate assignment of a P-sample respondent's census enumeration status, these sources are treated in other components of error.

After the $P$-sample interviewing is completed, a search of the census is conducted to determine if the respondents are enumerated. Then the P -sample respondents are designated as matching an enumeration in the census or as not enumerated in the census. Errors in assigning the enumeration status to P -sample persons which occur during the processing of the data are known as matching error. Errors may occur in either direction. People may be designated as matching a census enumeration although they are not in the census, called a "false match," or people may be designated as not enumerated although they are, called a "false nonmatch." Matching error will cause a bias in the estimate of the number of people in both the census and the P -sample population and thereby introduce a bias into the estimates of the number of people missed by the census.

### 5.2.2 Definition

The denominator $N_{11}$ of the dual system estimator is estimated from sample survey data, the P sample. The following were introduced in Section 4:
$A_{21}=$ the weighted number of people who were enumerated,
$B_{21}=$ the estimate of the number of people who match.

Then the net error due to incorrect classification of enumeration statuses, $m_{m}$, may be defined as $m_{m}=B_{21}-A_{21}$. The conditional expected value and variance of $m_{m}$ given observed value $\hat{M}$ are denoted by $\mathrm{E}\left(m_{m}\right)$ and $\operatorname{Var}\left(m_{m}\right)$.

### 5.2.3 Measurement

Measurement of $m_{m}$ is possible by processing a sample of the cases a second time i.e., by having highly trained personnel rematch them. The assumption underlying an independent rematch of a sample is that the personnel with more training make fewer mistakes in classifying enumeration statuses although they have the same materials and information available as the original workers. The original match codes and the evaluation match codes can be reconciled, and the discrepancies can be resolved.

Two evaluations of the clerical matching were conducted with the 1986 TARO data. One study evaluated the clerical matching for movers, and another evaluated the clerical matching for nonmovers.

In the evaluation of matching for nonmovers (Corby and Mulry 1988), a probability subsample of 35 blocks was chosen for a rematch by professionals from headquarters. The sample was stratified by match rate, and blocks with low match rates were sampled at a disproportionally high rates so that the quality control staff could learn as much as possible about matching errors. Adjacent blocks were not searched so the false nonmatches are possibly underestimated.

The second evaluation study considered matching error for movers (Childers et al. 1987). There were 90 movers who were not matched in TARO, and all of these movers were rematched. Eleven matches were found, two of which had been lost during the computer editing.

### 5.2.4 Estimation

We now use the results of the evaluation subsamples to estimate the moments of the distribution of $m_{m}$ from the PES sample. Not conducting an extended search in the evaluation for the nonmovers probably reduced the number of false nonmatches found. Experience with extended searches implies that adding an additional 20 percent of the net error of 70 (Hogan and Wolter 1988) is a conservative way to compensate for the lack of one. The results from the two evaluations yield a net error of 95 in the PES sample. Therefore, the net error rate is -.0055 . We apply the net error rate to only the P -sample cases with a resolved match status because the error in the imputation for the unresolved cases is covered in the Missing Data Section 5.7. The expected value of $m_{m}$ becomes $\mathrm{E}\left(m_{m}\right)=-1831$, when the overall sampling weight of 17 is used.

An estimate of the variance of the estimate of net matching error for nonmovers has not been calculated. The sample variance of the number of errors for movers is zero because all the nonmatched movers were rematched. However we do not believe that the true variance is zero. One way to obtain a variance specificiation would be to assume that the errors occurred in the manner of a mixture of Poisson processes, e.g., matching errors for movers followed one Poisson processs and matching errors for nonmovers independently followed another Poisson process. Treating the errors as arising from a simple Poisson process would then lead to a conservative estimate of variance; in this case the variance would be estimated by $17 \times 107$. However, the Poisson model may not be conservative if the errors occur in clusters. In an attempt to develop conservative estimates of variance, we have (somewhat arbitrarily) multiplied the variance estimate under the simple Poisson model by the overall sampling weight to obtain

$$
\operatorname{Var}\left(m_{m}\right)=(17)^{2} \times 107=30,923
$$

### 5.2.5 Summary

For the total error model, the first two moments of the posterior distribution of the net matching error for the PES sample are assumed to be $\mathrm{E}\left(m_{m}\right)=-1831$ and $\operatorname{Var}\left(m_{m}\right)=30,923$.

### 5.3 Quality of the Reported Census Day Address

### 5.3.1 Source of Error

Some of the respondents in the $P$ sample have moved between Census Day and their PES interview. The respondents may misreport whether they have moved during the time lapse. If they have moved, they may not report their previous address accurately, or their previous address may not be geocoded correctly by the staff. Any of these types of errors may cause the matching operation to search the census in an area other than where the respondent was enumerated. These errors may lead to assigning a nonmatch status to respondents who actually were enumerated because the matching operation is unable to locate their enumerations. Inappropriate assignment of the status of nonmatch will cause the estimate of the number of people missed by the census to be biased upward.

Circumstances under which inaccurate reporting of the Census Day address by a PES respondent will not cause a false nonmatch do exist. If the Census Day address is inside the search area for the reported address, and the reported address is geocoded correctly, then the matching operation will find the person.

### 5.3.2 Definition

The denominator $N_{11}$ of the dual system estimator is estimated from sample survey data, the P sample. The following were introduced in Section 4:
$A_{31}=$ the weighted number of people with an inaccurate Census Day address who are enumerated,
$B_{31}=$ the estimate of the number of people with an inaccurate Census Day address who match at another address.

Then the net error due to inaccurate reporting of the Census Day address, $m_{a}$, may be defined as $m_{a}=B_{31}-A_{31}$. The conditional expected value and variance of $m_{a}$ given the observed value $\hat{M}$ are denoted by $\mathrm{E}\left(m_{a}\right)$ and $\operatorname{Var}\left(m_{a}\right)$.

### 5.3.3 Measurement

Measurement of $m_{a}$ is based on a follow-up of a sample of P-sample respondents whose enumeration status is "not enumerated". Data from the follow-up are used to estimate the error that arises when people who were enumerated misreport their Census Day address when they respond to the PES.

An evaluation of the quality of the reporting of the Census Day address was conducted after the 1986 TARO. A post-production follow-up which reinterviewed a sample of 903 of the nonmatches was aimed at determining the number of nonmatches caused by misreporting mover status. Another search to match respondents who reported they in fact had moved within the test site was made at the new address.

### 5.3.4 Estimation

The sample cases found to have errors in their reported Census Day address may be used to estimate
$L_{e}=$ the weighted number of people who erroneously report their Census Day address in their P-sample interview.

A search of census enumerations at the newly reported addresses produces
$r_{a m}=$ the estimator of the percentage of people with errors in the location of their reported Census Day address who match census enumerations.
Then the expected value of the error $m_{a}$ is estimated by

$$
\mathrm{E}\left(m_{a}\right)=-r_{a m} L_{e} .
$$

The results of the post-production follow-up (Hogan and Wolter 1988) yielded a misreporting rate of at most 3.1 percent in the $P$ sample. A match rate of 33 percent was estimated for those who misreported their Census Day address and moved within the test site. If we assume the match rate for those who reported a census day address outside the test site is also 33 percent, then the expected value $\mathrm{E}\left(m_{a}\right)=-3481$.

An estimate of the variance of the error due to misreporting has not been made. Our professional judgment is that a conservative estimate of the variance at the PES sample level is 900 . Therefore, the variance at the TARO site level is

$$
\operatorname{Var}\left(m_{a}\right)=(17)^{2} \times 900=260,100
$$

### 5.3.5 Summary

For the total error model, the first two moments of the distribution of the error due to misreporting of Census Day address for the PES sample are assumed to be $\mathrm{E}\left(m_{a}\right)=-3481$ and $\operatorname{Var}\left(m_{a}\right)=260,100$.

### 5.4 Fabrication in the $P$ sample

### 5.4.1 Source of Error

Interviewers may fabricate people in P-sample housing units. Research has shown that interviewer fabrication during the PES may result in a substantial bias in the estimates of census coverage error based on the dual system estimator. Basically, the creation of fictitious individuals may decrease the PES match rate, causing the estimate of coverage error to be too large.

Experience at the Bureau of the Census has shown that fabrication of the members of a whole household is the problem for household surveys. Rarely is there a fabrication of the household member in a household where the other members are the real residents.

The quality control operation for the interviewing phase of the $P$ sample is designed to check for fabricated interviews and to interview the real household members. Therefore, no statistical correction for fabrication in the $P$ sample is made in the formation of the dual system estimates.

### 5.4.2 Definition

The $N_{11}$ and $N_{1+}$ in the dual system estimator are estimated from sample survey data, the $P$ sample. The following were introduced in Section 4:
$m_{f}=$ the weighted number of people who were replaced by fabricated P-sample interviews and who were enumerated,
$n_{p f}=$ the error in $N_{p f}$ due to households that were fabricated in the P sample.
The posterior expected values and variances of $m_{f}$ and $n_{p f}$ are denoted by $\mathrm{E}\left(m_{f}\right)$ and $\mathrm{E}\left(n_{p f}\right)$ and $\operatorname{Var}\left(m_{f}\right)$ and $\operatorname{Var}\left(n_{p f}\right)$.

### 5.4.3 Measurement

In the 1986 TARO, the estimate of the fabrication rate based on the quality control of the interviewing was approximately 0.6 percent. The estimate of the fabrication rate based on a post-production follow-up was approximately 1.2 percent (Hogan and Wolter 1988).

### 5.4.4 Estimation

We now estimate the moments of the posterior distributions of $n_{p f}$ and $m_{f}$ from the PES sample. We believe it is reasonable to assume $n_{p f}$ is negligible in TARO. Therefore, the expected value and variance are given by $\mathrm{E}\left(n_{p f}\right)=0$ and $\operatorname{Var}\left(n_{p f}\right)=0$.

The quality control data may be used to estimate $r_{f}=$ the rate at which P -sample interviews are fabricated.

The search of the census enumerations for people in the $P$ sample who were found by the quality control operation to not have been properly interviewed produces $r_{f m}=$ the match rate for people not interviewed because their household was fabricated in the $P$ sample.

In TARO, records were not kept so that the people who were discovered by the quality control not to have been interviewed properly could be identified. Therefore, no search was made for matching enumerations. Since we have no data available for a direct estimate of $r_{f m}$, we conservatively assume that the people not interviewed properly are like the people who were. We set $r_{f m}$ equal to the final overall P-sample match rate.

We use the conservative results from the post-production follow-up to yield a fabrication rate of 1.2 percent. The match rate for TARO is 88.6 percent (Diffendal 1988). Therefore, the expected value of the error $m_{f}$ is given by $\mathrm{E}\left(m_{f}\right)=-2502$.

An estimate of the variance of the estimate of fabrication error has not been calculated. Our professional judgment is that a conservative estimate of the variance can be derived by the reasoning discussed in Section 5.4.2. Thus, we estimate that the variance for the TARO site is

$$
\operatorname{Var}\left(m_{f}\right)=(17)^{2} \times 206=59,534
$$

### 5.4.5 Summary

For the total error model, the first two moments of the distribution of the net error due to fabricated interviews are assumed to be $\mathrm{E}\left(m_{f}\right)=-2502$ and $\operatorname{Var}\left(m_{f}\right)=59,534$. The net error due to fabricated interviews in is assumed to be negligible, and therefore, $\mathrm{E}\left(n_{p f}\right)=0$ and $\operatorname{Var}\left(n_{p f}\right)=0$.

### 5.5 Measurement of Erroneous Enumerations

### 5.5.1 Source of Error

Some enumerations may have been entered in the census as the result of mistakes. These enumerations are called erroneous enumerations. Since the dual system estimator requires estimating the number of distinct people captured in the census, a correction is made for erroneous enumerations in the estimate of total population. Subtracting the estimate of the number of enumerations that do not correspond to distinct people from the census count provides an improved estimate of the number of distinct people captured in the census. This estimated correction is obtained from the E sample in the PES.

The following types of enumerations are considered erroneous: (1) people who died before Census Day, (2) people who were born after Census Day, (3) enumerations that do not refer to real people, (4) people duplicated, (5) people enumerated outside the search area where the matching operation looks for their enumeration. The search area for a case includes the block for its address and the ring of adjacent blocks.

This component is caused by errors in measuring census error. An error in the estimation of the number of erroneous enumerations occurs either when an enumeration in the E sample
is designated as erroneous although it is correct, or when an enumeration is designated as correct although it is really erroneous. Therefore, both positive and negative error can occur in the estimation of the number of erroneous enumerations.

The types of enumerations that are the most vulnerable to misclassifiction as to whether they are erroneous include the duplicated and fabricated enumerations. These errors are the only ones considered because the others are either inconsequential or are treated separately. Errors in identifying enumerations for people who died before Census Day and people who were born after Census Day have a trivial effect. Errors in classifying the eumeration status because a person was enumerated outside the search area is covered in Section 5.6 on balancing the estimates of the gross overcount and the gross undercount.

### 5.5.2 Definition

The bias in the DSE due to misclassification of enumeration status is caused by error in the estimation of $N_{+1}$. In the formation of the estimate of the number of distinct people in the original enumeration $\hat{C}$, a correction is made for the number of erroneous enumerations, $\widehat{E E}$. $\widehat{E E}$ and therefore $\hat{C}$ are estimated from sample survey data, the E sample. Errors in the estimate $\hat{C}$ occur through the misclassification of the enumeration status of E-sample cases. Let
$c_{e}=$ the difference between the weighted number of erroneous enumerations misclassified as correct and the weighted number of correct enumerations misclassified as erroneous.
The expected value of $c_{e}$, conditional on the observed value $\hat{C}$, is denoted by $\mathrm{E}\left(c_{e}\right)$. The variance of $c_{e}$, conditional on the observed value $\hat{C}$, is denoted by $\operatorname{Var}\left(c_{e}\right)$.

### 5.5.3 Measurement

Processing error may be measured directly using a rematch of a sample of cases. Errors from other sources, such as duplications due to violations of census residency rules, can be assessed by viewing the frequency distributions of the erroneous enumerations. This is preferable to direct measurement of these errors because of the difficulties in obtaining accurate data in additional follow-ups. When tests confirm that the gross errors from these sources are under control, the net error can be assumed to be negligible. For example, the distribution of the erroneous enumerations by age group is expected to have a large number of duplications in the highlymobile groups of the population where there are more opportunities for the census residency rules not to be followed.

In the 1986 TARO, an evaluation of the E-sample processing was conducted in conjunction with the evaluation of the P-sample matching operation discussed in Section 5.2.3 (Corby and Mulry 1988). The data for the E sample from the same subsample of 35 blocks were reprocessed.

### 5.5.4 Estimation

We now estimate the moments of the distribution of $c_{e}$ from the PES sample. The results of the reprocessing (Hogan and Wolter 1988) yield a net error rate of 0.0007 in the identification of correct enumerations. The expected value of $c_{e}$ is $\mathrm{E}\left(c_{e}\right)=-238$. This estimate is based on the $E$ sample with a resolved enumeration status because the error in the imputation for the unresolved cases is covered in the Missing Data Section 5.7.

An estimate of the variance of net error has not been calculated. Our professional judgment is that a conservative estimate of the variance can be derived by the reasoning discussed in Section 5.2.2. Thus, we estimate that the variance for the TARO site is $\operatorname{Var}\left(c_{e}\right)=(17)^{2} \times 14=4,046$.

### 5.5.5 Summary

For the total error model, the first two moments of the posterior distribution of the net error in identifying correct enumerations are assumed to be $\mathrm{E}\left(\boldsymbol{c}_{e}\right)=-238$ and $\operatorname{Var}\left(c_{e}\right)=4,046$.

### 5.6 Balancing the Estimates of the Gross Overcount and Undercount

### 5.6.1 Source of Error

Both the E sample and the P sample measure enumeration errors in the census. The E sample measures the gross overcount in the form of erroneous enumerations. The $P$ sample measures the gross undercount in the form of those not enumerated. Ideally, the entire census would be searched before a P-sample person was declared to be not enumerated. Ideally, the entire country would be searched to determine if an E-sample enumeration is a duplicate or fictitious. Of course, such extensive searches are simply not feasible in the performance of the PES. These searches must be limited in the reasonable manner. The way chosen has to preserve the net error although the measured gross overcount and the measured gross undercount may increase due to limiting the search area. The gross overcount and the gross undercount have to balance to equal the net coverage error.

Failure to have procedures which balance the estimated gross overcount and the estimated gross undercount may cause an incorrect number of enumerations in the $E$ sample to be designated as erroneous when they are correct. This error may cause either an upward or downward bias.

Balancing is not an issue for the design of the PES planned for 1990 and tested in the 1986 TARO, as it was in 1980. The design calls for overlapping the $P$ sample and the $E$ sample. The same blocks are included in the $P$ sample as in the $E$ sample. The $P$-sample search area is, by definition, the proper search area. The E-sample search area is chosen to be consistent with the P -sample search area.

### 5.6.2 Summary

Error due to geocoding error is believed to be negligible in the 1986 TARO and will not be included in the total error model. The appendix contains a model for balancing arror.

### 5.7 Missing Data

### 5.7.1 Source of Error

Both the E sample and the P sample have missing data. The E sample has cases where the information required to determine whether the person is correctly or erroneously enumerated in the census is not available. The $P$ sample has cases where the information needed to determine whether the person is enumerated in the census is not available. The probability of being enumerated is imputed statistically to compensate for the inablility to resolve the case.

An unresolved status may occur in more than one way. The interviewer may be unable to obtain an interview during the $\mathbf{P}$-sample interviewing or during the PES follow-up. A P-sample or E-sample questionnaire may not have all the demographic and housing information required for the estimation. Even with all the information requested on the questionnaires, the circumstances may be so unclear that the enumeration status can not be resolved.

### 5.7.2 Measurement

We assess the error in the DSE caused by missing data instead of considering each component $c_{i}, m_{i}$ and $n_{p i}$ separately. Our approach is to perform a sensitivity analysis of reasonable alternative models for compensating for missing data. First a preferred method of imputation for
unresolved P-sample and E-sample enumeration statuses is specified prior to the implementation of the PES. Reasonable alternative treatments of the missing data can be suggested by problems that arise during the collection and processing of the PES data. The DSE can be computed under these alternative models for compensating for missing data. The range of the alternative estimates indicates the sensitivity of the DSE to the method of imputation. For example, a narrow range implies that the estimates are robust, and the missing data cause little uncertainty in the estimates.

### 5.7.3 Estimation

The effect of missing data on the estimates from the 1986 TARO was assessed by examining the range of estimates obtained when methods of imputation based on reasonable alternative assumptions were used in place of the preferred method. These included alternative treatment of proxy responses, movers, and designation of ficticious enumerations (Schenker 1988). The alternative treatment of the proxy interviews for P-sample cases classified them as noninterviews and applied the weighting adjustment. This essentially assigned proxy cases the same match rate as nonproxy cases. The alternative treatment of the $P$-sample movers reclassified them all as unresolved and imputed a match probability, instead of imputing for only those who were not resolved. This essentially assigned movers the same match rate as nonmovers. The alternative treatment of fictitious cases resulted from a review of the unresolved E-sample cases by experienced matching personnel who converted some unresolved cases to fictitious. This raised both the observed and imputed rates of erroneous enumeration.

Models 000 and 111 shown in Table 4 of Schenker's paper give the upper and lower bounds of the estimates of undercount rates, respectively. Both models differ from TARO in that they have inmovers as substitutes for outmovers. P-sample inmovers are P-sample respondents who moved into their housing unit between Census Day and PES interviewing. In the 1986 TARO the P -sample inmovers from areas outside the test site were omitted from the PES estimation. The omission of the outmovers from estimation essentially assumes that they had the same capture rate in the original enumeration as the included cases. Movers are believed to have a lower capture rate than nonmovers. Model 000 has the TARO treatments while Model 111 has all the alternative treatments.

### 5.7.4 Summary

The effect of missing data on the distribution of the total error is assessed by computing the distribution of the undercount rate under several reasonable imputation methods. The alternative methods which yield the upper and lower bounds for the undercount are used in the total error analysis.

### 5.8 Sampling Error

### 5.8.1 Source of Error

The observed DSE is subject to sampling error because $\hat{N}_{p}, \hat{C}$, and $\hat{M}$ are estimated from samples. The sample size for the PES is determined by the amount of sampling error and budget allowable. Other things being equal, the larger the sample size the lower the amount of sampling error introduced in the estimates. The sampling errror is affected by the estimator and the sampling design. In the TARO PES design, both the $P$-sample and the E-sample observations are collected from the same sample of blocks. All the people residing in the housing units in the selected blocks are included in the $P$ sample. All enumerations assigned by the census process to the sample block are included in the E sample. The estimation of the sampling error takes into account the tendency for census misses and erroneous enumerations to be correlated within blocks and within housing units. Experience has shown that many hard-to-enumerate areas have both a higher rate of omissions and a higher rate of erroneous enumerations.

### 5.8.2 Measurement

The standard randomization theory model for survey sampling is appropriate for estimating the variance of the DSE. The coefficient of variation which is the ratio of the square root of the variance of the observed DSE to the mean of the distribution of the DSE provides information on the amount of sampling error in the DSE.

The Taylor series estimator of variance for the observed dual system estimator (Moriarity 1987), $v\left(\bar{N}_{++}\right)$, is given by

$$
\begin{aligned}
\mathrm{v}\left(\hat{N}_{++}\right) & =\hat{N}_{++}^{2}\left(\mathrm{v}\left(\hat{N}_{p}\right) / \hat{N}_{p}^{2}+\mathrm{v}(\hat{M}) / \hat{M}^{2}-2 \mathrm{c}\left(\hat{N}_{p^{\prime}} \hat{M}\right) / \hat{N}_{p} \hat{M}\right) \\
& +\hat{N}_{p}^{2} \mathrm{v}(\hat{E}) / \hat{M}^{2}+2 \hat{N}_{++}\left(\hat{N}_{p} \mathrm{c}(\hat{E}, \hat{M}) / \hat{M}^{2}-\mathrm{c}\left(\hat{E}, \hat{N}_{p}\right) / \hat{M}\right)
\end{aligned}
$$

where

$$
\begin{aligned}
& \hat{E} \quad=I I_{2}+\widehat{E E}, \\
& \mathrm{v}(X) \quad=\text { the estimator of the variance of an estimator } X, \\
& \mathrm{c}(X, Y)=\text { the estimator of the covariance between } X \text { and } Y .
\end{aligned}
$$

The categories $I_{2}$, insufficient information for matching, and $\widehat{E E}$, erroneous enumerations, are treated as one group in the variance estimation. The variance and covariance estimators reflect the cluster sampling of blocks and block clusters.

### 5.8.3 Estimation

The standard deviation of the dual system estimate of 388,040 for the TARO site is $3,100.37$. The coefficient of variation is 0.008 . This implies the standard deviation for the estimated net undercount rate is 0.7 percent.

### 5.8.4 Summary

The sampling error for the TARO DSE is $3,100.37$, and the sampling error for the TARO net undercount rate estimate is 0.70 percent.

## 6. SYNTHESIS OF TOTAL ERROR

The combined effect of the component errors will be summarized by posterior distibutions for the net undercount rate. The bias in the estimate of net undercount rate, $\mathrm{B}(U)$, is estimated by the difference between and the mean of the posterior distribution. To construct the posterior distribution, we used a simulation method with 10,000 repetitions, generating pseudo-random component errors and adding them to the TARO estimates. Using the formulas in Section 5.1.2, we obtain the following formula:

$$
\begin{aligned}
N= & \left(\hat{N}_{p}-n_{p}\right) \cdot+(\hat{C}+c-(\hat{M}-m)) \\
& +\theta(\hat{C}-c-(\hat{M}-m))\left(\hat{N}_{p}-n_{p}-(\hat{M}-m)\right) /(\hat{M}-m) \\
= & (\hat{C}-c)\left(\hat{N}_{p}-n_{p}\right) /(\hat{M}-m) \\
& +(\theta-1)(\hat{C}-c-(\hat{M}-m))\left(\hat{N}_{p}-n_{p}-(\hat{M}-m)\right) /(\hat{M}-m) .
\end{aligned}
$$

Several different distributions were used to reflect alternative estimates of imputation error, alternative estimates of correlation bias (parameterized by $\theta$ ), and alternative marginal distributional forms for the components - normal, gamma, and uniform.

In this study, the estimate of percent net undercount for the TARO site is 8.42 with a sampling standard deviation of 0.7 . This estimate was selected because estimates of nonsampling error components are available only for the site as a whole. When a DSE is constructed for each post-stratum and then the DSEs are summed to give an estimate for the site, the percent net undercount estimate is 9.02 .

Table 6 displays the means and standards deviations of the error components for the PES sample. Recall that the DSE for the TARO site is $388,040, \hat{M}=298,204, \hat{C}=343,567$, and $\hat{N}_{p}=336,707$. The overall sampling weight, 17 , was used consistently throughout all the simulations so that comparisons of the effect of alternative assumptions such as correlation bias parameter values, error distributions, and imputation models are appropriate. The methodology generalizes to other applications where a different sampling weight is used in each stratum.

Table 7 displays the effects of the individual errors on the posterior distribution of the undercount when the TARO imputation is used. The net matching Census Day address, and fabrication errors are all errors in $\hat{M}$. Therefore, the presence of only one of them alone causes the bias in the estimate of percent net undercount to be positive. The net E-sample error is an error in $\hat{C}$. The presence of $E$-sample error alone causes the bias in the estimate of percent net undercount to be negative. The estimate for correlation bias, was chosen to be 2.7 , the median of Ericksen and Kadane's estimates. The presence of only correlation bias causes the bias in the percent net undercount estimate to be negative.

Table 6
Assumed Distributions of Error Estimates

|  | Mean | Standard <br> Deviation |
| :--- | :---: | :---: |
| Net Matching Error | -1831 | 176 |
| Census Address Error | -3481 | 510 |
| Fabrication Error | -2502 | 244 |
| Net E sample Error | -238 | 64 |

Table 7
Individual Effects of Errors on Posterior Distribution
of Percent Net Undercount and Bias in the Estimate of Undercount

|  | $\mathrm{E}(U)$ | Std. Dev. | $\mathrm{B}(\hat{U})$ |
| :---: | :---: | :---: | :---: |
| Net Matching | 7.86 | 0.06 | 0.56 |
| Census Address | 7.35 | 0.16 | 1.07 |
| Fabrication | 7.34 | 0.08 | 1.08 |
| Net E sample | 8.49 | 0.02 | -0.07 |
| Correlation Bias (2.7) | 10.61 | 0.00 | -2.19 |

Frequency


Figure 1. Percent Undercount when $\theta=2.7$

Table 8
Percentiles of the Posterior Distribution of Percent Net Undercount for $\theta=2.7$

|  | 1 | 5 | 10 | 25 | 50 | 75 | 90 | 95 | 99 |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Normal | 6.70 | 6.86 | 6.94 | 7.08 | 7.24 | 7.40 | 7.54 | 7.63 | 7.79 |
| Uniform | 6.75 | 6.86 | 6.93 | 7.07 | 7.24 | 7.42 | 7.55 | 7.62 | 7.73 |
| Gamma | 6.67 | 6.84 | 6.93 | 7.08 | 7.24 | 7.40 | 7.53 | 7.61 | 7.74 |

Table 9
Posterior Distribution of the Net Undercount Rate for Several Values of $\theta$

| $\theta$ | $\mathrm{E}(U)$ | St. Dev. | $\mathrm{B}(O)$ |
| :--- | :---: | :---: | :---: |
| 1.0 | 5.75 | 0.18 | 2.67 |
| 2.1 | 6.72 | 0.22 | 1.70 |
| 2.7 | 7.24 | 0.23 | 1.18 |
| 3.7 | 8.09 | 0.27 | 0.33 |

Simulations were conducted where the first two moments for error $n_{p}, c_{e}, m_{m}, m_{f}, m_{a}$, and $\theta$ were held constant, but the distributions were varied. We assessed the total error when all the error distributions were normal, all were gamma, and all were uniform. Varying the distributions had minor effects on the distribution of the percent net undercount. In each case the distribution of the percent net undercount was very close to normal. Figure 1 shows the distribution of the undercount when $\theta=2.7$, and it is illustrative of the results of the simulations.

Table 8 shows the percentiles of the distribution of the net undercount rate for different distributions for the component errors when $\theta$ is taken to be 2.7 and the TARO imputation is used. The standard deviation for the posterior distribution was 0.23 . In all the cases, a normal distribution is an adequate approximation. The percentiles differed by at most 0.02 for the percentiles between 5 and 95 . The 1 and 99 percentiles differed by at most 0.08 .

Varying the value of the estimate of $\theta$ for the correlation bias did affect the moments of the posterior distribution of the undercount. The variation appears in the mean and in the standard deviation. Table 9 shows the results for the different values of $\theta$, where the distribution for the errors are normal. The case where $\theta=1$ portrays virtually no correlation bias, while for the other sources of error are present. In the cases where $\theta=2.1,2.7$, and 3.7 , all the sources of error are taken into account. The distribution of the undercount shifts to the right as the estimate of $\theta$ for the correlation bias increases. The variance also increases as the estimate of $\theta$ increases. For all values of $\theta$ considered, the bias $\mathrm{B}(\hat{U})$ is positive although it decreases as $\theta$ increases.

The simulations were conducted with reasonable alternative models for the imputation for unresolved match status. Although there was some variation in the first two moments of the distribution of the net undercount rate, the estimate of net undercount rate in TARO appears robust to missing data. Table 10 illustrates the results of the simulations using models 000 and 111 described in Section 5.7.3. Models 000 and 111 yielded the upper and lower bounds of the undercount estimates under all the reasonable alternative imputation models. The bias in the estimate of the percent net undercount rate ranges from 0.93 to 2.79 . In other words, the bias is between 11 percent and 33 percent of the net undercount rate estimate of 8.42. Varying the imputation model has almost no effect on the standard deviation.

Table 10
Posterior Distribution of the Percent Net Undercount Under Reasonable Alternative Imputation Models When $\theta=2.7$

|  | $\mathbf{E}(U)$ | St. Dev. | $\mathrm{B}(\hat{U})$ |
| :--- | :---: | :---: | :---: |
| TARO | 7.24 | 0.23 |  |
| Model 000 | 7.49 | 0.23 | 1.18 |
| Model 111 | 5.63 | 0.22 | 0.93 |

The total variance of the estimated net undercount rate may be estimated by the sum of the sampling variance and the nonsampling variance. For the case where $\theta=2.7$, the standard deviation shown in Table 10 for both models 000 and 111 is 0.22 which translate to a nonsampling variance of 0.0005 when all errors are considered. The standard deviation of the estimate of net undercount rate is 0.70 which translates to a sampling variance of 0.49 . Therefore, the total variance is 0.0054 and standard error is 0.73 . The coefficient of variation of the net undercount rate is 0.083 . The nonsampling variance contributes very little to the total variance relative to the contribution by the sampling variance.

## 7. CONCLUSIONS

When the post-stratification is used in the estimation, the undercount estimate for TARO is 9.02 . The post-stratification increased the net undercount rate estimate by 0.6 , which is less than one standard deviation of 0.73 from the estimate of 8.42 . Although we expect the error in the post-stratified estimate is smaller, the result is consistent with the error analysis.

As we consider all the sources of error in the posterior distribution of the net undercount rate, we do not know the distribution of the correlation-bias parameter $\theta$. Although we could assume a prior distribution for $\theta$, others might disagree. If we were certain that $\theta$ is 2.7 , then our 95 percent confidence interval for the net undercount rate would be

$$
4.77<U<9.55
$$

We calculate this by taking the post-stratified estimate 9.02 and adjusting for the two bias estimates in Table 10, 2.79 and 0.93, and two standard deviations, $2 \times 0.73$. We feel this is a conservative estimate since we use two different bias estimates from imputation models 000 and 111. A very conservative 95 percent confidence interval for $U$ for any value of $\theta$ between 2.1 and 3.7 is $(4.43,10.32)$.

We believe the methodology described in this paper is applicable in the 1990 census with appropriate modifications. Areas for further research are nonsampling error estimates for poststrata, a distribution for the correlation-bias parameter, and models for address reporting error.

## ACKNOWLEDGEMENTS

The authors wish to thank Aref Dajani for developing the computer software for the simulations and Chris Moriarity for the variance computation. We are also grateful to Kirk Wolter and Howard Hogan for helpful discussions and suggestions during the development of the models. Bruce Spencer thanks the Undercount Research Staff at the Census Bureau for supporting this research under a Joint Statistical Agreement. The authors thank the referees for their helpful comments and suggestions.

## APPENDIX

## Definition of Balancing Error

The non-linearity of the dual system estimator makes an additive model inadequate for viewing the technical implications of the balancing of the estimated gross overcount and the gross undercount. Therefore, a more appropriated multiplicative model is developed in this section.

Limiting the E-sample and the P-sample search areas affects two parts of the DSE. One effect is a bias in the estimate of the number of erroneous enumerations, $\widehat{E E}$. The other is a bias in the estimate of the number of people in both the census and the $P$-sample population, $\hat{M}$.

The following definitions are needed for examining the effects of limiting the E-sample and the P -sample search areas in the TARO design on the dual system estimate:
$b=$ the proportion of the correct census enumerations that are in their P-sample search area.
$g=$ the ratio of the number of correct census enumerations that are in their E-sample search area to the number that are in their P -sample search area.
The proportion $g$ reflects error in the implementation of the survey committed when the $E$-sample search area is not equal to the $P$-sample search area. The way TARO was executed implies $g=1$. To show what would happen if $g$ does not equal 1 , we will carry $g$ through the discussion.

The limiting of the search area causes only a percentage $b$ of the $P$-sample people who are in both the census and the P -sample population to be designated as matching a census enumeration. Under these circumstances, a systematic bias equal to ( $1-b$ ) $N_{11}$ is introduced into the esimation of the number of people in both the census and the P -sample population. Therefore, the observed really estimates $b N_{11}$.

Likewise, the limiting of the search area causes only a percentage $b$ of the census enumerations to be available to be designated as correct. Then only a percentage $g$ of those, the ones whose search areas are consistent with the proper E-sample search areas, will be designated as correct. Under these circumstances a systematic bias equal to ( $1-b g$ ) $N_{1+}$ is introduced into the estimation of the number of distinct people in the census. This bias occurs in the estimation of the number of erroneous enumerations, $\widehat{E E}$. With this formulation, the observed number of distinct people in the census really estimates $b g N_{1+}$.

If $g=1$, no systematic bias is present in the estimation of the dual system estimate because $b g N_{+1} N_{1+} /\left(b N_{11}\right)=N_{1+} N_{+1} / N_{11}$.

The error in the estimation of $N_{+1}$ due to the failure to balance may be defined by
$c_{b}=$ the error in the number of erroneous enumerations due to the failure to define the E-sample search areas consistently with the P -sample search areas.

The error $c_{b}$ would be nonzero if $g$ does not equal 1. The ratio $g$ may be greater than or less than 1 . The error is given by $c_{b}=b(g-1) N_{+1}$.

## Measurement

In TARO, $c_{b}$ was evaluated by testing to confirm that balancing was not an issue and that the design was under control. The percentage of matching enumerations found within the sample block was large, which implies that the design was under control. Since the design was under control, $g$ is assumed to be approximately 1 , and $c_{b}$ is assumed to be negligible.

## Estimation

The geocoding appeared to be very good in the TARO test site. However, no formal measurement of the effects of any misassignment on the estimation of $\widehat{E E}$ was conducted. Therefore, $g$ is assumed to be 1 , which implies $\mathrm{E}\left(c_{b}\right)=0$ and $\operatorname{Var}\left(c_{b}\right)=0$.

## REFERENCES

CHILDERS, D., DIFFENDAL, G., HOGAN, H., SCHENKER, N., and WOLTER, K. (1987). The technical feasibility of correcting the 1990 Census. Proceedings of the Social Statistics Section, American Statistical Association, 36-45.

CORBY, C., and MULRY, M. (1988). Memorandum to K.M. Wolter, Subject: Matching Error Pilot Study. Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.

COWAN, C.D., and MALEC, D.J. (1986). Capture-recapture models when both sources have clustered observations. Journal of the American Statistical Association, 81, 347-353.

DIFFENDAL, G., (1988). The 1986 test of adjustment related operations in Central Los Angeles County. Survey Methodology, 14, 71-86.

ERICKSEN, E.P., and KADANE, J.B. (1985). Estimating the population in a census year: 1980 and beyond. Journal of the American Statistical Association, 80, 98-108, 129-131.

HOGAN, H., and MULRY, M. (1987). Operational standards for determining the accuracy of census results. Proceedings of the Social Statistics Section, American Statistical Association, 46-55.

HOGAN, H., and WOLTER, K. (1988). Measuring Accuracy in a Post Enumeration Survey. Survey Methodology, 14, 99-116.

MORIARITY, C. (1987). STSD 1986 Test Census Memorandum II-12, Subject: Documentation of the Calculation of the Los Angeles Post-Enumeration Survey Block Weights and Dual System Estimate Variances. Statistical Support Division, U.S. Bureau of the Census, Washington, D.C.

SCHENKER, N. (1988). Handling missing data in the estimation of coverage error for the 1986 census of Central Los Angeles County. Survey Methodology, 14, 87-97.
SEKAR, C.C., and DEMING, W.E. (1949). On a method of estimating birth and death rates and the extent of registration. Journal of the American Statistical Association, 44, 101-115.
SELTZER, W., and ADLAKHA, A. (1974). On the Effect of Errors in the Application of the Chandrasekaran-Deming Techniques. Reprint 14, University of North Carolina, Laboratory for Population Statistics.
SPENCER, B.D. (1986). Conceptual issues in measuring improvements in population estimates. Proceedings of the Second Annual Research Conference, U.S. Bureau of the Census, Washington, D.C., 393-407.

SPENCER, B.D. (1980). Implications of equity and accuracy for undercount adjustment: a decision theoretic approach. Proceedings of the 1980 Conference on Census Undercount, U.S. Bureau of the Census, Washington, D.C., 204-216.

WOLTER, K.M. (1986a). Some coverage error models for census data. Journal of the American Statistical Association, 81, 338-346.

WOLTER, K.M. (1986b). A Combined Coverage Error Model for Individuals and Housing Units. SRD Research Report Number Census/ SRD/RR-86/27, Statistical Research Division Report Series, U.S. Bureau of the Census, Washington, D.C.

# Representing Local Area Adjustments by Reweighting of Households 

ALAN M. ZASLAVSKY ${ }^{1}$


#### Abstract

Suppose that undercount rates in a census have been estimated and that block-level estimates of the undercount have been computed. It may then be desirable to create a new roster of households incorporating the estimated omissions. It is proposed here that such a roster be created by weighting the enumerated households. The household weights are constrained by linear equations representing the desired total counts of persons in each estimation class and the desired total count of households. Weights are then calculated that satisfy the constraints while making the fitted table as close as possible to the raw data. The procedure may be regarded as an extension of the standard "raking" methodology to situations where the constraints do not refer to the margins of a contingency table. Continuous as well as discrete covariates may be used in the adjustment, and it is possible to check directly whether the constraints can be satisfied. Methods are proposed for the use of weighted data for various Census purposes, and for adjustment of covariate information on characteristics of omitted households, such as income, that are not directly considered in undercount estimation.


KEY WORDS: Undercount; Raking; Local-area adjustment; Missing data.

## 1. HOUSEHOLD-LEVEL ADJUSTMENT BY WEIGHTING

A major research effort has been devoted to methods for estimation of the undercount in the 1990 Census in the United States (National Academy of Sciences 1985). In one of the primary methodologies that has been proposed, a Post Enumeration Survey (PES) would be conducted shortly after the Census in a sample of blocks. The fraction of persons in the PES who were omitted from the Census enumeration yields an estimate of Census undercoverage. Estimates of the undercount would be carried down to some geographical level (possibly the smallest geographical unit used by the Census, the block). These estimates would apply to classes formed on the basis of characteristics of persons, as well as possibly some household or block-level characteristics. The term "class" will be used henceforth to refer to estimation or adjustment classes or cells; the term "block" will refer to the smallest geographical unit for which undercount estimates are calculated. The 1980 Census found approximately one hundred million households in two to four million blocks, depending on the definitions used.

For each block, the outcome of the processes described above would be a vector of estimated undercounts, with $S$ components corresponding to the adjustment, or estimated number of persons omitted from the census in that block, from each of $S$ adjustment classes. The methods by which these estimates are arrived upon are beyond the scope of this paper. However, in our examples we shall assume that for each class within each block there is an undercount rate, expressing estimated omissions as a fraction of enumerated persons in that class and block. In this paper, the term "adjustment" refers to any process which incorporates the estimated undercount into the enumeration. The adjustment classes might be, but would

[^13]not necessarily be, the same as the post-strata formed in analysis of a Post-Enumeration Program. For forming simple marginal tabulations of persons by characteristics, this information might well be adequate. In particular, small-area counts used for various official and commercial purposes could be calculated from block totals.

However, for some purposes it would be desirable to place the added persons in households. We assume for these purposes that there is also an estimate of the number of omissions of whole households in each block. There might also be information distinguishing omissions of persons within enumerated households from those in omitted households.

If the resulting adjusted records are to be meaningful, the composition of the added households and the relationships of its individual members must be logically consistent and typical of the types of households found in that area. The term "composition" will be used to refer to the number of household members from each adjustment class. Thus, for example, a household consisting of a 20-year old white female head of household, a 75 -year-old Chinese male, and a 10 -year-old black daughter would not be a very plausible household, even if all of its members were from classes that are well represented in the block. Yet abstractly to describe these patterns and create new households that fit them is a daunting task.

## Example 1: Forming a roster of households.

Table 1 illustrates part of a census enumeration as it might appear on a microdata tape.
Table 2 represents the same roster, showing how the composition of the households might be summarized if there were only three estimation classes: (1) men over 20 years of age, (2) women over 20 years of age, and (3) children up to 20 years of age.

Table 1
A piece of a sample microdata file

| Name | Address | Sex | Age |
| :--- | :---: | :---: | ---: |
| John Smith | 328 Main Street | M | 34 |
| Mary Smith | 328 Main Street | F | 32 |
| Louise Smith | 328 Main Street | F | 7 |
| Nancy Chen | 330 Main Street | F | 62 |
| Jorge Ramirez | 332 Main Street | M | 21 |
| Juan Ramirez | 332 Main Street | M | 24 |

Table 2
Microdata file recoded by household, showing composition of households

|  | Count of persons by class |  |  |
| :--- | :---: | :---: | :---: |
| Address | Class 1 | Class 2 | Class 3 |
| 328 Main Street | 1 | 1 | 1 |
| 330 Main Street | 0 | 1 | 0 |
| 332 Main Street | 2 | 0 | 0 |

Essentially the same problem arises in many situations in which a household survey must be reweighted to match known marginal totals for various classes of individuals.

The essence of the method proposed in this paper is to assign weights to the households enumerated in the census lists for the block, so that the weighted totals of persons in each adjustment class and the weighted total number of households are precisely equal to the corresponding adjusted totals. Thus, although the weighting changes the proportionate composition of the block, all of the households are real and possess characteristics and relationships that are logically consistent and reasonable for that block. This weighting methodology is similar to the standard raking adjustment, in which the weight applied to counts in a cell of a contingency table is the adjusted count divided by the original count. The household weights are calculated after the block totals have been adjusted and will be consistent with those totals. For most Census purposes, the weighted records would be an adequate basis for forming published tables and sampled lists.

This proposal might be contrasted with imputation methods, in which undercounted units are represented by whole units added to the roster. The imputed units may be either persons or households. Although individual persons may be imputed into the block, the problem of fitting these persons into plausible households remains unsolved. Placing them in fictitious "group quarters," as was done in some tests of adjustment procedures, sidesteps this problem at the cost of creating a skewed picture of relationships in the block. Reweighting or imputation of individuals would be appropriate for residents of institutions or group homes, for whom the particular configuration of persons in the dwelling unit has no particular significance.

Another approach to imputation starts with probability models for omissions of households and of persons within households, and draws imputed households from the posterior distribution of the omissions given the enumerated households. This methodology is suited to the multiple imputation approach (Rubin 1987), in which the entire imputation process is repeated several times to represent the variability introduced by the underenumeration. However, in each block roster that is created, totals based on enumerated and imputed households would not necessarily be precisely equal to the desired adjusted totals. In this paper, our concern is with methods that give an exact fit to population estimates derived at a preceding stage.

The remaining sections of this paper develop methods for the proposed weighting adjustment. Section 2 gives a mathematical formulation of the objectives of the weighting scheme, while Section 3 explains how to fit the weights. Section 4 explains how to incorporate the distinction between omissions in enumerated and omitted households into the scheme. Section 5 introduces some refinements that improve the robustness of the procedure against the variability of small blocks. Section 6 describes simulation results. Section 7 discusses the use of weighted data for various Census purposes, while Section 8 considers the effects of the weighting adjustment on covariates that are not part of the scheme used in forming the adjustment classes. Finally, Section 9 summarizes some unresolved questions and areas for future research.

## 2. OBJECTIVES AND MATHEMATICAL FORMULATION OF A WEIGHTING PLAN

It is an essential goal of the proposed plan that the population of the block be assigned to valid household units, so that statistics for which the unit is the household are unambiguously defined. Thus, weights are assigned to households; the same weights apply to every person within the household.

In order that the counts in the weighted roster be those which are given by the predetermined adjustment, the following constraints must be satisfied:
(A1) Within each block, the sum of household weights equals the adjusted number of households.
(A2) Within each adjustment class and each block, the sum of weights for persons equals the adjusted number of persons.
In order that the weighted block roster be as similar as possible to the original block roster, we further require that:
(B) The weights should be, in some sense, as close to each other as possible.

With unit (or equal) weights, the composition of the block remains unchanged. If the weights are not very unequal, the census composition of the block is nearly preserved by the weighting scheme. To the extent that information about the undercount does not require a drastic revision of our view of the makeup of the block such a drastic revision should be avoided, consistently with good survey practise regarding weights.

We now turn to the mathematical formulation of these criteria. Suppose that in the block under consideration, there are $S$ adjustment classes and $I$ enumerated households, and household $i$ contains $C_{i s}$ members from class $s$. Suppose that $H$ is the desired total number of households in the adjusted roster for the block and $D_{s}$ is the desired total number of persons in class $s$. Let $\mathrm{W}_{\mathrm{i}}, i=1,2, \ldots I$, be the weights corresponding to the households. (Al) requires that

$$
\begin{equation*}
\sum_{i=1}^{l} W_{i}=H \tag{1}
\end{equation*}
$$

and (A2) requires that

$$
\begin{equation*}
\sum_{i=1}^{l} W_{i} C_{i s}=D_{s}, s=1,2 \ldots S \tag{2}
\end{equation*}
$$

These constraints can be represented by a matrix equation of the form $A W=B$, where
and 1 is a row of 1 's.
Objective (B) is represented by selecting some objective function that represents the distance between the weights $W$ and uniform weighting, and minimizing it. We will use the objective function $T=\sum W_{i} \log \left(W_{i}\right)$. This measure is proportional to the discriminant information (Kullback-Liebler information) of the discrete probability distribution (over households) with relative weights $W_{i}$ with respect to the probability distribution with equal weights, and is the same objective function that underlies the traditional "raking" (iterative proportional fitting) procedure for adjusting contingency tables (Deming and Stephan 1940; Ireland and Kullback 1968; Oh and Scheuren 1978 have a larger bibliography). Thus, our procedure may be regarded as an extension of raking. Scheuren (1973) applies raking to reweighting of households; Cilke and Wyscarver (1988) reweight to linear constraints but use a different objective function than
those considered here. Methods similar to those presented here were developed independently by Alexander (1987).

In the context of raking, initial counts $X$ are given for cells in a contingency table, and new cell counts $Y$ are calculated to minimize the objective function $\sum Y_{i} \log \left(Y_{i} / X_{i}\right)$. Then the weights of the original observations are the ratios $W_{i}=Y_{i} / X_{i}$. In our context, if $X_{i}$ households happened to have exactly the same composition we could regard them, in the same way, as forming a single entry in the roster with initial count $X_{i}$ and fit an adjusted count $Y_{i}$. However, with a large number of adjustment classes, it would be unusual for several households in the same block to have exactly the same composition. Thus we will not attempt to group households; rather, it is notationally and computationally simpler to list the households separately so that for each enumerated household composition the initial count $X_{l}=1$ and $Y_{I}=W_{l}$. Aside from this notational difference, the mathematical formulation here differs from that of a raking adjustment only in that the linear constraints do not have the special structure of margins in a contingency table. For brevity in the presentation of examples, we will sometimes include a count on a line to represent that number of identical lines in the roster of households.

In the contingency table setting, raking preserves cross-product ratios of cells, and preserves independence of variables when it holds in the original table. For these reasons, it has been called "structure-preserving estimation" in small-area estimation applications (Purcell 1979; Purcell and Kish 1979). See Section 10.1 for a further discussion of objective functions.

Our procedure differs from raking in that the linear constraints do not necessarily refer to margins in a contingency table. Our methodology includes raking as a special case, as well as the raking generalization of Oh and Scheuren (1978) in which different tables are used to fit each margin. In fact, constraints may be imposed on continuous as well as discrete covariates; applications of this sort are proposed in Section 8.3. Furthermore, the algorithms that are set forth allow direct determination of whether there are in fact any weights that are consistent with all of the given constraints. It is possible then to select constraints that must be relaxed in order to fit weights. These features give these methods potential applicability extending beyond the area of representing undercount.

## 3. FITTING THE WEIGHTS

The problem before us now is to determine weights satisfying the constraints $A W=B$, $W \geq 0$, minimizing the objective function $T=\sum W_{i} \log \left(W_{i}\right)$. To make $T$ a continuous function of $W$, we adopt the usual convention $0 \log 0=0$.

We will call any weight vector that satisfies the linear constraints (the equations and the inequalities) a feasible solution. As long as there is a constraint on the total weight of the households, the set of feasible solutions is bounded and therefore $T$ assumes a minimum value on it; furthermore, since $T$ is strictly convex, the solution is unique.

The problem of calculating weights then naturally is divided into three tasks: (1) determining whether the linear constraints $A W=B$ are consistent; (2) determining whether there are any feasible solutions; and (3) finding the feasible solution minimizing $T$. We will suppose that there are $I$ households and $p$ constraints, so $A$ is a $p \times I$ matrix.

## Example 2: Fitting weights.

Table 3 illustrates the roster of households in a block in which three classes are represented, as in Example 1; we may think of the classes as "men," "women," and

Table 3
A household roster

|  | Count per household by class |  |  |  |
| :--- | :---: | :---: | :---: | :---: |
| Line \# | Class 1 <br> (men) | Class 2 <br> (women) | Class 3 <br> (children) | Number of <br> households |
| $\mathbf{1}$ | 0 | 1 | 0 | 50 |
| 2 | 0 | 1 | 1 | 40 |
| 3 | 1 | 0 | 0 | 40 |
| 4 | 1 | 0 | 2 | 15 |
| 5 | 1 | 1 | 0 | 50 |
| 6 | 1 | 1 | 1 | 60 |
| 7 | 1 | 1 | 2 | 40 |

Table 4
Adjusted totals

|  | Raw <br> count | Adjustment <br> rate | Adjusted <br> count |
| :--- | :---: | :---: | :---: |
| Class 1 | 205 | .05 | 215 |
| Class 2 | 240 | .03 | 247 |
| Class 3 | 210 | .04 | 218 |
| Households | 295 | .02 | 301 |

and "children." This table may be regarded as a condensed version of a table with 295 lines, each representing one household.

The unadjusted and adjusted counts of households and of persons in each class are found in Table 4. The adjusted counts are calculated by applying the listed adjustment rates and rounding. The method by which the adjusted counts are obtained is immaterial, however, to the rest of the process.

### 3.1 Consistency of Linear Constraints

As long as the rows of $A$ are independent, the constraints $A W=B$ will be consistent. If any row is dependent on the others, the corresponding constraint is either inconsistent or redundant, depending on the values in $B$. Dependent rows can be identified by applying the $Q-R$ decomposition to $A^{\prime}$. If the corresponding constraints are redundant, they may be deleted without any loss of information; if they are inconsistent, the constraints must be reformulated in some way.

Example 2: (continued).
The $A$ matrix for this example has independent rows, and hence the constraints are consistent.
In Section 5, we consider circumstances in which inconsistent constraints are likely to appear and some methods for dealing with them.

### 3.2 Existence of Feasible Solutions

Determining the existence of feasible solutions is equivalent to determining an initial feasible solution in a linear programming problem, and the standard algorithms can be used. Suppose
our problem is to find a positive solution $W$ to $A W=B$, where $B \geq 0$. (If the latter condition does not hold it can be made true by reversing the sign of negative elements of $B$ and the corresponding rows in $A$.) Then create an augmented problem $[A \mid I]\left[W^{\prime} \mid Z^{\prime}\right]^{\prime}=$ $B, W, Z \geq 0$, where $I$ is a $p \times p$ identity matrix and $Z$ is a $p$ element vector variable. This problem automatically has an initial solution $W=0, Z=B$. Then apply the simplex method (as in Gass (1964) or any other linear programming text) to minimize $\sum Z_{i}$. If that sum can be reduced to 0 , the corresponding $W$ values are a solution to the original problem, while if it cannot, the original problem has no solution.

## Example 2: (continued).

A feasible (but not optimal) solution for this example gives total weighted counts of $86,54,29$, and 132 to the household compositions in lines $2,3,5$, and 6 respectively of Table 3. It may be verified that these counts yield the desired adjusted totals for households and for individuals in each class.

The problem of infeasibility is similar to that of inconsistency and is also discussed in Section 5.

### 3.3 Optimizing the Objective Function.

By the method of Lagrange multipliers, the minimizing solution must satisfy the equations $\partial T / \partial W_{i}=\log W_{i}+1=a_{i}{ }^{\prime} \lambda$, where $a_{i}$ is the $i$-th column of $A$ and $\lambda^{\prime}=\left(\lambda_{1}, \lambda_{2}, \ldots \lambda_{p}\right)$. Then $W_{i}=\exp \left(a_{i}^{\prime} \lambda-1\right)$; thus the model for the weights is log-linear in form, like that for a conventional raking adjustment. $\lambda_{s}$ represents the additional log-weight increment associated with a unit increment in the corresponding constraint coefficient $a_{i s}$, i.e. adding an additional household member from adjustment class $s$ to the household.

We can solve for $\lambda$ by Newton's method to satisfy $A W=B$. The iterative scheme we use is

$$
\begin{equation*}
\lambda^{(t+1)}=\lambda^{(t)}-\left(A W^{*} A^{\prime}\right)^{-1}(A W-B), \tag{4}
\end{equation*}
$$

where $W^{*}$ is the matrix with the elements of $W=W\left(\lambda^{(t)}\right)$ on the diagonal. A good starting value for $\lambda$ is $\lambda^{(0)}=\left(A A^{\prime}\right)^{-1} B$, which can be derived from a linear approximation around equal starting weights. A cyclic descent procedure for solving these equations, which is a generalization of iterative proportional fitting, is described in Section 10.2.

Example 2: (continued).
The weights per household and total weighted counts (weight times raw count) for each line in Table 3 are shown in Table 5. No household is upweighted by more than $8 \%$ or downweighted by more than $5 \%$.

Table 5
Optimal weights for Example 2

| Line \# | Weight | Weighted <br> counts |
| :---: | :---: | :---: |
| 1 | 0.9554 | 47.77 |
| 2 | 0.9557 | 38.23 |
| 3 | 0.9816 | 39.27 |
| 4 | 0.9823 | 14.73 |
| 5 | 1.0730 | 53.65 |
| 6 | 1.0734 | 64.40 |
| 7 | 1.0737 | 42.95 |

## 4. WHOLE-AND WITHIN-HOUSEHOLD ADJUSTMENTS

We now consider the distinction between within-household adjustments (that is, adjustments for omissions of persons within enumerated households) and whole-household adjustments (that is, adjustments for omissions of whole households). This distinction has previously been made for purposes of analysing the causes of undercount (Fay 1986). Our concern here is to use it to more accurately represent the undercount by an adjustment.

Within-household adjustments do not involve adding any households to the roster, but only shifting weight between households to increase the weighted totals of persons in the various classes. That is, households with few or no persons in a particular class are downweighted and those with many are upweighted, so that the total household weight remains constant. Thus, in this portion of the adjustment, some households will inevitably have their weights reduced. Whole-household adjustments, on the other hand, correspond to households that were omitted entirely from the census. These adjustments do not reflect on the accuracy of the enumerated households; thus they should be represented by adding households to the roster without taking weight away from the households that were enumerated.

We propose to separate these two portions of the adjustment. One set of constraints represents the within-household adjustment. The total household weights are here constrained to equal the enumerated count of households, while the total weights assigned to persons in each class are constrained to equal the enumerated count in that class plus the within-household adjustment for that class. $A W_{1}=B_{1}$ where $B_{1}$ consists of the enumerated household count and the counts of persons by class adjusted for within-household undercount.

A second set of constraints represents the whole-household adjustment. The total household weights are here constrained to equal the estimated omitted households, and the total person weights in each class are constrained to equal the estimated omitted persons in those households. $A W_{2}=B_{2}$ where $B_{2}$ consists of the count of added households and the counts of added persons by class for the adjustment for whole-household undercount.

After fitting two sets of weights corresponding to the two sets of constraints, the two weights for each household are added to obtain weights that incorporate both parts of the adjustment ( $W=W_{1}+W_{2}$ ). The distinction between whole- and within-household adjustments contains information which may lead to a different set of adjusted weights than would be calculated if the adjustments were combined, as is illustrated in Example 3. However, if this distinction is not made in the estimation of the undercount, an adjustment can still be calculated in a single step.

## Example 3: adjustments for whole-household omissions.

Suppose there are only two adjustment classes, and a hypothetical block has the composition described in the first three columns of Table 6.

Suppose now that to the 30,010 households enumerated, we must add 231 persons each in Class 1 and Class 2, and 121 households. The last three columns of Table 6 show the adjusted counts under alternative assumptions: (1) the omitted persons may belong to any household, enumerated or omitted, and (2) all of the omitted persons were in the omitted households.

When the omitted persons could have been in any household, the algorithm downweights the households with only one person from each class $(1,1)$ and upweights households with two from one class and one from the other ( 1,2 and 2,1 ). While the households with two persons from each class are substantially upweighted (by a factor of 1.354 ), only a small portion of the added persons appear in those households since

Table 6
Hypothetical raw and adjusted household counts for Example 3

| Household composition |  | Raw count (number of households) | (1) Omitted persons in any household | (2) Omitted persons in omitted households only |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Class 1 persons | Class 2 persons |  | Adjusted counts | Counts of omitted households | Adjusted totals, omitted and enumerated households |
| 1 |  | 10,000 | 9904.54 | . 01 | 10,000.01 |
| 1 | 2 | 10,000 | 10106.46 | 10.99 | 10,010.99 |
| 2 | 1 | 10,000 | 10106.46 | 10.99 | 10,010.99 |
| 2 | 2 | 10 | 13.54 | 99.01 | 109.01 |

the original count for that composition is so small.
When the omitted persons appear only in the omitted households, weights are calculated first to fit $231 \times 2=462$ persons into 121 additional households, and then these weights are added to the unit weights in the raw counts. While no household composition is downweighted, the $(2,2)$ households are upweighted extremely (by a factor of 10.901 ). In fact, it is mathematically impossible to accommodate 462 persons in 121 households of two to four persons each without having at least 99 households with 4 members. Thus, the information that the added persons (or some known fraction of them) belong in the omitted households substantially changes our view of the appropriate adjustment.

## 5. FEASIBILITY OF CONSTRAINTS

In the preceding sections we have assumed that feasible solutions exist to the constrained optimization problem. Here we will consider situations in which the solutions will not exist or will be unsatisfactory, and some alternative methods to deal with these situations.

### 5.1 When Will Constraints be Non-feasible?

There are three ways in which the constraints may fail to allow of satisfactory solutions: (1) when the constraints are actually inconsistent, (2) when the constraints are consistent but there are no positive weights that satisfy them, and (3) when there is a feasible solution but it involves an extreme adjustment to some household weights. The issues associated with these three failure modes are fairly similar.

One could write down constraints that are intrinsically inconsistent, for example that all classes of men are adjusted upward by $2 \%$ while men in total are adjusted upward by $4 \%$. In our procedure each constraint applies to the number of persons in a distinct adjustment class and so there are no inconsistencies of this sort. However, a contingent inconsistency is still possible, that is to say one that depends on the particular collection of household compositions that appears in a block. The following are examples of contingent inconsistency, infeasibility, or unsatisfactory weights:
(1) Proposed undercount estimation methods envision defining over 100 adjustment classes. In a small but diverse block the number of classes represented might be larger than the
number of households; hence the number of constraints would be larger than the number of weights to be fitted. An inconsistency is then almost inevitable.
(2) If all households in the block have exactly the same number of members from a particular adjustment class (e.g. every household has one young Hispanic girl), then the number of members of this class represented is unaffected by the distribution of weights.
(3) The adjustment of the number of households may be too large or small to accommodate the adjustment of persons in some class. This may represent a failure of the model for adjustment of the number of households. For example, suppose that the number of men to be added by the whole-household adjustment is greater than the number of households to be added, but no household in the block has more than one man. The constraints then might be consistent but infeasible, since they could be satisfied only by assigning negative weights to some households without men.
(4) The block may have had omission rates atypical of blocks in the PES on which omission rates were estimated. For example, suppose that in most blocks (including most of the PES sample blocks), adult males with certain characteristics tend to be heavily undercounted, but the block being adjusted is atypical in having adult males of this class present in most households and well counted. The class undercount estimate might lead to an extreme upward adjustment that could not be accommodated within the existing households.
(5) Some adjustment may require giving substantial additional weight to households containing persons from a combination of adjustment classes that appears in only one household, so that household receives an extreme weight. In this case the problem is feasible but the solution is not very satisfactory.
Problems of infeasibility may also arise where the difficulty cannot be so easily traced to a particular inconsistency in the adjustment.

### 5.2 Making the Constraints Feasible

Regardless of the stage of the fitting procedure at which the infeasibility is discovered, several methods are available to relax the constraints and make them feasible. In this section, we survey several such methods, drawing out both the intuitive logic of each choice and the computational methods required.

### 5.2.1 Methods Based on Dropping Rows (constraints) of $\boldsymbol{A}$

When checking for consistency of constraints, some rows may be found to be linearly dependent on the previous rows and hence either redundant or inconsistent. If these rows are simply dropped from the $A$ matrix, a consistent set of constraints is obtained; thus, no further computational effort is required.

If the constraints are arranged in sequence from the most important to the least important, than the less important constraints will be dropped when they are inconsistent with the more important ones. This ordering makes the most sense if the original constraints on distinct adjustment classes (defined by a multi-way classification of the population) are reframed in an ANOVA-like manner as constraints on total population ('grand mean''), classes defined by one classification variable ("main effects"), and classes defined by interactions. For example, if there are ten adjustment classes defined by two sexes and five age ranges, the reframed constraints in order of importance might be: total population ( 1 constraint), population by sex ( 1 more constraint), population by age ( 4 more constraints), age-sex interactions (the remaining 4 constraints). The 4 age constraints could be further broken down as old-vs.-young ( 1 constraint) and 3 further constraints within those larger groups.

A similar procedure can be applied at the stage of checking feasibility of the constraints. If it is not possible to make all of the $Z_{i}=0$, the objective function in the linear programming problem can be modified to be $\sum c_{i} Z_{i}$, with the coefficients $c_{i}>0$ corresponding to the most important constraints made larger. Then a maximal set of feasible constraints can be identified, and the remaining constraints dropped.

The outcome of this procedure would be weights that give the correct block totals on the coarser classifications of persons, while failing to be correct on all cross-tabulations.

### 5.2.2 Methods Based on Adding Columns (households) to $\boldsymbol{A}$

When constraints are only contingently infeasible (in the previous sense that infeasibility depends on the particular set of household compositions in the block), they become feasible when households are added that have the required composition. The simplest application of this principle is to work at a higher level of geographical aggregation than a block. A few adjacent blocks may be combined when problems arise in fitting, or the entire roster may be grouped at, for example, the enumeration district level before weighting. The larger the unit, the broader the range of household compositions that will be represented and the less likely that problems of infeasibility will arise.

A more sophisticated procedure would use a hot-deck of households from adjacent "donor" blocks to enrich the pool of households to which weight can be assigned. Computational simplicity is important here since it may be necessary to scan through a long list of households to find the one or ones which will make the constraints feasible. In the consistency-checking stage, if row $j$ of $A$ is dependent on the previous rows, then if the column for the added household is independent of the columns of $A$ (with regard only to the first $j$ rows), row $j$ of the augmented $A$ will be independent. In the stage of checking for feasibility, if the algorithm halts because no reduction can be made in the objective function $\sum Z_{i}$, the search for basic columns can be extended to columns corresponding to households in the hot deck. Finally, if some household's fitted weight is extremely high, the hot deck can be scanned for other households that would also receive high weights with the current values of $\lambda$ (that is, columns $a$ such that $a^{\prime} \lambda$ is large). If these are added to the block they will draw off some of the weight from the overweighted households when the weights are refitted, since they are likely to also have members in the same adjustment classes.

The intuition behind this method is that the household compositions that are enumerated in a block are only a sample of those which actually could have appeared there had the enumeration been complete. The observed distribution of household compositions is smoothed by mixing it with the distribution for adjacent blocks, which contain households that are also typical for that area. Thus, conceptually this method is related to Bayesian smoothing methods that improve estimation of some quantity for one unit by borrowing strength from its distribution in similar units. This Bayesian rationale is developed in terms of a block-level randomeffects model by Zaslavsky (1989).

The donor blocks could be chosen by a sequential hot deck procedure; then, the donor blocks would tend to be geographically close to the adjustment block and no particular set of blocks would have undue influence on the entire census. By detailed stratification of blocks, the donor blocks could be selected to be similar to the block being adjusted on characteristics such as mean income, types of housing units, and racial balance.

### 5.2.3 Combined Methods

The two types of methods outlined above can be combined by an appropriate reframing of constraints. The principle here is to satisfy all constraints in the larger geographical units,
while satisfying only the more important constraints in the smaller units. This type of compromise may make it possible to get a fairly good fit to the desired distribution without having to add additional records to the roster.

Suppose that the $A$ matrices for several blocks have been reframed similarly as sequences of rows representing main and interaction constraints. Then a single large $A$ matrix representing all of the constraints can be formed. The rows for the more important constraints can be kept separate, while rows for subsidiary constraints can be combined across blocks. For example, suppose there are ten adjustment classes, defined by sex ( 2 levels) and age ( 5 levels), and two blocks. Altogether there are eleven constraints (one for number of households and one for each adjustment class) in each block. If these are combined into a single matrix, keeping main effects and two-way interactions, the constraints are: block household counts ( 2 constraints), block populations ( 2 constraints), sex ( 1 constraint), age ( 4 constraints), block $\times$ sex interaction ( 1 constraint), block $\times$ age interaction ( 4 constraints), and sex $\times$ age interaction ( 4 constraints) in the combined blocks. Here 4 constraints have been eliminated (block $\times$ sex $\times$ age interaction); in a more realistic problem with more blocks, classification variables, and levels, the reduction would be much greater.

## 6. SIMULATION RESULTS

Simulations were performed to answer two classes of questions:
(1) The first set of questions is concerned with evaluation of the success of the algorithm in terms of its own constraints and objectives. Does the reweighting algorithm give an answer? In real problems, is there a solution to the weighting constraints? How much do the weights vary? Is the amount of computation required within reasonable limits?

To answer these questions, "feasibility simulations" were performed in which the weighting algorithm was applied to simulated blocks made up of real households, using real adjustment rates. This procedure thus closely parallels the practical application of the algorithm.
(2) The second set of questions is concerned with evaluation of the success of the algorithm in improving the quality of inferences based on a micro-data set: does the weighted microdata set more accurately describe the real world than the raw, unweighted data?

To answer these questions, simulated blocks made up of real households were drawn, representing the true (but unobserved) compositions of households in blocks. For each "true" block, omissions were imposed using real estimated undercount rates and a plausible model for the distribution of undercount among households. The weighting algorithm was applied to the "enumerated" blocks generated in this way. Summary statistics describing household composition were calculated for the simulated "true" blocks and for the simulated observed blocks with undercount, both unweighted and weighted for undercount adjustment. The goal of these "inference simulations" was to determine whether the reweighting brought the statistics closer to their values in the "true" blocks; in other words, did reweighting correct the biases caused by the undercount?

The source of households for all simulations was the $1 \%$ ' $B$ ' Public Use Microdata Sample (PUMS) from the 1980 Census (Bureau of the Census 1985). Households were extracted from sections of Los Angeles County, California that include the site of the Test of Adjustment Related Operations (TARO) of the 1986 Test Census.

Undercount rates were those calculated from the 1986 TARO (Diffendal 1988, Table 7) for adjustment classes defined by sex, age (five levels), race (Hispanic, Asian, or 'other race''),
and tenure (owner or renter). Adjustment factors calculated from the given undercount rates ranged from 0.982 to 1.211 .

Each household was coded as a vector of counts representing the number of individuals in that household from each of the 60 adjustment classes.

Further details on the simulation procedures and on a larger set of simulations are in Zaslavsky (1989).

### 6.1 Feasibility Simulations

For each of four block sizes ( $20,50,100$, and 200 households), 50 simulated blocks were drawn from the full sample and 50 were drawn from only those households with no Asian members. For each block, simulations were attempted using two levels of the household adjustment rate (the factor by which the number of households in the block is adjusted).

The algorithms of Section 3 were applied. To recapitulate, the linear constraints were checked first for consistency, and then for feasibility (existence of a positive solution); finally, weights were calculated using Newton's method. As no data were available distinguishing withinhousehold and whole-household omissions, no effort was made to separate them in these or other simulations.

The results of these simulations are summarized in Table 7.

## Consistency and feasibility:

The columns headed "incons", "infeas", and "OK" represent the number of simulated blocks (out of the 50 trials) in each simulation that fell into each of the following categories respectively: (1) the constraints were inconsistent (could not be satisfied by any weights), (2) the constraints were consistent but not feasible (could not be satisfied by any positive weights), or (3) the constraints were both consistent and feasible.

In the "non-Asian"' simulations there are 41 constraints to be satisfied (some of which may be trivial, i.e. when the corresponding adjustment classes are unrepresented in the block). Thus with 20 -household blocks, the constraints were never consistent; with 50 -household blocks, the constraints were sometimes consistent and then usually feasible. The constraints were usually feasible in 100 -household blocks, and always in 200 -household blocks.

The numbered columns at the right represent the order of the simplest marginal constraint that could not be satisfied, in the sense of the heirarchical reparametrization in Section 5.2.1. Thus, column (1) indicates the number of simulated blocks for which a "main effect" constraint (marginal total of persons classified by one stratifying variable) could not be satisfied, column (2) indicates the number of trials for which a two-way interaction constraint could not be satisfied, etc. Even when the constraints were inconsistent with 50 - or 100 -household blocks, the main-effect constraints and often the two-way or even three-way interactions were feasible. This suggests that pooling of blocks for higher-order interactions, as described in Section 5.2.3, might be a successful strategy for dealing with problems of infeasibility.

The results were less encouraging for simulations using the full samples. Even with 200household blocks, only rarely were the constraints consistent and feasible. With increasing block size the lower-order constraints were more likely to be feasible. This is explained by the small number of households with Asian members (approximately $5 \%$ in each sample). Out of 200 households, the expected number of Asian households would be about 10, an insufficient number to satisfy the 20 possible constraints for the Asian adjustment classes. Such a situation in which some groups of adjustment classes are poorly represented in a certain region or in particular blocks would surely not be unusual in practise. This would require pooling of blocks on a large scale for the corresponding constraints, while the constraints for the better-

Table 7
Feasibility simulation results

| Non-Asian Households |  |  |  |  |  |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| size | HH | rate | incons | infeas | OK | maxW | minW | varW | iters | (1) | (2) | (3) | (4) |
| 10 |  | 1.00 | 50 |  | 0 | NA | NA | NA | NA | 22 | 28 | 0 | 0 |
| 10 |  | 1.05 | 50 | 0 | 0 | NA | NA | NA | NA | 8 | 42 | 0 | 0 |
| 20 |  | 1.00 | 50 | 0 | 0 | NA | NA | NA | NA | 0 | 50 | 0 | 0 |
| 20 |  | 1.05 | 50 | 0 | 0 | NA | NA | NA | NA | 0 | 50 | 0 | 0 |
| 50 |  | 1.00 | 47 | 1 | 2 | 1.921 | 0.200 | 0.142 | 3.00 | 0 | 3 | 37 | 8 |
| 50 |  | 1.05 | 47 | 0 | 3 | 1.550 | 0.620 | 0.036 | 1.33 | 0 | 3 | 36 | 8 |
| 100 |  | 1.00 | 10 | 0 | 40 | 2.068 | 0.429 | 0.088 | 2.03 | 0 | 0 | 8 | 2 |
| 100 |  | 1.05 | 10 | 0 | 40 | 1.573 | 0.753 | 0.020 | 1.90 | 0 | 0 | 8 | 2 |
| 200 |  | 1.00 | 0 | 0 | 50 | 2.434 | 0.543 | 0.063 | 2.18 | 0 | 0 | 0 | 0 |
| 200 |  | 1.05 | 0 | 0 | 50 | 1.749 | 0.821 | 0.015 | 2.00 | 0 | 0 | 0 | 0 |

Full Sample

| size | HH | rate | incons | infeas | OK | $\operatorname{maxW}$ | $\min W$ | varW | iters | (1) | (2) | (3) |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 100 |  | 1.00 | 49 | 0 | 1 | -- | -- | -- | -- | 0 | 34 | 15 |
| 200 |  | 1.00 | 49 | 0 | 1 | -- | -- | -- | -- | 0 | 2 | 43 |

represented classes might be satisfied on a smaller scale.

## Weights:

The maximum and minimum household weights and the variance of the weights were calculated for each simulated block for which the constraints were consistent and feasible. For each simulation condition, the average value of these quantities (across simulated blocks) is displayed under the heads "maxW", "minW", and "varW." The following observations characterize some of the effects of the simulation design factors on the fitted weights.
(1) For simulations with household count adjustment factor of 1.05 , in every case, the average variance of the weights was smaller, and the average of the minimum weights and of the maximum weights were closer to unity, than with household adjustment factor 1 . This is intuitively reasonable since almost all class adjustment factors exceed 1 , and it requires a more extreme adjustment to add individuals to existing households than to add individuals and households to accommodate them. For example, if the adjustment factors for households and for every adjustment class are all equal, every household would be upweighted equally.
(2) Fixing other factors, the variance of the weights becomes smaller as the number of households per block increases. Again, this is intuitively reasonable because the pool of households is richer in a larger block; the probability of finding exactly the households needed to represent undercounted individuals is higher. The trends for the extreme weights are less clear-cut than for variances; here, the narrowing of the variance is offset by the larger sample over which the extreme is calculated in the larger blocks.
(3) The average variances for simulations with 200 -household blocks were at most .063 . Thus the reweighting is generally not extreme.

## Computational costs:

The mean number of Newton steps required to fit the weights (from the starting values given in Section 3.3), shown under the heading "iters", is usually about two. These iterations were sufficient to satisfy all constraints with error no greater than .001 . Using this information, a rough estimate can be given of the number of floating point operations required to apply the algorithm. Computational costs of the modified raking algorithm are discussed in Section 10.2.

Assume that blocks are of sufficient size that it is not necessary to check consistency and feasibility of the constraints in every case (but perhaps only when the weight fitting does not succeed in a few steps). Then the key calculation is fitting the weights. For production runs, data structures and programs should be devised which take advantage of the sparseness of the $A$ matrix (due to the fact that only a few classes are represented in each household). Then if $S_{1}$ is the total number of nonzero entries in $A$ and $S_{2}$ is the sum (through the block) of the squares of the number of nonzero entries for each household, each Newton step requires about $5 S_{1} / 2+S_{2} / 2$ multiplications (plus a term independent of the number of households per block). In the samples studied here, $S_{2} \approx 5 S_{1} ; S_{1}$ is bounded by the total population of the block. Thus the bound on the number of multiplications is approximately $15 \times$ population total (counting the start as an iteration); the number of additions is comparable.

In an era in which even microcomputers have megaflop arithmetic capability, $8 \times 10^{9}$ floating point operations to reweight an entire census does not seem unreasonable. The calculation of weights might well take less computer resources than the "bookkeeping" data processing required in any method of incorporating undercount. Of course, if the procedure were applied to a sampled database, as in forming a public-use sample, the costs would be reduced correspondingly.

### 6.2 Inference Simulations

For the inference simulations, pseudo-blocks of 50 households each with only Hispanic members were drawn. These were treated as if they represented true blocks. Then simulated omissions were imposed on the these households, assuming that each member was (independently) omitted with probability equal to the undercount rate from Diffendal (1988), with two negative undercount rates truncated to 0 .

The entire distribution of the "enumerated" block was represented by including in the pseudo-Census roster the true composition and the possible compositions obtained by omission of one or more household members, each weighted by its probability under the model.

The pseudo-Census roster with undercount was then reweighted to the original pseudo-block totals for number of households and of individuals in each adjustment class. Both the pseudoCensus roster and the reweighted roster were compared to the original pseudo-block.

The purpose of organizing the simulation in this manner was to remove variability due to randomness in the rate of omissions in a block (around the mean undercount rate) and in the distribution of the omissions among the households in the block. Furthermore, feasibility is guaranteed because the original households are always included (with weights) in the pseudoCensus roster. One way of regarding this setup is that each simulated block represents a very large population in which observed undercount rates and the distribution of observed compositions approach their expectations.

Several sets of statistics were used in evaluation of the reweighting procedure. These were all chosen because they summarized household characteristics that are not functions of the populations by adjustment class. The first set was the distribution of sizes (number of members) of households. Note that the mean number of persons per household, like any function of the class totals and household count, will automatically be adjusted to the correct (pre-undercount) values; the distribution of sizes, however, is not controlled by the adjustment procedure.

The second set of statistics was the distribution of number of adult (over 14 years old) members in households with one or more children (up to 14 years old). In this case, the mean is not automatically adjusted to the correct value, since it depends on the joint distribution of counts from different classes within households as well as on marginal totals.

The last two sets of statistics were the distribution of the age group (coded from 1 to 5 as in the formation of the adjustment classes) of the oldest male in the household (coded 0 if no male is present), and the same distribution for households with one or more children. Again, neither the distribution nor its mean are directly constrained to their true values.

The results of these simulations are summarized in Table 8. Because almost all of the differences noted here are highly significant (relative to between-pseudo-block variances of the differences), standard errors are not shown in the tables. The lines of each table are labelled "true" (for the original pseudo-blocks), "enum" (for the simulated enumerated blocks, i.e. after omissions due to undercount), and "adjust"' (enumerated blocks after adjustment for undercount). Every column except the means should be read as a percentage of households in the block.

Table 8
Inference simulation results

| Size distribution |  |  |  |  |  |  |
| :--- | ---: | :---: | :---: | :---: | :---: | :---: |
|  | size 1 | size 2 | size 3 | size 4 | size $5+$ | mean |
| true | 7.240 | 16.200 | 20.240 | 22.600 | 33.720 | 3.971 |
| enum | 10.349 | 19.631 | 21.772 | 20.690 | 27.558 | 3.632 |
| adjust | 7.372 | 16.421 | 20.596 | 21.392 | 34.219 | 3.971 |

Size distribution (number of adults) for households with children

|  | size 0 | size 1 | size 2 | size 3 | size 4 | size $5+$ | mean |
| :--- | ---: | ---: | :---: | :---: | :---: | :---: | :---: |
| true | 0.000 | 6.925 | 58.404 | 17.214 | 9.125 | 8.332 | 2.585 |
| enum | 1.736 | 18.309 | 49.874 | 15.965 | 7.677 | 6.439 | 2.323 |
| adjust | 0.924 | 13.277 | 48.557 | 18.223 | 9.810 | 9.209 | 2.562 |

Age of oldest male (by five age classifications)

|  | none | age 1 | age 2 | age 3 | age 4 | age 5 | mean |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: | ---: |
| true | 7.080 | 4.000 | 28.680 | 33.800 | 21.960 | 4.480 | 2.730 |
| enum | 9.981 | 7.388 | 26.296 | 30.972 | 21.160 | 4.203 | 2.585 |
| adjust | 7.853 | 5.989 | 26.307 | 33.439 | 21.931 | 4.480 | 2.690 |

Age of oldest male (by five age classifications) for households with children

|  | none | age 1 | age 2 | age 3 | age 4 | age 5 | mean |
| :--- | :---: | ---: | :---: | :---: | :---: | :---: | :---: |
| true | 3.602 | 6.214 | 30.744 | 42.649 | 15.843 | 0.949 | 2.638 |
| enum | 5.809 | 11.723 | 27.321 | 39.096 | 15.158 | 0.894 | 2.488 |
| adjust | 4.272 | 9.069 | 27.242 | 42.038 | 16.418 | 0.962 | 2.601 |

Household size distribution was biased downwards in the enumerated blocks. As well as correcting the mean, adjustment brought the estimated percentage for every size substantially closer to the true percentage.

The distribution of number of adults in households with children was also biased downwards. The majority of these households had contained two adults, so this size category was most understated by the enumerated statistics. Due to the log-linear structure of the adjustment, however, the most extreme adjustments were made to the largest and smallest households. Thus, the highest size categories were slightly overadjusted and intermediate categories were underadjusted; the "size 2 " category was adjusted a small amount in the wrong direction. Nonetheless, the mean of the adjusted distribution was much closer to the "true" value than the adjusted mean was.

The story is similar for the distributions of age of oldest male. Although these statistics are only indirectly related to the counts by class, in almost every case the adjusted distributions and means are closer to the "truth" than are the unadjusted distributions and means.

In summary, these simulations suggest that these weighting adjustments can improve estimates of measures of household structure as well as the aggregate counts for which they were intended. However, reweighting does not provide accurate adjustments with certain configurations of the data, such as the many households with two adults noted above; to deal with these situations may require a model-based imputation method such as that outlined by Zaslavsky (1989).

## 7. THE USE OF WEIGHTED DATA

The product of the methods of the preceding sections would be a census roster in which households have weights, persons in households have weights adopted from their households, and institutionalized persons have individually assigned weights. This section outlines the use of these rosters for various Census purposes.

### 7.1 Formation of Tables of Counts

As with any data set of weighted observations, the sum of weights replaces the simple count of observations in forming tables. The only problem created by the use of weights is that of obtaining integer entries in the tables. This problem arises even before the calculation of household weights: when the estimated omissions are calculated, the counts in each class will not in general be integers.

If the adjusted totals by class are rounded to be integers, any table that aggregates classes (for example, a count of adult males that is a sum of counts of adult males from different classes) will also contain integers, since it must be consistent with those totals. For tables that are not based on those totals, summing the weights in a particular group may not necessarily generate integer counts. For example, if a class combines women of ages $20-40$, a sum of weights for women aged $20-30$ would not necessarily be an integer. In any case, it seems unlikely that all class weights would be rounded since this might well lose the entire adjustment to roundoff error. However, it should be possible to use existing Census Bureau integerizing methods ("controlled rounding'') to deal with these problems, especially where non-disclosure requires that published counts be rounded anyway (Cox et al. 1986; Cox 1987).

### 7.2 Formation of tables of sums and means

Generally, sums (of continuous quantities) and means are not expected to be integers, so
the issue of rounding does not arise. Also, tables based on long-form information are already derived from a sample so an additional source of weights should not change the process much. A deeper issue is that of the values of non-classification covariates to be assigned to households that are "weighted in" to the census; this is discussed in Section 8.

### 7.3 Public Use Samples

The public use tapes are a sample of census records that are released for further analysis by consumers of census data.

To generate these samples from weighted census rosters requires only that the sampling procedure be modified slightly to make sampling probabilities proportional to weights. Even on the $5 \%$ tape (the highest sampling rate), the weighted sampling probabilities should be smaller than 1 . Once these tapes are produced, the user would not have to be aware of the adjustment and weighting process that had gone into generating them.

The public use tapes are the source of data for many of the more complicated analyses by sociologists, economists, planners, etc. in which the details of household composition, as well as counts of persons, are of importance. It is important that these tapes could be generated easily and used like raw census data.

As a service to those users of the public use tapes who wish to check the sensitivity of their analyses to the undercount adjustment, the tape should include factors (the inverse of the adjustment weights attached to the household records in the original census rosters) that would allow the user to reconstruct the equivalent of the unadjusted census.

## 8. ADJUSTMENT OF COVARIATES THAT ARE NOT USED IN CLASSIFICATION

The methods described above guarantee that weighted block totals by variables used in classification, such as sex, race, and age group, will equal the adjusted block totals. However, these lists will also be used to accumulate totals or counts for variables such as income and education that might not be used in the classification scheme. This section will consider the effect of these adjustment methods on such statistics. For concreteness of exposition, income will be used as the main example. Income is an important non-classification variable; some research suggests that revenue allocation programs may be most affected by errors in measurement of income. (National Academy of Sciences 1985).

In general, there are two possible sources of bias in the estimation of a non-classification covariate: (1) bias in adjustment of household composition, and (2) systematic differences between fully enumerated households and households with similar composition that are omitted (entirely or in part). However, if we have an estimate of mean income for the block, we can make the weighted mean for households in the block equal the estimated (adjusted) mean in much the same manner we make the weighted counts of individuals in the block equal the estimated (adjusted) counts.

### 8.1 Household Composition Bias

In this section we will assume that the average income level associated with a certain household composition is the same for fully enumerated households and those which are partly or wholly omitted from the enumeration. In other words, we consider here the case in which omission is noninformative for income.

Suppose that household income is a sum of independent contributions from persons of each class in the household (i.e. suppose that the contribution to income from persons in each class are independent of what other members are in the household). Then weighted household income totals would be an unbiased estimate of the true income totals (when adjustment rates are correct), since the sum of incomes would be a linear function of class counts for the block. However, under the more realistic assumption that linearity does not hold, misallocation of persons between households (and corresponding misrepresentation of household composition in the adjustment) could lead to bias in income estimates. Thus, for example, the average income of households with two children might not be the mean of the average income of one-child and three-child households (with the same composition of adult members). Then the weighting procedure might introduce the correct number of children but if, on the average, too many (compared to the truth) two-child households were created relative to one- and three-child households, estimates of household income would be biased.

Our procedure tends to fit weights that make the "adjusted-in"' households similar in composition to those that are common in the enumeration. However, the adjustment is described only by adjustment class totals, which do not carry detailed information on the composition of the omitted households. Thus, if certain household compositions are disproportionately undercounted they may be underrepresented in the weighted lists, and if these compositions are associated, for example, with lower incomes, the total income estimates will be biased upwards.

This is essentially a problem of potential lack of fit of the model used in adjustment to the patterns in the data. The most severe biases might appear in statistics that refer specifically to household composition, such as the number of single-parent families.

If composition bias were found to be a serious problem, one approach to controlling it would be to augment the class adjustment rates with additional information that describes the joint omissions of persons from different classes (or grouped classes).

### 8.2 Response Bias

It is not unreasonable to think that, of a group of households with the same composition, those which are missed in the census will differ systematically in some characteristics from those that are enumerated. In other words, omission may be a form of nonignorable nonresponse. For example, households with lower incomes and educational levels may be more likely to be missed altogether, or to omit some members from their roster; income and education are not classification variables and therefore are not directly adjusted.

Whole-household adjustments are represented in the proposed methods by upweighting households, preserving the values of all covariates. The implicit assumption is that the omitted households do not differ on these covariates from enumerated households with similar composition. There is no information available in the block being adjusted to contradict this assumption. However, it should be possible to collect information in the PES on the differences between enumerated and missed households, which could be incorporated into the adjustment. For example, the income of wholly omitted households might be related to the mean income of enumerated households with the same composition by a linear regression; then the added (weighted-in) households could be imputed the income obtained by applying the linear regression function to the income of the enumerated donor household. Little and Rubin (1987) discuss relevant methods for missing data problems with informative nonresponse. Another approach that is integrated with the weighting adjustment methodology is described in the next section.

Within-household adjustments are represented by downweighting a household with certain enumerated characteristics and upweighting another household that contains an additional
member or members. In the absence of further adjustment, the characteristics of the upweighted household, rather than those of the enumerated household from which the weight was taken, will apply to the "weighted-in" component.

This poses problems that cannot be resolved without collecting some data (from a subsample of the PES). For example, if a child were omitted from the household roster, there is no reason to think this would lead to misreporting of income. If households with more children had a higher mean income than those with fewer children, then the weighting would tend to over-estimate mean incomes.

If an adult were omitted from the roster, this might also mean that the same adult's income (if any) would be left out of the reported household income. It is plausible that the mean unreported income in this situation would be positive but less than the mean income of the corresponding adults in households where all aduit members appear on the roster. For a stereotypical example, consider a family on public assistance that does not report an adult male member, whose income would otherwise be deducted from the assistance level, and whose residence is somewhat inconsistent. That member's income is likely to be less than that of a permanently resident adult male in a family that does not depend on public assistance. Thus, neither the income of the enumerated household nor that of the "weighted-up"' household would be an accurate imputation for the adjusted household.

No direct correspondence is established between households that are down-weighted and those that receive additional weight. Thus an unadjusted income cannot be carried over directly from the enumerated household to the "weighted up" household. However, with some research comparing the incomes of enumerated and missed households, the incomes of down-weighted households could be used in adjusting incomes. For example, the mean household income of the reweighted block could be constrained to be equal to that of the block before adjustment.

### 8.3 Weighting Adjustment of Non-classification Characteristics

Suppose that adjusted summary information (by block) is available on some characteristics of households other than counts of individuals by adjustment class. For example, we might have an adjusted estimate of mean income or of the proportion of single-parent families, possibly from a regression model. As long as the summary statistic can be represented as a weighted sum of covariate values for each household, then conformity to the desired adjusted value can be imposed by a linear constraint on weights which can be made part of the weighting adjustment methodology of this paper. Thus, in the income example, we would constrain the weighted sum of incomes to equal the product of the number of households and the adjusted mean income. To adjust the proportion of single-parent families, we would constrain the weighted sum of $0-1$ indicators for that status to the desired total count.

### 8.4 Summary and Implications

The methodology proposed will upweight households, and without further consideration of possible biases, will carry along the characteristics of the upweighted households. If the size of the adjustment and the biases introduced in household characteristics are both of small order, the overall bias in estimated block characteristics will be of second order and should not be a major problem. Some simple regression adjustments might make it possible to reduce the biases by an additional order of magnitude.

## 9. SUGGESTIONS FOR FUTURE RESEARCH AND DEVELOPMENT OF METHODOLOGY

This section summarizes a number of suggestions for implementation and further development of this adjustment methodology.

### 9.1 Post Enumeration Survey (PES) Data-gathering and Statistical Modeling

Omissions of persons in enumerated and omitted households should be distinguished in the PES and the two omission rates modeled separately for each adjustment class. Rates of omissions of whole households should also be modeled (Section 4). A variety of measures (as in Section 6.2) could be used to compare the composition of "weighted-in" households to that of omitted households found in the PES; if research found that "composition bias" was a significant problem, higher-order statistics should be developed (Section 8.1). A sample of PES households that were omitted in the Census should be administered the long form, so that the relationship between omission and covariates such as income and education could be modeled for the adjustment (Sections 8.2, 8.3).

### 9.2 Feasibility of Adjustments

The methods of Section 5 should be tested and compared using PES data.

### 9.3 Multiple Imputation

Although the procedures proposed in this paper operate deterministically, there are a number of sources of uncertainty in statistics based on the weighted records. These include: uncertainty in estimation of undercount rates; variability in class undercount rates from block to block around the national mean; binomial variability in the actual number of omitted households or individuals around the expected number given the undercount rate; uncertainty regarding differences between covariate values for omitted households and for enumerated households that are weighted up to replace them.

For research uses, files could be prepared that would represent all of these forms of uncertainty by multiple imputation (Rubin 1987). Two or more versions of the reweighted data set could be represented by including several sets of weights on the file. Researchers could repeat their analyses using each set of weights in turn. The variability among the statistics calculated on the different versions gives an estimate of the variability introduced by the process of undercount adjustment. Zaslavsky (1989) discusses procedures for multiple imputation in this setting.

## 10. SUPPLEMENTS

### 10.1 Choice of Objective Function for Weighting

A number of objective functions have been proposed for calculating an optimal fitted table (usually in the context of contingency tables, $c f$. Fagan and Greenberg 1988). In each case the function takes the form $T=\Sigma T_{1}\left(W_{i}\right)$, where $T_{1}$ takes one of the forms displayed in Table 9 . Each of these functions can be standardized to an equivalent function $T_{0}$ by multiplication by a constant coefficient and adding a linear function of $W$, so that $T_{0}(1)=0, T_{0}^{\prime}(1)=0$, $T_{0}^{\prime \prime}(1)=1$. Since $\sum W_{i}$ is constrained to a given value, the optimum weights will be unaffected. Then the standardized objective functions agree through the second term of their Taylor expansions about 1 , and should give similar results when the weights are close to 1 .

Table 9
Comparison of objective functions for table fitting

| Name of fitting <br> procedure | Objective <br> function <br> $T_{l}(W)$, usual <br> form | Objective <br> function $T_{0}(W)$, <br> standardized <br> form | Second <br> derivative <br> $T_{0}^{\prime}(W)$ |
| :--- | :---: | :---: | :---: |
| Least squares <br> (minimum variance) | $(W-1)^{2}$ | $(W-1)^{2} / 2$ | 1 |
| Raking | $W \log W$ | $(W \log W-W+1$ | $1 / W$ |
| Maximum likelihood | $-\log W$ | $W-1-\log W$ | $1 / W^{2}$ |
| Minimum $\chi^{2}$ | $(W-1)^{2} / W$ | $(W-1)^{2} / 2 W$ | $1 / W^{3}$ |

in the degree of asymmetry between the costs of downweighting and upweighting, determined by the exponent of $W$ in the second derivative, $T_{0}^{\prime \prime}(W)=W^{-k}$. The least squares procedure ( $k=0$ ) treats up-and down-weighting completely symmetrically and therefore may yield zero or negative weights. As $k$ increases, the cost of upweighting becomes smaller relative to that of downweighting. All of the other objective functions $(k>0)$ give every observation in the raw data a positive weight; in the case of the "raking' function, this is obvious from the form of the weights as shown in Section 3.3. The use of the "raking" function here in preference to maximum likelihood or minimum $\chi^{2}$ is motivated by the simple form of its solution and by the analogy to raking in contingency tables. Cressie and Read (1984) systematically study the properties of this family of measures of fit.

### 10.2 A Cyclic Descent Methodology for Fitting Weights

In this section we present a fitting methodology analogous to iterative proportional fitting (IPF) in contingency tables. In IPF, the cell counts are transformed multiplicatively in such a way that the cross-products are preserved (the condition for minimization of the objective function) while the table is made to conform to each set of marginal constraints in turn. The algorithm converges to a table that satisfies all of the constraints, and perforce preserves the cross-products as well (Bishop, Feinberg and Holland 1974; Ireland and Kullback 1968).

In our setting, the weights are required to have the log-linear form $W_{i}=\exp \left(a_{i}^{\prime} \lambda-1\right)$ derived in Section 3.3 while satisfying the constraints $A W=B$. In this exposition we will assume that the total weight constraint $\sum W_{i}=H$ is omitted from $A W=B$, and that $A$ is of dimension $p$ (constraints) $\times I$ (number of household compositions). We will proceed through a series of steps in each of which each weight $W_{i}$ is multiplied by $c \rho^{a_{j i}}$ to obtain a new weight $W_{i}^{\prime}$, thus preserving the log-linear structure; $c$ and $\rho$ are chosen so that the constraints $\Sigma W_{i}^{\prime}=H$ and $\sum W_{i}^{\prime} a_{j i}=b_{j}$ are satisfied. By proceeding cyclically so that $j=1,2, \ldots p$ indexes each constraint in turn, the algorithm eventually converges to weights that satisfy all of the constraints.
 $j=1$ by using the last weights from the last cycle, $W_{i}^{(t, 0)}=W_{i}^{(t-1, p)}$ ). Then $c$ and $\rho$ must satisfy

$$
\begin{equation*}
\sum_{i} c \rho^{a_{j i} W^{(t, j-1)}}=H, \sum_{i} a_{j i} c \rho^{a_{j i} W_{i}^{(t, j-1)}}=b_{j} \tag{5}
\end{equation*}
$$

Eliminating $c$ from these equations, $\rho$ is a root of

$$
\begin{equation*}
\sum_{i}\left(H a_{j i}-b_{j}\right) W_{i}^{(t, j-1)} \rho^{a_{j i}}=0 \tag{6}
\end{equation*}
$$

We must have $H a_{j, \text { min }} \leq b_{j} \leq H a_{j, \max }$ where $a_{j, \min }$ and $a_{j, \max }$ are respectively the minimum and maximum values of $a_{j i}$. If this were not the case, constraint $j$ could not be satisfied with any weights. Thus there must be at least one root $\rho$, and if the $a_{j i}$ are non-negative, the expression is increasing in $\rho$ so this root is unique. The actual value of $\rho$ is determined then by Newton's method, or by a closed-form formula for the roots of a polynomial (since with the original $A, a_{j i}$ is the number of class $j$ members in a household, which is an integer rarely exceeding 2 ).

While we have not yet proven that this algorithm always converges, we have found it to be successful in practice. This algorithm does not require any matrix inversion, and if the $a_{j i}$ are small integers, then at each step, the recalculation of the weights involves calculating only a few integral powers. Furthermore, if some constraints take the form of simple marginals, the adjustment for those constraints takes the form of a conventional raking step.

If the original constraint matrix $A$ is used, the procedure may take advantage of the sparseness of $A$ (which is a consequence of the fact that only a few classes are represented in each household). At each step (say, adjusting to fit margin $b_{j}$ ), only the weights corresponding to non-zero $a_{j i}$ need be modified; thus only $S_{1}$ multiplications (the number of nonzero entries in $A$, which is less than the population of the block) and perhaps $3 S_{1}$ additions are required per cycle, as compared to $5 S_{1}+S_{2}$ operations per iteration with Newton's method. On the other hand, the rows of $A$ tend to be highly dependent, so convergence may be slow (typically 20 cycles in our simulations); orthogonalization of $A$ destroys the sparse structure of the coefficients. Thus, unless $S_{2}$ is much larger than $S_{1}$ (or unless some other method is devised to accelerate the algorithm), raking is not faster than Newton's method.

## ACKNOWLEDGEMENTS

This research was supported by Joint Statistical Agreements $86-8$ and $87-7$ between the U.S. Bureau of the Census and Harvard University. The author has benefited from comments by Donald B. Rubin and other participants in the seminar on census undercount at the Harvard University Department of Statistics, and Nathaniel Schencker, Nash Monsour, and other members of the Undercount Research Staff in the U.S. Bureau of the Census. Simulations used data made available by the Inter-university Consortium for Political and Social Research and originally collected by the U.S. Department of Commerce, Bureau of the Census.

## REFERENCES

ALEXANDER, C.H. (1987). A class of methods for using person controls in household weighting. Survey Methodology, 13, 183-198.
BISHOP, Y.M.M., FIENBERG, S.E., and HOLLAND, P.W. (1974). Discrete Multivariate Analysis. Cambridge: M.I.T. Press.
BUREAU OF THE CENSUS (1985). Census of Population and Housing, 1980: Public Use Microdata Samples.

CILKE, J.M., and WYSCARVER, R.A. (1988). The Individual Income Tax Simulation Model, in Office of Tax Analysis, Compendium of Tax Research 1987, Washington: Government Printing Office.
COX, L. (1987). A constructive procedure for unbiased controlled rounding. Journal of the American Statistical Association, 82, 520-524.
COX, L., FAGAN, J., GREENBURG, B., and HEMMIG, R. (1986). Research at the Census Bureau into disclosure avoidance techniques for tabular data. Proceedings of the Section on Survey Research Methods, American Statistical Association, 388-393.
CRESSIE, N., and READ, T.R.C. (1984). Multinomial goodness-of-fit tests. Journal' of the Royal Statistical Society, Series B, 46, 440-464.
DEMING, W.E., and STEPHAN, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. Annals of Mathematical Statistics, 11, 427-444.
DIFFENDAL, G. (1988). The 1986 Test of Adjustment Related Operations in Central Los Angeles county. Survey Methodology, 14, 71-86.
FAY, R.E. (1986). Implications of the 1980 PEP for future census coverage evaluation. U.S. Bureau of the Census, unpublished.
FAGAN, J.T., and GREENBERG, B. (1988). Algorithms for making tables additive: Raking, Maximum Likelihood, and Minimum Chi-square. Proceedings of the Section on Survey Research Methods, American Statistical Association (forthcoming).
GASS, S.I. (1964). Linear Programming: Methods and Applications. New York: McGraw-Hill.
IRELAND, C.T., and KULLBACK, S. (1968). Contingency tables with given marginals. Biometrika, 55, 179-188.
LITTLE, R.A., and RUBIN, D.B. (1987). Statistical Analysis with Missing Data. New York: Wiley.
NATIONAL ACADEMY OF SCIENCES (1985). The Bicentennial Census: New Directions for Methodology in 1990. Washington: National Academy Press.
OH, H.L., and SCHEUREN, F.J. (1978). Multivariate ratio raking estimation in the 1973 Exact Match Study. Proceedings of the Section on Survey Research Methods, American Statistical Association, 716-722.
PURCELL, N.J. (1979). Efficient estimation for small domains: a categorical data analysis approach. Ph. D. dissertation, University of Michigan.
PURCELL, N.J., and KISH, L. (1979). Estimation for small domains. Biometrics, 35:365-384.
RUBIN, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: Wiley.
SCHEUREN, F.J. (1981). Methods of estimation for the 1973 exact match study. In Studies from Interagency Data Linkages, Washington: Social Security Administration.
ZASLAVSKY, A.M. (1989). Representing Census undercount at the household level. Ph. D. thesis, Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts.

# QUID, A General Automatic Coding Method 

JACQUES LORIGNY ${ }^{1}$


#### Abstract

The QUID system, which was designed and developed by INSEE (Paris) Institut National de la Statistique et des Études Économiques - National Statistics and Economic Studies Institute, is an automatic coding system for survey data collected in the form of literal headings expressed in the terminology of the respondent. The system hinges on the use of a very wide knowledge base made up of real phrases coded by experts. This study deals primarily with the preliminary automatic standardization processing of the phrases, and then with the algorithm used to organize the phrase base into an optimized tree pattern. A sorting example is provided in the form of an illustration. At present, the processing of additional coding variables used to complement the information contained in the phrases presents certain difficulties, and these will be examined in detail. The QUID 2 project, an updated version of the system, will be discussed briefly.


KEY WORDS: Automatic coding; Natural language variables; Phrase matching; N-grams.

## 1. INTRODUCTION

The QUID (abbreviation of QUestionnaires d'IDentification - Identification Questionnaires) system is an automatic coding system designed and developed by the Institut National de la Statistique et des Études Économiques (INSEE - National Statistics and Economic Studies Institute) in 1979-1980.

## Review of the Problem

The problem consists of automatically classifying an individual surveyed into a job defined in accordance with an existing nomenclature (for example, the nomenclature of the professions). In order to do this, the system uses mainly the natural language answer given in response to a direct question (for example, "What is your present profession or trade?"), as well as additional information contained in the survey form, which is assumed to have been previously coded (for example, the Economic Activity code for the firm where the individual works).

In our terminology, a direct answer in natural language is called the "literal heading", or simply "heading". Any additional encoded information is represented by the generic term 'additional variables".

In the next section, we will discuss the basic approach of the QUID system and the results of its implementation at INSEE. In section 3, we will describe the present version of the system. Finally, in section 4, we will examine the problems surrounding the processing of additional variables, and will discuss the new version of the system (QUID 2), which should help resolve the difficulties encountered.

[^14]
## 2. THE PRINCIPLE BEHIND THE METHOD

### 2.1 The Basic Approach

The basic approach of the QUID system consists of building a very large data base made up of typical respondent headings accompanied by a corresponding code assigned by an expert. The data base is as large as possible in order to make it possible to obtain a high matching rate, and new headings are added to the base as they appear.

In our terminology, the data base is called a "knowledge base" or "knowledge file" (KF), because it has the ordinary structure of a flat file in its raw state. Most often, the knowledge file is set up on the basis of a survey carried out during a previous year, which has already been coded either manually or using an interactive method. Each base heading is accompanied by its code (which is a priori assumed to be accurate), and its "frequency of occurrence" in the KF; that is, the number of individuals who responded using this heading.

The management task of the knowledge base (auditing, expansion) is completely separate from the operation of coding the survey under way. It is the responsibility of a central office staffed by expert coders, while the coding operation itself is most often regionally decentralized.

The difficulties of an approach of this type derive from the rapid increase in the time required to search the base as it grows in size. In order to solve this problem, the QUID system uses mathematical results derived from Information Theory (Shannon 1948; C.-F. Picard 1972; B. Bouchon-Meunier 1978; M. Terrenoire 1970; D. Tounissoux 1980), which can be used to minimize search time by organizing the base in the form of an optimized tree structure.

The basic approach of the QUID system also makes it possible to opt for a set of general programs; that is, those that can be used with all semantic fields, for example, professional, food products, or municipal headings.

### 2.2 Results

The system has been tried for various INSEE tasks and is presently being used to code the CS (socio-professional category) code in order to process DADS (Déclarations annuelles de données sociales - Annual social information) data provided by all firms that employ paid labour. The following figures provide an idea of the orders of magnitude involved.

At present the knowledge file for the DADS application contains 122,000 headings (representing a knowledge base population of 650,000 wage earners). Its optimized organization consists of a tree with about 100,000 nodes (of which 86,000 represent certainty nodes; see section 3.2). It has been used to code a population of 570,000 wage earners with an average effectiveness of $90 \%$, varying between $85 \%$ and $95 \%$, depending upon the region. By 'effectiveness", we mean the percentage of cases where the system provides a single answer which is accepted on principle under the conditions of this application. At present, since we do not have a precise measurement of the validity of these single answers, we estimate that the error rate is likely to be in the order of $5 \%$ to $10 \%$. However, the knowledge base is being audited by the Dijon Expert Centre, according to which a significant proportion of the error rate should normally decrease. Once this has been achieved, we will have more accurate figures to report.

From the point of view of data processing limitations, the optimized tree is loaded into 3,300 kilobytes of central (virtual) memory and the automatic coding time for an individual case is in the order of 40 ms in an IBM 4341 central processing unit.

For the last few months, we have had available a variant of the coding program itself. This has been designed for use with mini-computers and can load the tree by sections, depending upon available memory space.

In applications other than DADS data, effectiveness is not as high, no more than $75 \%$. It all depends upon the quality and comprehensiveness of the knowledge base.

## 3. THE PRESENT VERSION OF THE QUID SYSTEM (QUID 1)

### 3.1 Preliminary Standardization of Headings

Before constructing the optimized tree, the raw headings are first standardized in accordance with a set of external parameters chosen by the user for his application.

The words are separated and fitted into predetermined zones whose length (a single one for all words) and maximum number (a single one for all headings) are parametrized. It is advisable to choose a larger value on the basis of these two parameters, and allow the optimization algorithm itself to select the significant elements of the heading (see section 3.2). For example, the DADS application (see section 2.2) chose 4 zones of 12 characters each.
"Empty words" are eliminated. The list of empty words is an external parameter provided by the user for his application. Most often, it includes articles, prepositions, etc., and is significantly dependent upon the application.

Initials are standardized (I.N.S.E.E. becomes INSEE, S N C F becomes SNCF).
Finally, the user may process the table of separate words in any way he wants (in the form of a subprogram in the PL/1 language). In fact, this is rarely necessary and seldom used (except to code municipal codes from municipal headings).

Once word processing has been completed, the words are divided into bigrams (blocks of two consecutive letters) or trigrams (blocks of three consecutive letters), etc. Choosing the type of blocking is parametrized (however, a single parameter is used for the entire application). In practice, blocking into bigrams is the only type that has been used until now; however, the idea of blocking into trigrams should be tested. For the purposes of this study, we will only consider blocking into bigrams.

### 3.2 The Algorithm Used to Set Up the Optimized Tree Pattern

Let $T=\left(t_{1}, t_{2}, \ldots, t_{j}, \ldots, t_{n}\right)$ the code to be coded, for example all the modalities of the Profession code.
$Q=\left(q_{1}, q_{2}, \ldots, q_{i}, \ldots, q_{m}\right)$ all the bigrams resulting from the standardization of the headings (for example, $m=24$ when the number 4 has been chosen as the "number of words" parameter, and 12 characters as the "word length" parameter).
$X=$ the tree pattern to be constructed, which we call a "QUID" (questionnaire d'identification - identification questionnaire).
The algorithm constructs $X$ by parsing down from the root node $x_{0}$ (which by convention is placed at 'level 0 ') to the nodes in levels 1,2 , etc.

At root node $x_{0}$ it links the entire KF, and searches for the best bigram to query first; that is, that which can discriminate best for the desired code $T$ in the entire KF.
$N\left(x_{0}\right)$ represents the total frequency of occurrence associated with the entire KF; that is, the sum of frequencies accompanying the base headings,
$N\left(x_{0}, j\right)$ is the frequency of occurrence of code $t_{j}$ in the entire KF.
We assume that the knowledge population is statistically representative of the population to be coded (we should recall that, in practice, the KF is often the survey file for a previous year).

Thus, we can estimate the probability of finding code $t_{j}$ the population to be coded on the basis of the following formula:

$$
\operatorname{Pr}\left(t_{j} \mid x_{0}\right)=N\left(x_{0}, j\right) / N\left(x_{0}\right)
$$

The a priori ambiguity for $T$ is measured on the basis of Shannon's entropy:

$$
H\left(T / x_{0}\right)=\sum_{j} \operatorname{Pr}\left(t_{j} \mid x_{0}\right) \log 1 / \operatorname{Pr}\left(t_{j} \mid x_{0}\right)
$$

Let us assume that a bigram, for example $q_{i}$ is allocated to node $x_{0}$. To each of its modalities in the KF, we associate the sub-base made up of the headings that have this modality.

Let $\left(a_{i}^{1}, a_{i}^{2}, \ldots, a_{i}^{k}, \ldots\right)$ represent the modalities captured by bigram $q_{i}$ in the KF. For each of these modalities, thus, for each of the sub-bases generated, we create a node $y$, which follows immediately after $x$ and is located at level 1 of the tree.

The information provided by bigram $q_{i}$ (which is assumed to be assigned to root node $x_{0}$ ) is measured by the average reduction in the ambiguity of $T$ when we go from $x_{0}$ to one of the $y$ nodes.

That is:

$$
I\left(x_{0}, T, q_{i}\right)=H\left(T \mid x_{0}\right)-\sum_{y \in \Gamma\left(x_{0}\right)} \operatorname{Pr}(y) H(T \mid y)
$$

where
$\Gamma\left(x_{0}\right)$ represents all the successive $y$ nodes at level 1 below node $x_{0}$
$H(T \mid y)$ the conditional entropy of $T$ at node $y$.
(same formula as above but replacing $x_{0}$ by $y$ ).
$\operatorname{Pr}(y)=N\left(x_{0}, a_{i}^{k}\right) / N\left(x_{0}\right)$ if $a_{i}^{k}$ is the modality of bigram $q_{i}$ which generates node $y$, and $N\left(x_{0}, a_{i}^{k}\right)$ is the frequency of occurrence of modality $\mathrm{a}_{\mathrm{i}}^{\mathrm{k}}$ of bigram $q_{i}$ in the entire KF.

The algorithm carries out this data calculation for all bigrams $q_{1}, q_{2}, \ldots, q_{m}$, because at root node $x_{0}$ they are all possible candidates for selection as the first bigram to be queried.

The algorithm chooses the bigram which maximizes $I\left(x_{0}, T, q_{i}\right)$. For example, in the case of $q_{i 0}$, it effectively divides the base into as many sub-bases as there are modalities of bigram $q_{i 0}$ in the base. This effectively creates the y, nodes that follow $x_{0}$ at level 1 , and the construction of level 1 of $X$ is thus completed.

For each sub-base obtained (thus, for each $y$ node), the algorithm carries out exactly the same operation as that which we have just described for root node $x_{0}$, and so on.

The process stops for a given node:
(1) when there is only one heading at the node; in this case, the conditional entropy is zero; or
(2) when there is only a restricted number of headings that differ in terms of the remaining bigrams, but which have all the same code; or
(3) when there are two headings or more, but they have different and not distinguishable codes.
Cases (1) and (2) are known as "certainty nodes", and case (3) is known as the "uncertainty node". Together, they represent the "terminal nodes".

Standardized Heading:


Query the content of bigram no. 2 in this node of the tree.

The content of bigram no. 2 is EF.


In this node of the tree, we may determine the profession code: its value is 7622 (1975 Trade Nomenclature).

In this example, the raw heading is that of the profession entered by the individual surveyed. The objective of the system is to find the corresponding profession code in the 1975 Trade Nomenclature.

Initially, we extract the first ten characters of the three most significant words. In this way, we obtain the standardized heading, which is then blocked into pairs of letters (these are called bigrams and are numbered from 1 to 15). Then, we query the system. This operates in accordance with a chain of questions and answers optimized by a mathematical algorithm based on information theory. This calculation takes place during the course of a preliminary phase which determines the first bigram queried as a function of a given knowledge file, and then the following sequence of questions depending upon the answer obtained each time. At this point, the computer queries first bigram no. 12, which contains TR. At this stage, it ascertains that it can without ambiguity determine that this repesents the Profession 7622 code (Technical Staff and Technicians). On the average, processing time takes a total of 41 milliseconds of computer time in an IBM 370/148, and the amount of central memory used is 380 Kbytes.

Raw Heading: head, maintenance team.
Figure 1. Example of Classification of a Heading in the Tree.

The construction of the tree $X$ continues from level to level until the KF has been exhausted. In fact, we have never gone beyond level 15, but there is no set limit for the system itself. An example of classification in the tree is shown in Figure 1.

### 3.3 The Use of the Coding Itself

In order to code a heading in the current survey, we start by standardizing the information in accordance with section 3.1. Then, the bigrams obtained are matched against those of the Quid loaded into the computer. The exploration leads to three possible results.

### 3.3.1 Certainty Node

The system provides a single code but this may well be wrong if the knowledge base is not comprehensive enough. For example, during one of our first tests in 1979, we obtained a certainty node for level 1 on the basis of bigram $2=C C$, since the only heading obtained had been VACCINEUR VOLAILLES (POULTRY VACCINATOR).

When we later had to code the heading RACCOMMODEUR VÊTEMENTS (GARMENT MENDER) the single code obtained was that representing agricultural service professions, and the error was obvious.

Thus, we added to the system a control procedure based on single echoes. This process is known as "redundancy control" and consists of verifying, after the detection of a single echo, the content of the first three bigrams of each word. A single echo (obtained on the basis of the vector leading to a certainty node) is said to be non-ambiguous, when the cluster of headings in the certainty node contains at least one heading that has the same redundancy bigrams as those of the heading to be coded. Otherwise, the echo is said to be ambiguous, and consequently treated as an anomaly of the automatic system. Experience has shown that this arrangement tends to consolidate significantly the reliability of the system without appreciably overburdening the tables in memory or increasing processing time (even in large applications, the number of redundancy formulas per certainty node is, on the average, in the order of one, and rarely goes beyond ten).

In order to be thorough, we should add that this redundancy control is not rigidly set once and for all. The user has two external parameters: the list of bigrams over which he intends to exercise control, and the (maximum) number of bigrams retained. In this way, he can keep in check the severity of the matching control, depending upon his objectives in terms of the quality and "effectiveness' of automatic coding.

### 3.3.2 The Uncertainty Node

The system provides various possible codes (most often two codes), and displays their respective frequencies of occurrence at the node under consideration. In this case, the officer who has the file of the survey being processed will then manually reject one of the two.

### 3.3.3 The Case of an Unknown Response

If, during the course of exploring the Quid, the modality sought is not found in the modalities captured by the bigram queried, the search will fail and this also represents a case of rejection that must be processed manually.

New cases encountered during the course of processing will be stored in memory, centralized in the expert centre, verified, and then incorporated into the KF in order to produce a new expanded version of the Quid.

At present, for purposes of convenience, the knowledge iteration takes place once a year, but nothing prevents it from being organized so that it takes place more often so that applications can progress faster, for example in the case of population surveys.

## 4. THE PROBLEM OF PROCESSING ADDITIONAL VARIABLES

In the present version, QUID 1, additional variables are simply structured into bigrams and processed in the same way as literal data. This leads to certain difficulties and problems that made it necessary to develop a new version, QUID 2, which operates in two stages:

- in the first stage, QUID 1, which is reserved for processing the literal heading and producing either the final code (when this is totally determined by the heading), or an internal code designating a rule or decision table that can be applied to the additional variables to achieve the calculation;
- in the second stage, the rules or decision tables achieve the determination of the final code.


## Detailed Examination of the Difficulties Encountered

At times, certain nomenclatures that are particularly complex, such as the PCS Code (Nomenclature of Professions and Socio-Professional Categories) call upon a combination of the literal heading and various additional variables.

For example, the coding of the PCS code uses the Professional Category additional variable (which is abbreviated to CPF). The following is the question such as it appears in the 1982 Population Census Individual Form:

Indicate the professional category of your present job:

- unskilled or semi-skilled labourer 1
- labourer - semi-skilled labourer ( $\mathrm{OS}, \mathrm{O} 1, \mathrm{O} 2, \mathrm{O} 3, \ldots$ ) 2
- skilled labourer (P1, P2, P3, TA, OP, OQ ...) 3
- clerk 4
- technician, draftsman 5
- supervising workers or clerks 6
- foreman - supervising other foremen or technicians 7
$\begin{array}{ll}\text { - engineer or professional staff } & 8\end{array}$
The additional question was made necessary by the fact that the heading alone is not always enough to classify the individual in accordance with PCS nomenclature.

For example, a LUMBER COMPANY WORKER

- must be classified into 6916 (lumber company or forestry worker) if his CPF is $1,2,3$, or 4
- and into 4801 (Managerial and supervisory staff of agricultural or lumber operations) if his CPF is $5,6,7$, or 8 .

The present system considers these additional variables as if they were literal data. They are placed at the end of the heading and structured into bigrams in the same way (for example, the CPF variable with the addition of a blank space is placed into the $(m+1)$ th bigram). However, this solution is not satisfactory and leads to various errors:

Error No. 1. When there is not enough information in the KF, this may lead to many cases of unknown responses.

For example, if the KF has only one LUMBER COMPANY EMPLOYEE with a CPF $=2$ and another with a CPF $=7$ the file will be unable to find a LUMBER COMPANY WORKER
with a CPF other than 2 or 7 (that is, a priori in 6 cases out of 8 ). This error is made worse when the additional variable is very diluted, for example, in the case of the variable representing the Economic Activity of the undertaking (which is abbreviated as additional variable AE).

Error No. 2. When there is not enough information in the KF, this may lead to miscodings.
For example, if the KF has only one LUMBER COMPANY WORKER with a CPF $=2$, the CPF bigram will not discriminate or appear in the search key, so that a LUMBER COMPANY WORKER with a CPF $=7$, will be classified into PCS $=6916$ instead of 4801. This is a case of miscoding

In order to correct this defect in the present system, the only measure we can take is to apply the redundancy control to the additional variables (and thus obtain an ambiguous or questionable case which is rejected or corrected manually, instead of allowing the error to remain undetected). However, here again, this is only a last resort. In fact, the additional variables lead to an unchecked expansion of the KF. Each KF reference has its own cross combination of modalities of additional variables, and it is not very likely that we would find the same combination for a new individual to be coded. Thus, this will lead to many uncertain cases and automatic coding rejections, which will reduce the practical benefits of mass exploitation.

The two errors, no. 1 and no. 2, are related to the relative incompleteness of the KF. For example, it would be enough to enter into the KF eight LUMBER COMPANY WORKER titles and add in each case one of the possible CPF modalities ( 1 to 8), in order for the two errors to disappear. However, in the case of real applications, we find that the relative incompleteness of the KF decreases quite slowly, as it grows to reach its operating pace. Contrary to the lexicographic space of literal headings, which tend to become dense rather quickly, the cross checked space of the additional variables remains a vast frontier for a long time, and goes very slowly from a density of occupation of 0 to a density of 1 (one individual).

Error No. 3. There is a third category of errors that are not caused by the incompleteness of the KF but by the excessive sensitivity of the QUID in relation to errors inevitably contained in the file (and this always in relation to the additional variables).

Let us take a simple example. Let us assume that the SENIOR SECRETARY heading must be coded PCS $=4615$ (senior secretarial staff), regardless of the value of all the additional variables. Let us consider the following KF, in which an error has slipped by (for example, the failure to assign the PCS code):

| Heading <br> Senior | CPF a.v. | AE a.v. | PCS Code |
| :--- | :---: | :---: | :---: |
| Secretary | $\boxed{7}$ | $\lfloor 49 \mid 11]$ <br> (fashion design, <br> haute couture) | 4615 |
| Senior <br> Secretary | 77 | $\lfloor 83 \mid 43$ <br> (loan <br> cooperative) | 4616 |
|  |  |  | 1 <br> error |

Even though the AE additional variable should not be used to code the PCS code, the QUID algorithm uses it to separate the two certainty nodes.

- One in favour of 4615 in view of bigram $\mathrm{AE} 1=49$.
- And the other in favour of 4616 , in view of bigram AE1 $=83$.

The result is that, during the coding stage itself, all senior secretaries belonging to economic sectors other than those starting by 49 or 83 will appear as "unknown cases". Moreover, those in all sectors starting by 83 will obviously produce errors. However, it is mainly the first phenomenon that interferes with accuracy, because it affects an area that is much larger than that affected by the initial error.

Error No. 4. Finally, the present QUID algorithm is excessively rigid in terms of choosing the optimal question. Most often, this results in a simple inversion of the order of the questions in the course of the search, in relation to the order that would have been preferred by the designer. Thus, the effect is secondary, since the final results are identical. However, this may also lead to more serious distortions.

Let us take the following (partly fictitious) example. Let us assume that, according to the nomenclature, the SENIOR SECRETARY heading should be coded either PCS $=4615$ as above if the CPF additional variable CPF $=1$ to 7 , and $P C S=3726$ (current managerial staff in other administrative business services), if CPF equals 8.

Let us examine the KF containing the following two references:

| Heading | AE a.v. | CPF a.v. | PCS Cod |
| :---: | :---: | :---: | :---: |
| Senior |  |  |  |
| Secretary | 49\|11 | 8) | 3726 |
| Senior |  |  |  |
| Secretary | 83\|43 | 7 | 4615 |

Thus, the two references are correctly coded. When the QUID algorithm arrives at a node where it has examined all the possible bigrams of the literal heading, it must now choose one bigram in the additional variables, in order to separate the two final results: $P C S=3726$ and PCS $=4615$. In this simple but not altogether unrealistic example, the three possible bigrams: AE1, AE2, and CPF, provide the same quantity of information (one bit). In our algorithm, the arbitrary convention is that in cases of equality, the program should choose the first question in the order in which the additional variables were presented in the form. However, in this example, this will be deceiving, since we would encounter the aberration discussed above (error no. 3). However, it is not possible to determine an order of additional variables that would prevent this type of error in all important cases. We can only seek an order of questions that will be statistically the least invalid, by groping our way on the basis of the order of conceptual splits, the negentropic capacity of each additional variable, etc.

## 5. CONCLUSION

In its QUID 1 version, the present QUID system provides very valuable services to INSEE. Nevertheless, it still has certain weak points regarding the processing of additional variables.

The new QUID 2 version should improve processing while remaining faithful to our "basic approach'" to the automatic coding problem, which could be summarized in two points:

1. Separation of the knowledge base (in this case, a base of rules and decision tables that are written in natural language, are independent of each other, and are audited and managed by an autonomous expert centre), and the use of automatic coding programs (in this case, loading and table exploration programs).
2. Construction of general programs; that is, programs that are independent of the semantic field processed.

At least, these are the objectives that we try to attain.

## ACKNOWLEDGEMENTS

I would like to thank the reviewers for their invaluable help during the preparation of this paper.

## REFERENCES

BOUCHON-MEUNIER, B. (1978). Sur la réalisaiton de questionnnaires. Doctoral thesis. Paris.
KNAUS, R. (1987). Methods and problems in coding natural language survey data, Journal of Official Statistics, 3, 45-67.
LORIGNY, J. (1982). Mesures d'entropie et d'information pour les systèmes ouverts complexes. Doctoral thesis, Paris.
LORIGNY, J. (1985). Manuel d'utilisation du système QUID. Institut National de la Statistique et des Études Économiques, Direction de la production, Paris.
PICARD, C.-F. (1972). Graphes et Questionnaires. Paris: Gauthier-Villars.
SHANNON, C.E. (1948). A Mathematical Theory of Communication. Bell Systems Technical Journal, 27, 379-423, 623-656.
TERRENOIRE, M. (1970). Un modèle mathématique de processus d'interrogation: les pseudoquestionnaires. Doctoral Thesis, Grenoble.
TOUNISSOUX, D. (1980). Processus séquentiels adaptifs de reconnaissance de formes pour l'aide au diagnostic. Doctoral Thesis, Lyon.

# ACTR A Generalized Automated Coding System 

M.J. WENZOWSKI ${ }^{1}$


#### Abstract

A generalized implementation of a method for performing automated coding is described. Traditionally, coding has been performed manually by specially trained personnel, but recently computerized systems have appeared which either eliminate or substantially reduce the need for manual coding. Typically, such systems are limited in use to those applications for which they were originally designed. The system presented here may be used by any application to perform coding of English or French text using any classification scheme.


KEY WORDS: Automated coding; Classification; Text searching.

## 1. INTRODUCTION

Automated coding refers to the process by which text is machine analysed in order to assign it a classification, or code. To be practical, automated coding systems must be capable of coping with such problems as: rearranged words, plural vs singular forms, missing words, extraneous words, spelling variations, synonyms, abbreviations, inconsistent hyphenation and variable punctuation and syntax. In addition, in searching a text database for a match, they should be capable of determining the closest match when no identical match can be found.

Generalized systems provide all of the features required, packaged within an easy to use, flexible, and efficient framework. To use a generalized system for a particular application, no development or conversion effort is needed to tailor it to the application specific requirements. As well, no application sponsored support for the maintenance of a generalized system is necessary, since the package is supported and maintained by a central agency.

ACTR (an acronym for: Automated Coding by Text Recognition) employs techniques similar to those employed in other automated coding systems currently in production at Statistics Canada (Landry and Pidcock 1984), but is unique in that it has been generalized to allow it to be used by any application to assign codes based on the input of English or French text according to any classification scheme.

The methods which ACTR uses to perform automated coding are based on techniques which were originally developed at the U.S. Bureau of the Census (Appel and Hellerman 1983). Basically stated, the method consists of searching through a collection of text previously associated with correct codes. If the subject text is successfully located, the associated code is returned and the process ends. Otherwise, the search continues, but uses an algorithm to locate the closest match, and subsequently assign its associated code.

[^15]
## 2. USING ACTR

To use ACTR in an automated coding application users first need to define the text and associated codes which they intend to use as a standard for matching. While there are many sources for this information, the best is a set of text which is representative of the text which will most likely be encountered in a matching run. For a survey, this generally means the responses and manually assigned codes from a previously completed survey. Although great care should be taken to ensure that the correct codes have been assigned, the text should be left as is, complete with spelling, grammar and syntax errors, since in this form it is most representative of the text which will be encountered in subsequent surveys.

After having defined a file of text and correctly assigned codes, they must be loaded into a matching database. ACTR provides the software required to perform this task and so automatically transforms the file into a matching database.

ACTR has been designed to allow an iterative approach to developing an automated coding application. Accordingly, text and codes can be added, changed or deleted at any time during the life of the application. In addition, the parsing strategy (discussed in detail below) can be altered at any time. Thus, users are presented with a software framework which, through cycles of database updates and matching runs, will allow for as many iterations as is necessary to obtain the matching quality desired. Users are encouraged to use ACTR in this manner, since ultimately it leads to higher quality and more economical coding operations.

## 3. PRINCIPLES OF OPERATION

In the case of a human being performing a coding operation, the similarity between occupations described as "Computer Programmer" and "Programming Computers" is so great that they would generally be judged as identical. However intuitive this reasoning may seem, computer systems in general would rate the two as unequal. Unfortunately, natural language (for example, English or French) frequently provides a large number of ways to express the same meaning. So, for a computer based system to be able to cope with this variance, there must be some means by which a degree of similarity can be determined.

This is the essence of ACTR: text is rated according to how similar it is to some other text. In the preceding example, ACTR treats the two occupation descriptions as identical since, after suffixes are truncated, double letters are removed and word order is ignored, both phrases become "Comput Program" and as such are clearly equal.

The steps employed in reducing the above phrases to a standard form are part of what is known in ACTR as the parsing strategy. ACTR's parsing strategy is entirely user controlled and may be changed at any time during the life of an application. Users exercise control over the parsing strategy employed in their applications by supplying the data which is to be used to direct the process. This means that all steps are entirely controlled by the user, even to the extent of allowing a step to be skipped.

## The Parsing Strategy

Parsing is the ACTR process which is responsible for the reduction of phrases to a standard form. Ideally, the resulting form should be such that any two phrases with the same words will be identical in their ACTR representation regardless of their syntactical and grammatical differences. Returning to the previous example, the two phrases "Computer Programmer" and "Programming Computers" when properly parsed, should ideally result in a set of identical words for each phrase. For example, both phrases could be reduced to "Comput Program".

The parsing process employed may involve the reduction of plural forms, elimination of trivial words, removal of suffixes and/or a number of other steps. Although the order of the parsing steps applied is fixed by ACTR, users control how, if at all, each step is executed. For further information on the order of parsing, the interested reader should consult Connor, Salloum and Wenzowski (1988).

Basically, the parsing process can be thought of as having the following two major subcomponents:

1. TEXT PROCESSING. In this stage of parsing, the text supplied is processed as a continuous stream of characters. Although one may think of the text as containing words, spaces and punctuation, none of these is given any special consideration at this point in the parse. This view is necessary in order to allow for the recognition of particular character strings exactly as they occur in situ.
2. WORD PROCESSING. When this stage of the parse begins, the text has already been broken down into words and so further processing is performed on a word by word basis. This view is necessary since a large amount of text standardization occurs on the basis of defined words.

## Text Processing

As already discussed, these steps are performed regardless of context. Thus, the following steps are performed on a character by character basis.

Exclusion Clauses: Exclusion clauses are ignored in matching, but are used in database updating to indicate the intention of allowing controlled duplication of phrases. By default, ACTR will not allow identical phrases to be loaded into a matching database.

By providing a means of controlling duplication, users are able to load phrases which could have more than one code assigned, even though they are identical after having been parsed. Although not used in matching, exclusion clauses are stored along with the phrase in the matching database and can subsequently be used to manually resolve multiple matches.

The syntax of an exclusion clause is defined entirely by the user. Both beginning and terminating strings must be provided. These and any information enclosed by them are ignored during matching.

As an example, consider an exclusion clause syntax defined with a beginning string of "(Except"' and a terminating string of ")". With this in place, the two phrases "Computer Programming (Except As An Employee)" and "Computer Programming (Except As SelfEmployed)" could co-exist in the matching database, even though their ACTR representations are identical. Subsequently, if a match for "Computer Programmer" is requested, both of these phrases would be returned. Since exclusion clauses are stored along with the original phrase text, they can be displayed to a reviewer, who could then manually resolve the match.

Deletion Strings: If any deletion string supplied by the user is found in any position in a phrase, ACTR will remove it from consideration before continuing the parse.

As an example, in English processing, this is a way in which the apostrophe can be removed. For example, the two phrases "Electrician's Apprentice" and "Apprentice Electrician" would become identical with the removal of the apostrophe.

Note that if this step were not performed, the apostrophe would most likely be used as a word delimiter. This would yield three words for the first phrase and two for the second, of which only one word would be common to both.

Replacement Strings: This facility is most useful for standardizing abbreviations. This is desirable since abbreviations commonly include characters which, although useful to the abbreviation, would be viewed as word separators at a later stage in the parse. If this were allowed to happen, information loss would most likely occur.

As an example, if the string "T.V." was defined with a replacement value of: "Television" then any occurrence of the original string would be translated to the replacement value before continuing the parse.

Note that if this step were not performed, the result of parsing "T.V." would most likely be the two letters " T " and " V ". This is clearly undesirable, since the meaning of the abbreviation has been completely lost.

Word Characters: ACTR defines a word as any contiguous sequence of characters in a phrase which are all members of the set of characters contained in the word character list. Any characters not in this list will be used as word delimiters and will be dropped from further consideration.

Typically, the set of word characters used contains all of the letters of the alphabet and all of the numeric characters. With this in place, a phrase of "Farmer/Fisherman" will result in two words, since "/" is not a word character and is therefore used as a word delimiter.

## Word Processing

At this point, ACTR begins to treat the text as a collection of words. Thus, the following processing steps are applied on a word by word basis.

Hyphenated Words: Any hyphenated words supplied are replaced by the subtitute word(s) also provided. This feature is very useful in providing for the recognition of words and word groups which are inconsistently hyphenated.

As an example, if the user defines "Take-Out" as a hyphenated word with a substitute word of "Takeout" then this substitution will be made. If, on the other hand, this definition had not been made, then two words would result if the hyphen was not a word character.

Illegal Word Characters: If any of the strings supplied are found to exist in any word in any position, then that entire word is removed from further consideration.

As an example, some applications use this feature to eliminate words which contain numeric characters. So, if the set of numeric digits was given as illegal word characters, then a word like "DEPT716A" would be removed from further consideration.

Replacement Words: This feature provides a synonym capability in order to ensure that two dissimilar words will be recognized for matching purposes. This can also be useful to overcome commonly occurring spelling mistakes.

As an example, if the phrases "Automobile Repairs" and "Car Repairs" were processed with the word "Car"' given as a replacement word for "Automobile" then the two phrases would be made identical.

Double Words: This feature forces ACTR to consider not only the occurrence of the two word grouping, but their order as well. This can be useful to overcome inconsistencies in word spellings and also to preserve word order.

As an example, consider the phrase "Take Out Restaurant". Although this would yield three perfectly acceptable words, the words "Take" and "Out" would not match to either of "Takeout" or "Take-Out". However, if a double word combination of "Take Out" was defined with a replacement of "Takeout" then the first case in the example given is addressed.

We are presented here with an example of how steps in the parsing strategy can be used together. If the hyphenated word example given above was also entered, then all of the hyphenated, double word, and single word cases would match.

Trivial Words: If any word in this set is encountered in the course of parsing, then it will be removed from further consideration.

As an example, if the set of trivial words contained "A", "Am" and ' I ", and the two phrases 'I Am A Computer Programmer" and "Computer Programmer" were encountered, then the phrases would match.

Suffixes: At this point, words are scanned right to left looking for the longest defined suffix such that the remaining word, after the suffix is removed, will be at least five characters in length. If a defined suffix is found, it is removed.

As an example, if the suffixes "ing" and "er" are defined, then the phrases "Computer Programming" and "Computer Programmer" will match.

Replacément Suffixes: Replacement suffixes are searched for in a word by scanning right to left for the presence of the longest defined replacement suffix. If one is found, it is removed and the substitute supplied is used in its place.

As an example, the user may wish a plural form to be reduced to a singular one so that the singular suffix will be recognized in the suffix truncation step. This is demonstrated with the phrases "Battery Manufacturing" and "Manufacturing Batteries". If the suffix "ies" is changed to " $y$ " then not only will the phrases be the same, they will be processed in the same manner at suffix truncation time.

Double Letters: At this stage in the parse, each word is examined for the presence of any double character occurrences which are contained in the (user-defined) double letter set. If any are found, they are reduced to a single occurrence.

Typically, the double letter set used is the full set of alphabetic characters. If this is the case, then the words 'Programer"' and "Programmer" would match, in spite of the spelling error.

Root Words: At this point, words are scanned for the presence of any of the root words supplied. The scan is applied from left to right in the word, and searches for the longest defined matching root word. If one is found, then its substitute is used as a replacement for the word and the suffix truncation and replacement steps are skipped.

As an example, the languages "Slavee"' and "Slavic" differ only in their last two characters. So, if the suffixes defined include "ee" and "ic" then an information loss occurs, since both words will become identical. Although generally, suffix truncation works well for most applications, it quite clearly fails for this particular example. To overcome this problem, if root words of "Slave"' and 'Slavi"' are defined, then the suffix truncation step is bypassed for these cases only. Thus, as suffix truncation problem cases are identified, root words and their substitutes can be defined to overcome them.

Duplicate Words: Finally, the set of words resulting from the parse of the supplied text is examined for the presence of duplicates.

Note that words which are duplicates at this point may not have appeared as duplicates before the text was parsed. Only one occurrence of each word defined at this point in the parse is kept.

## 4. SEARCHING AND MATCHING METHODS

ACTR always processes the supplied text according to the parsing strategy defined before attempting a match. If after doing this, ACTR is able to locate a phrase on the matching database with all of its words in common with all of the words in the supplied text, then the match found is referred to as a "Direct Match". If a direct match cannot be found, ACTR may, as a user option, continue to search the database for the closest match. This latter type of match is called an "Indirect Match". Although they share a common foundation in that they are both based on parsed text, the two matching methods used by ACTR differ greatly in their mechanisms for both locating and assigning a match.

## Direct Matching

In direct matching, only a $100 \%$ match is searched for. Recall that matching is based on parsed text, so phrases which are $100 \%$ matches may not appear to be identical in their original form. This is a direct effect of the parsing strategy in use.

In terms of database access techniques, the fastest path to an item is through the use of a key. Unfortunately, the roadblocks to keyed access of ACTR phrases exactly as they occur include a maximum phrase length of 200 characters and an upper limit of 20 on the number of parsed words. These two items make keyed access impractical since the extreme length of the key would negate any benefit derived. The only alternative to keyed access is sequential access, but this is undesirable because of the time required to search through the large volumes of information generally contained in a matching database.

So, we are presented with no other alternative but to somehow reduce the size of the key, thus making keyed access viable. There are many well known data compression techniques which could be used to do this, a general survey of which can be found in Reghbati (1981). In ACTR, the required data compression is achieved by forming the "compressed phrase key" or CPK. How CPK's are actually formed is discussed below. Accept for now that CPK formation results in a key which is approximately $35 \%$ of the original size of the phrase. The CPK can thus be used to access the matching database with an efficiently sized key in order to determine whether any direct matches exist.

The use of the CPK in ACTR is significant in the following ways:

1. All $100 \%$ matches will always be located using this method.
2. Since ACTR is able to locate direct matches by using the most efficient means possible, matches made by using this method are both faster and cheaper to perform.
3. As applications mature, the proportion of direct matches generally increases due to ongoing database update activity on the part of the user. Thus, overall matching costs for an application can actually decrease as the application matures, even though the size of the matching database may increase.

## CPK Formation

The CPK is formed by first ordering the words defined in parsing. The actual order is arbitrarily chosen and so is not significant, as long as the same ordering applies for all CPK formations. (The order used happens to be in ascending order of the collating sequence in use.)

After ordering, the words are concatenated into a single string which contains no blanks. This string is then compressed in order to form a short enough string to allow for efficient use as a database retrieval key. The compression of the string is based on the following:

1. The words resulting from parsing generally contain only characters from the 26 alphabetic character set and the 10 character numeric set. (Recall that the actual set of characters which may be encountered in words is user-defined.) However, characters are stored internally (ie. in memory and on disk) using an 8 bit code. Thus, there are $2^{8}$ or 256 possible 8 bit code combinations while ACTR words typically use no more than 36 of these. This leaves a 220 code surplus which could be used for other purposes.
2. Certain double and triple letter combinations are known to occur more frequently than others in English and French text samples. In ACTR, the double letter combinations are known as "digrams", and the triple letter combinations are known as "trigrams".
3. The 220 "free" codes can then be used to replace the digrams and trigrams described above as they occur in text samples.
4. Starting with the concatenated, parsed words, ACTR scans for the presence of any of the predefined digrams and trigrams. If any are found, they are replaced with the associated 8 bit code. The result is that a character sequence which formerly required 16 or 24 bits of storage, now requires only 8 bits.

## Indirect Matching

Like direct matching, indirect matching begins with the set of words resulting from the parsing process. However, indirect matching can never be as efficient as direct matching since the concept of closest match is relative. That is, we cannot find the closest match without first performing an exhaustive search through all of the possible matches.

In order to perform indirect matching, the matching database must first be searched for each of the words resulting from the parsing process in order to determine which, if any, are known. Following this step, for each word in the supplied phrase which is known to the database, all phrases containing the word must be retrieved and evaluated.

The nearest matching phrase is determined by calculating a score for each of the possible matches. Scores are based on the weights of the words which are in common with the database and subject phrases. Of all database phrases evaluated in this manner, the highest scoring phrase is the one which is considered to be the closest match.

## Word Weight Calculation

For each word known to the database, ACTR calculates a matching heuristic, or weight. These weights are an indication of the usefulness of a word in assigning a code and act as components in the phrase score calculation process.

The method by which word weights are calculated is based on: $n$, a count of unique codes, whose associated phrases contain this word; $V_{i}$, the relative frequency of code $i$ from previous surveys; $X_{i}$, a count of the number of word occurrences for phrases with code $i ; P_{i}$, the proportion of this word in code $i$, calculated as $V_{i} \times X_{i} / \Sigma_{j=1}^{n} V_{j} \times X_{j} ; E W$, the entropy of the word, calculated as $-\sum_{i=1}^{n} P_{i} \times \log _{2} P_{i} ; K$, the total number of word occurrences for code $i$, calculated as $\Sigma_{i=1}^{n} \times X_{i}$; EU, the entropy of a uniformly distributed variable with $K$ unique values, calculated as $\log _{2}(\mathrm{~K})$; and finally EO, a small value to avoid division by zero, calculated as $-K / K+1 \times \log _{2} K / K+1$.

From the preceding, word weights are calculated as: EU-EW $+\mathrm{EO} / \mathrm{EO}+\mathrm{EW}$.

## Phrase Score Calculations

For each database phrase which is evaluated for an indirect match, a score is calculated. The score is based on: $n$, the number of words the phrases have in common; $w_{k}$, the weight for word $k ; m$, the number of words in the subject phrase; and $l$, the number of words in the database phrase.

From the preceding, phrase scores are calculated as: $n^{3} \times \Sigma_{k=1}^{n} w_{k} / m \times 1$.

## Matching Parameters

After calculating a score value for each potential match, ACTR compares the score against user supplied values for the following parameters and takes the action indicated.

## 1. UPPER THRESHOLD

If the resulting score is greater than or equal to this value, then a winner is considered to have been found.

## 2. LOWER THRESHOLD

If the resulting score is greater than or equal to this value, but less than that supplied for the upper threshold value, then a possible match is considered to have been found.

## 3. PER CENT DIFFERENCE

If more than one winner is found, and their scores are within the supplied value for this parameter, then multiple winners are considered to have been found.

## Limiting the Search for an Indirect Match

ACTR searches the matching database for possible matches using the known words in the subject phrase. That is, these words are used to search for database phrases which contain them. The search proceeds in order of the ascending frequency of occurrence of the known words. Thus, the known word which occurs the least frequently in the database is used to start the search, the next lowest is used to continue the search, and so on.

As can readily be appreciated, finding a match by the indirect process has the potential of being time consuming and very expensive. Unfortunately, attempts to find matches by indirect means are unavoidable since a nearest matching feature is an essential component of any automated coding system.

While performing a search in this manner, ACTR maintains a list of database phrases which have already been evaluated. After a database phrase has been evaluated, it will not be reevaluated in a subsequent iteration for the currently executing matching effort. This ensures that a database phrase which contains more than one of the known words will not be evaluated more than once.

As a further search optimization, ACTR makes use of the user supplied matching parameters. With these, it constructs a table of optimistic scores for each iteration of the word based search:

1. For the first known word, the optimistic score is based on the possible occurrence of a database phrase with the same number of words as the number of known words and with all of its words in common with the subject phrase's known words.
2. For the second word, a similar assumption is made, but since the first word has already been used in the preceding search iteration, we know that any phrase containing the first word has already been evaluated. So, the optimistic score is based on the presence of the second and subsequent words only.
3. Optimistic scores for succeeding iterations are based on the presence of the current and succeeding unsearched words only.

The formula used to calculate the optimistic scores is based on: $a$, the number of known words in the subject phrase; $b$, the number of words in the subject phrase already searched; $c$, the total number of words in the subject phrase; and $d$, the number of known words not yet searched, calculated as $a-b$;

From the preceding, optimistic phrase scores are calculated as: $\left(d^{2} \times \sum_{i=d}^{a} w_{i}\right) / c$.
With the table of optimistic scores in place, ACTR evaluates the potential score at each iteration before performing a database access. Thus, hopeless searches are never attempted.

To summarize, the search for an indirect match is terminated when any of the following conditions are met:

1. The maximum potential score for the current iteration does not meet or exceed the threshold defined for possible matches.
2. At least one match has been found and the maximum potential score for the current iteration cannot produce another.
3. The maximum number of possible matches requested by the user has already been found and the maximum possible score for the current iteration does not exceed that of the lowest scoring phrase.

## 5. SUMMARY

A flexible and efficient automated coding methodology, embedded in a generalized software system has been presented. The system can be used to perform automated coding for any application in English or French or both, using any classification scheme. In doing so, it makes use of a powerful generalized parsing strategy and significant performance optimizations. For further information on ACTR, the interested reader is directed to Connor, Salloum and Wenzowski (1988).

## 6. ACKNOWLEDGEMENTS

While a complete list of all those who have been involved in the ACTR project would be too large to be presented here, the author would specifically like to acknowledge the contributions made by: John Connor and Bill Salloum, the principal programmers for the project; Victor Estevao, who was instrumental in helping to design the parsing strategy; Malvinder Rakhra and Paul Surman, who performed a great deal of system testing; and Don Royce, who was the project manager for the research effort.

## REFERENCES

APPEL, M., and HELLERMAN, E. (1983). Census bureau experiments with automated industry and occupation coding. Proceedings of the Section on Survey Research Methods, American Statistical Association, 32-40.
CONNOR, J., SALLOUM, B., and WENZOWSKI, M. (1988). ACTR Documentation Set. (System Overview, User's Guide, Tutorial, Guide to the Parsing Strategy, Default Parsing Data, Message Guide, Command Language Guide, Searching and Matching Methods \& Programmer's Guide) Internal Documents, Statistics Canada, Research and General Systems Subdivision, Ottawa, Canada.
LANDRY, L., and PIDCOCK, J. (1984). Business Register Automated SIC Coding System, System Proposal and Design. Internal Document, Statistics Canada, Informatics Services and Development Division, Ottawa, Canada.
REGHBATI, H. (1981). An overview of data compression techniques. Computer, 14,4, 71-75.

# Quality Control Processing System for Survey Operations ${ }^{1}$ 

WALTER MUDRYK ${ }^{\mathbf{2}}$


#### Abstract

The methods used to control the quality of Statistics Canada's survey processing operations generally involve acceptance sampling by attributes with rectifying inspection, contained within the broader framework of Acceptance Control. Although these methods are recognized as good corrective procedures, they do little in themselves to prevent errors from recurring. As this is of the utmost importance in any quality program, the Quality Control Processing System (QCPS) has been designed with error prevention as one of its primary focuses. Accordingly, the system produces feedback reports and graphs for operators, supervisors and managers involved in the various operations. The system also produces information concerning changes in the inspection environments which enable methodologists to adjust inspection plans/procedures in accordance with the strategy of Acceptance Control. This paper highlights the main tabulation and estimation features of the QCPS and the manner in which it serves to support the principal quality control programs at Statistics Canada. Major capabilities from a methodological and systems perspective are discussed.


KEY WORDS: Quality control processing system; Process control; Acceptance sampling; Acceptance control; Skip-lot sampling.

## 1. INTRODUCTION

This paper deals primarily with the features of the Quality Control Processing System (QCPS) that is presently being used at Statistics Canada. However, in order to show how this system fits into the overall quality picture for surveys, the paper begins with a brief discussion of the survey process and the role that quality assurance and quality control play in this process. The paper then identifies the specific quality control methods and strategies that are used for processing operations at Statistics Canada and how the QCPS serves to support this activity. The paper then proceeds to describe the system features and provides a summary of its major achievements.

### 1.1 The Survey Process

The requirement of ensuring quality in the overall survey process has always been considered a high priority at Statistics Canada. In a very general sense, it may be viewed as being achieved through the application of a series of quality assurance (QA) and quality control (QC) measures at the appropriate stages of a survey process. It is important to distinguish between these two activities since in our environment, they involve very different approaches and procedures that are normally applied at different points in the process. A simplified overview of the survey process at Statistics Canada includes the following stages:

[^16]- planning
- design
- implementation
- processing
- publication.

It is important to note that every one of these stages is subject to some error. It should also be realized that the further into the survey process the errors are discovered, the more impact they have on survey timeliness, cost and accuracy. Therefore, it is good practice to put a strong emphasis early in the process, on the development of measures and procedures that would prevent or reduce their occurrence. This should occur at the planning and design stages of the survey process. These measures and procedures are also known as quality assurance.

### 1.2 Quality Assurance

A general approach to establishing quality assurance is to try to anticipate problems very early in the survey process and take appropriate steps to prevent or minimize them. The anticipation can be based on experience, reviews, evaluations, debriefing exercises, feasibility studies, etc. The steps could include improving sampling frames/designs, modifying data collection methods, improving questionnaire design, providing clearer processing procedures, etc. A comprehensive list of such steps may be found in Statistics Canada's Quality Guidelines (1987).

This approach is extremely important since effectively it moves quality upstream and thereby helps to prevent many potential problems from occurring. Furthermore, in so doing, it assures better quality at the least cost by "getting it right the first time". Despite our best efforts however, there are some situations when error levels continue to be unacceptably high. In these situations we consider the use of quality control.

### 1.3 Quality Control

In contrast with QA, statistical quality control has been found to be highly applicable at the processing stage of the survey cycle. At this stage, the work usually has the following characteristics:

- labour intensive and repetitive in nature;
- assigned to individuals or operators with varying abilities;
- normally grouped into batches or lots of similar work units.

As such, these survey operations are more prone to the occurence of errors. Examples of these operations include:

- coding/transcription
- manual editing/reviewing
- data capture/entry
- corrections/reconciliation
- updating/profiling, etc.

For many reasons, which include complexity of tasks, abilities of operators, turnover of staff, etc., the amount and significance of error varies between operations, between operators within an operation, and at times within operator. Statistical quality control is used to identify and reduce this variability and ensure that the outgoing quality of each operation falls within acceptable levels.

## 2. QUALITY CONTROL STRATEGY

### 2.1 Methods of Quality Control

Of the two main methods of quality control available, namely, process control charts and acceptance sampling, we have found the latter methodology applied in the broader context of Acceptance Control, to be the more appropriate method for on-line quality control of survey processing operations. The reasons for this are as follows:

- prior control or stability of process cannot be assumed initially nor always attained in the long run;
- assignable causes of error are not always known since we are dealing with people (vs. say machines);
- processes cannot readily be stopped and adjusted for assignable causes, even if they are known;
- with many operators and large "between operator" variabilities, many individual control charts requiring immediate updating (i.e., after each sample observation) would be required on-line to the survey operation; this would be operationally difficult to achieve.

Therefore our quality control strategy generally consists of using varying acceptance sampling procedures (with rectification) applied at the operator level, as a screening device for correcting substandard quality, with the aim of continually reducing inspection as the inspection results support this action. This is coupled with an emphasis on operator and supervisor feedback to establish error prevention. In this manner both error correction and subsequent prevention are exercised at the error source, where they can have their greatest impact. Furthermore, between operator variations are automatically dealt with as each operator is effectively treated as a process in the following sense. During a period of low to moderate stability, acceptance sampling is applied to each lot processed. During a period of high stability coupled with good past inspection results, less acceptance sampling and even spot checking may be applied under the broader strategy of Acceptance Control.

### 2.2 Acceptance Control

After a quality control program has been operating for some time, operator processing abilities tend to improve and in many cases, a stabilization of quality occurs. In an effort to take advantage of this improved situation and to enable our quality control designs to be more economical, we have adopted the strategy that Schilling calls Acceptance Control (1982). Under this approach, acceptance sampling procedures are continually modified and adapted as changes in the inspection environment are identified. This is in accordance with one of QC's main pioneers, H.F. Dodge who states (1950):

> "A good product with a history of consistently good quality requires less inspection than one with no history or a history of erratic quality. Accordingly, it is good practice to include in inspection procedures provisions for reducing or increasing the amount of inspection, depending on the character and quantity of evidence at hand regarding the level of quality and the degree of control shown."

In fact the ultimate aim of acceptance control is to continually reduce inspection to the level of spot checks or process controls as the quality history improves and stabilizes. At Statistics Canada, two specific approaches are used to achieve this principle:

- Graduated Inspection Plans. These are obtained by raising or lowering the quality index for the sampling plan as changes in the process average are observed and then closely monitoring the impact on the resulting average outgoing quality estimates.
- Cumulative Results Plans, more specifically Skip-Lot Sampling (Stephens 1982). Here, the extent of skipping lots depends on the stability and level of expected incoming quality.
Both approaches are part of our acceptance control strategy and require a good quality history which would indicate not only the underlying level of processing quality (i.e., at the operator level) but also the extent of stability (i.e., degree of control) that can be expected in the process. Accordingly, the inspection process must provide:
- good data (accurate error estimates);
- quick results (monthly, weekly, daily);
- incentive for improvement (feedback reports);
- quality history (time series of error quality).

Essentially these have been the motivating influences in developing the Quality Control Processing System (QCPS). It should be noted that changes are currently being made to the system to expand the existing operator quality history. This should provide the data to enable greater implementation of spot checks and/or process control for selected operators with exceptional and stable performances.

## 3. SYSTEM DESCRIPTION

Based on the strategy identified above, the QCPS has been developed to achieve the following objectives:

- process any single acceptance sampling transaction;
- provide output by operator where each operator can be treated as the error source;
- provide feedback to four levels of staff with current and historical quality control information;
- support the acceptance control strategy by enabling the processing of skip-lot sampling results and providing an extensive operator quality history;
- support the major QC objectives of error correction and prevention while enabling inspection costs to continually be minimized.


### 3.1 Methodological Features

## a. Inspection Schemes

The system can process any quality control transaction resulting from the application of single acceptance sampling. This naturally includes normal, reduced and tightened plans as well as any skipped lots resulting from skip-lot sampling. The system will also process any lot whose plan designation is $100 \%$ inspection.

## b. Lot Status Codes

The system determines the treatment of incoming QC transactions by using lot status codes which indicate the state of completeness of the intended inspection. There are codes for the following lot situations:

- sample inspected and accepted;
- sample inspected and rejected (remainder inspected);
- $100 \%$ inspected;
- any of the above not completed ( 3 codes);
- no sample inspection due to skip-lot.


## c. Attributive Quality Measures

The system will produce estimates for various quality measures which include percent defective, defects per hundred units and weighted error equivalents. For the latter quality measure, the system allows errors to be weighted according to a pre-defined error seriousness classification scheme. Typically, under these more complex measures, errors are categorized and assigned weights from 0 to 1 depending on their relative magnitude and seriousness. For purposes of simplicity, no more than four error categories are generally defined, as follows:

| Category | Weight |
| :---: | :---: |
| Critical | 1.0 |
| Major | $0.4-0.6$ |
| Minor | $0.2-0.3$ |
| Insignificant | $0.0-0.1$ |

## d. Estimates

The system provides estimates and their associated standard errors (where applicable) for many key quality control indicators. The most important of these are:
(i) Error Rates

Error rates are calculated which relate to the individual operator, a specific sampling plan or the overall application. These estimates are provided for various time frames (e.g., daily, weekly, monthly, quarterly, etc.), and various subsets of the application, such as specific lot categories (e.g., rejected lots) or sub-groups (e.g., regional offices).

## (ii) Operator Process Average

An estimate of an operator's processing ability at any particular point in time is provided by the operator process average. This estimate is calculated using an empirical Bayes approach (MacMillan and Mudryk 1988) which essentially shrinks the current operator sample error rate estimate part way towards the grand average error rate of the last four periods for that operator. The basis of shrinkage is determined by the ratio of the sampling variance of the current sample estimate to the total variance of the grand average estimate. This quantity has been found to produce good estimates for qualifying operators onto minimum inspection sampling plans.

## (iii) Rejection Rates

Actual and expected rates of rejection are calculated for each sampling plan for purposes of statistical comparison and operational evaluation. The expected rates are obtained assuming Poisson probabilities.
(iv) Inspection Rates

Inspection rates are calculated at various levels as a general indicator of relative costs. These rates are determined with and without skip-lot effects on an actual and expected basis. The expected rates are a natural extension of the expected rejection rates discussed above.

## (v) Average Outgoing Quality

An estimate is provided of the Average Outgoing Quality (i.e., AOQ) rate resulting from the application of quality control to the operation. This estimate projects the observed error rate at the operator level to the uninspected volume for that operator, and then aggregates all operators to determine the overall estimate.

## e. Analysis

The system provides tabulations and outputs which enable analyses to be performed at various levels which help to subsequently fine tune the application parameters and/or modify the plans. These include:

- operator profiles that enable a sampling plan/procedure qualification analysis;
- individual sampling plan evaluations that provide an overall QC plan analysis;
- summaries of key indicators that enable a QC cost-benefit analysis;
- a Pareto analysis of operator and error code contributions;
- group charts of operator process averages that provide an operations performance analysis.


## f. Reports

The system produces 8 reports and 5 graphical outputs (through its link to SASGRAPH) for each application run. Tabulations can also be produced for specified sub-groups (e.g., Statistics Canada's regional offices) with a summarizing feature over all sub-groups of each report.

Each set of output reports is designed for and disseminated to four levels of staff, namely: operator, supervisor, manager and QC designer. Examples of the output reports are available from the author.

### 3.2 Software Features

## a. Operator Capacity

For each application, the system can handle up to 108 operators in its historical file, each containing up to three previous periods of error information. A unique self-maintaining feature of this file is that any operator who has not been active during at least one of the last 4 consecutive months of processing is dropped. This makes room for new operators on the file and thereby increases the effective file capacity.

## b. Historical Updates

The system updates each operator error quality history (of up to 4 consecutive periods) with new information as it becomes available. This is currently being increased to 6 consecutive time periods. If an operator has not processed during a particular month, blank data for that month is inserted. Likewise, application year-to-date and quarterly totals are updated with the addition of each new month of QC data.

## c. Year-End Rollover

Most of the QCPS applications are maintained on a calendar year basis. When this option is specified, the system will zero out the previous monthly totals and commence a new application time series (usually starting in January). The quarterly totals and the operator error quality time series however, are not re-set at this time and continue to be maintained as usual.

## d. Recovery

If a tabulation run is made and errors are subsequently discovered, another run can be made using the recovery feature with the corrected data, to automatically produce the corrected outputs.

## 4. SYSTEM BENEFITS

The QCPS is aimed at servicing the needs of four levels of staff which interface with each QC application. Accordingly, the major achievements of this system can best be described under these same headings:

## a. Operator Level

The QCPS provides extensive feedback to the individual processing operators on their current and historical performance. The operators are then able to track their own progress, compare their own performance with that of their peers, and identify explicitly where their errors are being made. The result of this feedback generally leads to:

- improvement in operator processing ability;
- increased motivation with respect to peers;
- greater quality consciousness;
- higher operator morale.


## b. Supervisor Level

The system provides operational information to the supervisors which enables them to better manage their operation in terms of:

- effective resource allocation and work distribution;
- identifying problem operators and/or areas;
- determining training needs.


## c. Management Level

The system provides data summaries on key quality control indicators for management which enables them to:

- receive an assurance of quality;
- track the progress of the application in terms of quality and costs;
- recommend changes to operational objectives.


## d. QC Design Level

The system provides extensive information (e.g., estimates, quality histories) which is used to analyze the quality control design and fine tune or enhance the methods and procedures of each application. When this data has been established and maintained over a sustained period of time, it can lead to:

- improvements in QC methodologies and procedures;
- sampling plan and/or inspection procedure adjustments;
- minimization of inspection costs.


## 5. CONCLUSIONS

The QCPS is being used at Statistics Canada to support the Quality Control programs of many production oriented survey processing operations. As the ultimate aim of each program is to exercise error prevention to the extent possible, as well as, to progressively reduce inspection to the level of spot checks, a good and flexible processing system is essential. The QCPS achieves these objectives by providing good data and quick results to the various levels of staff that are involved in each operation, as well as, supporting the various inspection methods that fall under the general strategy of Acceptance Control.

The system is particularly attractive to our user community since it can easily handle large volume operations involving many operators, quickly and at a low cost. Furthermore, by treating each operator individually, the system focuses attention to each relevant error source and supports this with necessary feedback to the appropriate levels of staff. In this manner the system enables our quality control methods to be both preventive and corrective in an efficient and economical manner.

## ACKNOWLEDGEMENTS

The author would like to thank the referees and Jeffrey Smith for their constructive comments in reviewing this paper. Examples of output reports relating to this system are available from the author.

## REFERENCES

DODGE, H.F. (1950). Inspection for quality assurance. Industrial Quality Control, 7(1), 8.
MacMILLAN, J.H., and MUDRYK, W.V. (1988). A non-parametric Empirical Bayes approach for estimating a process average in quality control. Paper presented for the Section on Physical and Engineering Sciences, American Statistical Association Annual Meeting, New Orleans, Louisiana.
MUDRYK, W.V., and BOUGIE, R.W. (1987). Quality control processing system (QCPS) - Users Manual. Internal document. Statistics Canada, Ottawa.
SCHILLING, E.G. (1982). Acceptance Sampling in Quality Control. New York: Marcel Dekker.
STATISTICS CANADA (1987). Quality Guidelines. 2nd Edition, Ottawa, Canada.
STEPHENS, K.S. (1982). How to Perform Skip-Lot and Chain Sampling. Volume 4, ASQC Basic
References in Quality Control: Statistical Techniques. American Society for Quality Control, Milwaukee, Wisconsin.

# Postal Address Analysis 

YVES DeGUIRE ${ }^{1}$


#### Abstract

When we examine postal addresses as they might appear in an administrative file, we discover a complex syntax, a lack of standards, various ambiguities and many errors. Therefore, postal addresses represent a real challenge to any computer system using them. PAAS (Postal Address Analysis System) is currently under development at Statistics Canada and aims to replace an aging routine used throughout the Bureau to decode postal addresses. PAAS will provide a means by which computer applications will obtain the address components, the standardized version of these components and the corresponding Address Search Key (ASK).


KEY WORDS: Postal addresses; Administrative data; Parsing; Standardization; Search key.

## 1. INTRODUCTION

Postal address analysis can be defined as the process of identifying the basic components of an address which appears in free format, standardizing those components, and generating an identifier for that address. This process can be used, for example, in the pre-processing step of any record linkage application that uses an address field or in the generation of a key for database access. Statistics Canada, as part of its 1991 census research program, is conducting a study on the implementation of a national Address Register. Such a register contains basically, postal address information. This information must be analyzed carefully in order to produce a register and to assess its quality. The Address Register Research Team has recognized that fact and research into the area of automated postal address analysis was initiated.

This paper presents the results of this research on postal address analysis. The nature of an address and its related problems will be described. Also, some computer considerations will be discussed to explain why new software is needed for the Address Register and Statistics Canada. Finally, we will examine PAAS (Postal Address Analysis System); a system currently under development at Statistics Canada.

## 2. POSTAL ADDRESSES: STATEMENT OF THE PROBLEM

A postal address can be defined as a string of characters representing a location where an individual can pick up his mail. By location, we mean a physical place where the deliverer (like a postman) and the receiver agree in the matter of mail reception. It can be a dwelling, a postal box, a street or a rural route. To restrict our field of study, we are going to examine the addresses that are Canadian (French and English), that represent residential locations and that should result in correct mail delivery.

[^17]As one would expect, the flexibility in the address definition results in problems for any computerized application having to deal with postal addresses. Even a person is likely to encounter some problems with addresses with which he/she is not familiar. Three major problems are analyzed here.

### 2.1 The Syntax of a Canadian Postal Address is Complex

A postal address is composed of tokens (lexical items which can be considered as basic units in an address). A token can be either a delimiter, a term (or keyword), a word, a letter or a number. Figure 1 illustrates an example of token decomposition. Tokens can be combined to get address components which are larger address structures. In turn, a component can fall into three groups: designators, qualifiers and secondary words. Figure 1 gives also an example of a component decomposition. Valid addresses are composed of both a set of valid combinations of components and a set of valid combinations of tokens. However, it is more practical for implementation purposes, to define an address with token patterns (combinations of tokens). Token patterns can be generated from a formal postal address grammar (written in BNF for example) and used directly for constructing a postal address.

This syntax is fairly complex. First of all, the grammar is sizeable. We have analyzed a national sample of 30,000 addresses taken from six different administrative files. In these addresses, we found around 4,900 different token patterns. This is substantially higher than what is reported in Drew(1987) because we have analyzed addresses from many different files, not just one. Other interesting results concern the distribution of those patterns. Only 37 patterns are necessary to cover $50 \%$ of the addresses. So, there are a few common patterns, but most of the patterns are rather rare. Nevertheless, this analysis illustrates the complexity of postal address syntax by demonstrating that it is not restricted to just a few patterns. Secondly, as much as 600 different terms can be found in a good national sample of addresses. Thirdly, an address is usually in free format, i.e. the components (and the delimiters) can occur in any one of several positions.

### 2.2 Addresses Don't Follow Precise Standards

Addresses representing the same address location can be written in many ways as illustrated in Figure 2. The reason for this situation is the flexibility in postal address syntax and also human nature. In fact, people write addresses as they like and follow the "standards" in use in their immediate environment.


Figure 1. Two ways of decomposing a postal address.

### 2.3 Ambiguities Occur in Postal Addresses

A postal address can't be regarded only from a syntactic point of view. Its semantic (i.e. the meaning a postal address) must be examined as well. Sometimes, one address can potentially represent more than one location. We then face an ambiguity since we don't know how to interpret it. To do so, more knowledge is required in order to exclude the locations that don't exist and to identify the correct location. However, this knowledge doesn't always permit us to narrow down the location; we then face an unresolvable ambiguity. Figure 3 shows an example of an ambiguous address.

## 3. COMPUTER SYSTEMS CONSIDERATIONS

Now that we have a better understanding of postal addresses as well as their related problems, we will concentrate on the use of postal addresses in computer systems.

### 3.1 Computer Applications Requiring Address Information

Several types of application require address information. Some record linkage projects link individuals or dwellings (like in the construction of an Address Register) based on their postal addresses. Their linkage rules perform essentially on standardized address components. On the other hand, databases and computer files storing postal addresses are numerous. For example, postal addresses information for an Address Register must be stored in some fashion, either in a stand alone flat file or in some kind of integrated database. But what information is stored? Address components (standardized or not) could be. For follow-up or historical purposes, the original input address could be kept as well. However, retrieval from a large database (or a large flat file) requires an Address Search Key (ASK) to allow direct access (or direct matching) to a record identified by a postal address. Mailing labels processing is another area where postal addresses is a big concern. Address components, standardized or not, can form mailing labels.

1) 32 main st apt \#1, Ottawa, Ontario
2) 32 main st apt \#1, Ott., Ontario
3) 32 main st 1, Ottawa, Ontario

4) 860 first st, Ottawa, Ontario
5) 8601 st, Ottawa, Ontario
6) 8601 st, Ott., Ont

Figure 2. Examples of Addresses Which Represent the Same Location.


1) Apt 976, Fort St-John, BC
2) Apt 976 Fort, St-John, BC
3) Apt 976 Fort ST, John, BC

if at least two are an existing address, we have an unresolvable ambiguous address.

Figure 3. Example of an Ambiguity.

### 3.2 Three Basic Information Components

Therefore, three basic information components need to be derived from a free format postal address: the address components, the standardized components and the Address Search Key (ASK).

## 1. THE ADDRESS COMPONENTS

They represent recognizable and useful portions of an address. The major address components are street number, street name, street direction, street designator, postal designator, postal qualifier, municipality name, province name, and postal code.

## 2. THE STANDARDIZED COMPONENTS

They are the standardized version of the address components, where any style variations are removed.

## 3. THE ADDRESS SEARCH KEY (ASK)

This is a compressed string, unique for a given address.

### 3.3 Postal Address Analysis System

A complete Postal Address Analysis System (a computer system that generates the three basic information components we need) represents an expert system in the field of postal addresses. Expert because you replace a specialist (like a postman) in address recognition. At Statistics Canada in the 1970's, two routines were developed to analyze postal addresses. ENCODA (component decompositon) and ASKGEN2 (standardization and ASK) were implemented for the Business Register Maintenance System. They served well until recently. With the advent of powerful computers, new software development techniques and the Address Register itself, they don't perform to today's standards.

- The encoding success rate is too low. A study using a national sample of addresses from many administrative files shows that ENCODA cannot properly decode an address, on average, $15 \%$ of the time. This is not acceptable since it could lead, in the case of the creation of a national Address Register, to over one million encoding failures.
- The user interface is poor. There is no comprehensive status produced at the completion of the analysis. As well, very few utilities are provided in order to ease programming burden.
- The functionality is incomplete. Standardized components and ASK are mixed up in the same data structure. Standardized components are truncated to allow data compression but ASK is very long because it is stored in fields of fixed length. Also, the software doesn't recognize address ambiguity.
- Maintenance of the software is a nightmare. New address patterns are difficult to incorporate into the routines because these are complex and tend to become more and more so with time. This is a sign of aging software.

To fulfill the requirements of an Address Register and of Statistics Canada in the area of address analysis, the development of a completely new system was initiated. The problem this time has been approached with expert system techniques, modular design and full scale implementation. This new system is called PAAS, for Postal Address Analysis System.

## 4. A POSTAL ADDRESS ANALYSIS SYSTEM: PAAS

PAAS is currently under development. Therefore, some results are preliminary, but in general very encouraging. We will review here the four basic functions of the system.

### 4.1 Address Parsing

The parsing function is the most important and complex function of PAAS. Here, PAAS accepts as input a free format address, scans it (breaks it into lexical items) and parses it (analyzes the syntax) to decode it into address components.

This parser generates the following items for every address processed (Figure 4 illustrates two examples of this output):

- A comprehensive Address Status code; such as V for valid, E for syntax error, etc.
- Identification of components in the input address.
- Components classification: every component is classified using a detailed code, so it is easy to understand the meaning of a component. This code is divided into three sub-codes:
- TYPE code: indicates the group of components to which a component belongs. Example of TYPES are those for province (PR), municipality (MU), street (ST), etc.
- CAT code: refines the group of components indicated by TYPE. Examples for the street TYPE (ST) are name (NA), number (NU), designator (DE), etc.
- CLASS code: classifies a component by examining its characteristics. Examples are avenue (AV) or road (RD) classification of a street designator.
- Ambiguity detection: the PAAS parser flags any component that could change because of an ambiguity.

The PAAS parser was implemented using MPL. MPL is a meta-programming language. It allows us to generate programs or subroutines used for syntax analysis and automatic translation. The input to MPL is a set of specifications divided into the scanning (token recognition), the syntax rules and the semantics. The scanning represents the lexical analysis where the input is broken down into tokens. The syntax specification is similar to a BNF grammar specifications: the right-hand side symbols of a syntax rule are defined by the left-hand side symbols. Figure 5 gives examples of syntax rules. Finally, a semantic action can be associated with any rule and is used to handle some complex aspects of the syntax, as well as to perform other actions (such as updating a table of components). The MPL language is well suited to writing translation specifications and has been used at Statistics Canada to implement STATPAK (retrieval

|  | ADDRESS | COMPONENT | TYPE | CAT | CLASS | AMB_FLAG |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| (1) | 32 Main st, Ottawa, Ont | 32 | ST | NU | ** |  |
|  |  | Main | ST | NA | ** |  |
|  | ADDRESS_STATUS $====>\mathrm{V}$ | st | ST | DE | ST |  |
|  |  | Ottawa | MU | NA | * |  |
|  |  | Ont | PR | NA | 35 |  |
| (2) | 32 Main st Ottawa Ont | 32 | ST | NU | ** |  |
|  |  | Main | ST | NA | ** |  |
|  | ADDRESS_STATUS $====>\mathrm{A}$ | st | ST | DE | ST | * |
|  |  | Ottawa | MU | NA | ** | * |
|  |  | Ont | PR | NA | 35 |  |

Because the second example misses the commas to delimit the address, an ambiguity is flagged by PAAS.
Figure 4. Examples of the PAAS Parser Outputs.
and tabulation system for the census), NYSIIS (name encoding routine) and NAMEPARS (name parser). It saves development time (e.g. you don't need to write a detailed and custom program in a traditional programming language such as PL/1). The specifications in BNF are much easier to understand than is a program with a complex logic.

The PAAS parser involves a rather complicated syntax analysis and represent a fairly important MPL application. For example, a dictionary containing more than 600 terms assist in the scanning of addresses. As well, more than a hundred syntax rules implement the syntax analysis. In this syntax analysis, the initial tokens are transformed from a rule right- hand side to a rule left-hand side and become higher level address fragments (this is known as forward chaining) until the address is completely analyzed. During this process, the address components are identified and stored in a table by the semantic action of a rule. The invalid addresses are found whenever no rule is applicable. A sample set of rules to decode an address is illustrated in Figure 5. Finally, for some complex addresses, a special analysis is peformed through the use of the MPL semantic facility. This is required anytime an ambiguous term is encountered. In this case, PAAS analyses the surroundings of the ambiguous term.

In comparison with ENCODA, the PAAS parser is an improvement in the following are as:

- The quality of the parsing: the PAAS parser is able to decode more addresses successfully than ENCODA does. A series of parallel runs over identical national samples of addresses showed that PAAS is successful on more than $97 \%$ of addresses, while ENCODA properly handles only $85 \%$ of them.
- The indication of an address status: the status is more complete than ENCODA's which provides for only two possibilities: decoded address or blank address!
- The components: PAAS generates much more comprehensive component information than does ENCODA.
- The maintenance: the utilization of MPL helps in making the PAAS parser a lot easier to maintain than a huge algorithm such as is used by ENCODA.


### 4.2 Components Standardization

The standardization aims to remove any style variation in the address components defined in the parsing phase.

Unlike ASKGEN2, PAAS doesn't truncate any component and retains all the information in the components. This standardization is achieved basically in three different ways depending on the nature of the component:

1. CODABLE COMPONENTS

Every component for which a limited number of values exist is standardized by replacing its value with the CLASS code of the component (this code uniquely identifies the standardized value of the component). Falling into this category are components such as the province name, street designator, etc.
2. NAME COMPONENT NOT NUMBERED

To standardize a non-numbered name component, several rules must be applied to transform the original value into a standardized value. The rules vary from the removal of useless characters (e.g. quote, hyphen, etc.) to abbreviation replacement (e.g. Mtl becomes Montreal).
3. NAME COMPONENT NUMBERED

A numbered name component is standardized by returning its name as a number. For example First becomes 1, Second 2, etc.

## Ask

The Address Search Key should be unique and short.
Uniqueness is accomplished by concatenating in a pre-determined order the standardized components of an address (rather than a table as with ASKGEN2). We must note here that the ASK doesn't necessarily represent a unique identifier for dwellings. In rural areas for example, a postal address quite often represents many dwellings (e.g. RR \#1 Ottawa Ontario).

Address to parse: 100 Rideau st Ottawa Ont K1N5X2 At some point, we have a string of address fragments which will be transformed by five rules. The " $\mid$ " denotes a "OR" and [] is an optional syntax element.


Figure 5. Rules for a Sample Address Syntax.

To shorten the key, different compression techniques can be used. However, compression takes time and we have to choose a technique that will be efficient. We are experimenting with two different techniques.

## 1. TRUNCATION

here, the name components are truncated. This technique is not real compression and could affect the uniqueness of a key. However, it is simple and fast.
2. REAL COMPRESSION
a compression technique that we are looking at consists basically of replacing common combinations of characters by a character code not in use for writing an address. Here, we will preserve the uniqueness but increase the complexity of generating and using a key. Therefore, a longer ASK calculation time is expected with this technique.

### 4.3 Ambiguity Resolution

Once an ambiguity is determined from the parsing, it must be resolved, either manually, or automatically by the PAAS system. PAAS uses a municipality name file (this file covers the whole country with around 6000 names and has as its source in the Postal Code Directory tape from Canada Post) in an attempt to resolve an ambiguity.

This methodology is limited to the problems related to municipality names. This is not so bad since these problems account for a good portion of the ambiguous situations, and are easy to detect and to resolve (they don't involve a large amount of data). Future work could examine the usefulness of detecting and resolving more situations.

Finally, no matter how good the software becomes, the unresolvable and the non-existant addresses will remain a problem and should be followed-up manually.

## 5. CONCLUSION

The results of postal address analysis as accomplished by PAAS are encouraging. It decodes a vast majority of addresses, outputs a very informative code for every component, standardizes and generates an ASK properly, and handles ambiguities. Also, PAAS integrates utilities and interfaces for users and maintainers.

Users have access to an interface which processes their addresses through the four basic functions as well as a facility that handles the addresses in error (on-line processing). A file processor program is also provided.

Also integrated into PAAS is a quality assurance tool for PAAS maintainers. PAAS will evolve in the future with the discoveries of new addresses and obsolete addresses. Making sure that the changes to the system are applied properly is tricky. This maintainance tool ensures that a change to the software doesn't jeopardize any valid addresses properly analyzed in previous versions of the system.

## ACKNOWLEDGEMENTS

The author would like to thank J.P. Lozano and M. Vriends who worked on the implementation of PAAS, and B.E. Hill and M. Elsaesser for their help in writing this paper. Acknowledgement is also due to J. Armstrong who experimented with PAAS as well as providing comments on an earlier version of this paper.

## REFERENCES

BARRETT, WILLIAM A., BATES, RODNEY M., GUSTAFSON, DAVID A., and COUCH, JOHN D. (1986). Compiler Construction. Science Research Associates Inc.

CANADA POST CORPORATION (1986). Postal Codes Directory: Atlantic, Quebec, Ontario and Western regions.
DEGUIRE, Y. (1987). Research into the parsing and standardization of free format addresses at Statistics Canada. Internal report, Statistics Canada.
DREW, J. DOUGLAS, ARMSTRONG, JOHN, VAN BAAREN, ALEX, and DEGUIRE, YVES, (1987). Methodology for construction of address registers using several administrative sources. International Symposium on Statistical Uses of Administrative Data, Ottawa.
HILL, TED (1986). MPL A Translator Writing System. System Documentation, 1-4. Statistics Canada. LOZANO, J.P. (1987). Postal Address Analysis System Study. Internal Report, Statistics Canada.
STATISTICS CANADA (1986). Record Linkage Software User Guide. System documentation, Research and General Systems.
STATISTICS CANADA (1988). Postal address analysis system (PAAS): Project charter (draft). Internal report. Research and General Systems.

# A Brief Note on SQL 

DAVID N. EMERY ${ }^{1}$


#### Abstract

This note portrays SQL, highlighting its strengths and weaknesses. KEY WORDS: Relational database management system; Database query language.


## 1. INTRODUCTION

A great deal of media attention has been focused on relational database management systems and SQL (pronounced see-quel), the most popular of the associated database query languages. To a large extent, SQL has been cast in the role of panacea for all the ills associated with data management. Unfortunately, this leads to a great deal of misconception on the part of potential users of SQL. These people are then sometimes disappointed with SQL when they eventually get a chance to use it.

The intent of this note is to clear up some of this misconception by providing a realistic portrayal of SQL, highlighting its inherent strengths and weaknesses. No attempt will be made to elaborate the advantages of the relational data model itself. These advantages have been adequately documented elsewhere (Date 1985).

## 2. SQL - WHAT IS IT?

The interaction which takes place between a user (whether systems developer or end user) and a database management system can be broadly categorized according to the function taking place:

- data definition;
- data control (i.e. authorization and control of data integrity);
- data retrieval; and,
- data modification (i.e. insert, update, and delete).

A database management system must provide interfaces for carrying out each of these functions. Depending on the particular system, these interfaces take the form of utilities, query languages, and/or subroutine libraries for programming languages.

SQL addresses these four functions in a single well-defined, rigidly structured language. SQL is the interface used to communicate, to the database management system, how relations (i.e. logical files or tables) are to be subdivided and/or combined to create new relations.

The key to understanding SQL's capabilities is an appreciation of the fact that SQL addresses exactly these four roles - no more and no less. Any other functionality must be supplied by the application which initiates the SQL statement.

Consider the following example. The table, DWELLING, contains information about dwellings such as number of occupants, type of dwelling, where it is located, type of heating,

[^18]and age of dwelling. In order to impute the type of dwelling, one might want to obtain a set of potential donor dwellings which are in the same geographic area, are the same age, and use the same heating method. The following SQL statement could be issued to obtain a donor set:

```
SELECT DWELLING_ID, TYPE_OF_DWELLING (Query 1)
    FROM DWELLING
    WHERE HEATING TYPE = 'GAS' AND
        AGE \(=20\) AND
        LOCATION_CODE = 'XYZ';
```

SQL does not provide a mechanism for manipulating the set of retrieved donor records. Selecting the n'th record, every second record, or a random record are all beyond the capability of SQL. Similarly, SQL has no mechanism for manipulating a table to affect its appearance on a terminal or printer. These are capabilities one would rightfully demand of a programming language, and hence the term database query language. Calling SQL a fourth generation language (4GL), then comparing it to products which incorporate only the data retrieval and data modification functions into a programming language, only adds to the confusion. It is really an apples and oranges comparison since both are 4GLs, but of very different flavours.

Given this very focused functionality, the obvious question then has to be - why all the fuss about SQL?

## 3. SQL - ITS BENEFITS

### 3.1 Implementation Transparency

A SQL query indicates nothing about how the data is actually organized and stored on the database. The query states what is to be retrieved, modified, or stored; the database management system determines the best way to do it. Issues such as:

- which data columns are indexed (a performance improvement feature);
- whether the table/column is actually stored or merely an execution time combination of other tables; and,
- the data's internal representation (i.e. floating point, packed decimal, binary)
have no bearing whatsoever on a SQL statement's syntax. Consequently, the user is immune to changes in the database's organization and structure. Changes to the underlying structure of the database can be made at will without changing the query. A query is immediately able to take advantage of improvements in the database structure or optimization algorithms.

Similarly, when formulating a SQL query the user does not specify the order in which processing is to take place to satisfy the query. That is the responsibility of the query processing software's optimization algorithms. This software evaluates the query against the current structure and organization of the database to determine the most efficient way of satisfying it.

### 3.2 Non-proprietary, Internationally Accepted Standard

Both the International Standards Organization (ISO) and the American National Standards Institute (ANSI) have recently adopted a common standard for SQL (ISO 1987). The existence of this standard, with a commitment to it by a number of relational database management system vendors, gives software developers access to a much broader market without significantly extra development effort. By building their applications on top of standard SQL, they have removed their reliance on a particular database management system. As a result, the creation
of software tools, built upon an interface to this standard version of SQL, has become a major growth industry. For example, natural language interfaces, fourth generation programming languages, data dictionary software, data entry/validation packages, and spreadsheet software, all layered on top of ANSI/ISO SQL, are beginning to appear on the market.

The active interest in SQL has also had a very positive impact on the SQL standard itself; it is continuing to evolve. The most recent draft revision to the ISO Standard for SQL incorporates the specification of referential integrity constraints into SQL's data definition statements. The significance of this extension to SQL is best illustrated by a further elaboration of the DWELLING example. Assume that the database also has a table PERSONS which contains detailed information about individuals including a dwelling code which indicates the dwelling where they currently reside. One might define a integrity constraint stipulating that each person must be associated with exactly one dwelling. Consequently, it would be an error to delete a DWELLING record which still had any PERSONS records referencing it, or to add a PERSONS record which referenced a nonexistent DWELLING record. Currently, logic to detect and prevent these inconsistencies must be inserted into each application program capable of deleting a DWELLING record. With the incorporation of referential integrity specifications into SQL, this program logic will no longer be required. The DBMS software assumes responsibility for detecting and terminating any attempt to remove a DWELLING record which still has associated PERSONS records.

### 3.3 Ease of Extension

One of the major differences between the various vendors' versions of SQL is the number and variety of supported functions. This is to a large extent due to the ease with which extra functionality can be incorporated into SQL, without change to its overall structure. For example, the SQL standard documents the grouping functions of average (AVG), maximum (MAX), minimum(MIN), enumeration (COUNT) and aggregation (SUM) for unweighted data. Referring again to the earlier DWELLING example, one could generate various summary statistics about number of occupants, broken down by geographic location:

```
SELECT AVG (NO_OF__OCCUPANTS), MAX (NO_OF_OCCUPANTS),(Query 2)
    MIN (NO_OF_OCCUPANTS), SUM (NO_OF_OCCUPANTS),
        COUNT (NO__OF_OCCUPANTS)
FROM DWELLING
GROUP BY LOCATION_CODE;
```

Some Vendors have augmented these functions with others such as variance (VARIANCE) and standard deviation (STDDEV). With these extra functions the identification of outliers, more than one standard deviation from the mean, is a very straightforward exercise:

```
SELECT DWELLING_ID FROM DWELLING
WHERE NO_OF_OCCUPANTS <
(SELECT AVG (NO_OF_OCCUPANTS) - STDDEV (NO_OF_OCCUPANTS)
FROM DWELLING)
OR
NO_OF_OCCUPANTS >
(SELECT AVG (NO_OF_OCCUPANTS) + STDDEV (NO_OF_OCCUPANTS)
FROM DWELLING);
```


### 3.4 Single Interface to the Database

When interrogating a database from within a host language program such as PL/1, FORTRAN, or C, one also uses SQL statements. These statements are virtually identical to
those used when interrogating the database interactively via a SQL statement processor. The only difference lies in the fact that the host language interface requires an additional INTO clause to indicate the program variables receiving the results of the query.

By using an identical interface to a host programming language, one is able to separate the program development and debugging exercise into two distinct activities:

- testing of the database retrieval storage statements (i.e. the SQL statements themselves), and - testing of the program code which manipulates the data.

The first of these activities can be carried out using a SQL command interpreter even before the host language program has been written. The optimal SQL statements can then be moved directly into the host program where the testing effort can be focused on the logic associated with manipulating the data.

Since the SQL statements embedded in the host language are interpreted at execution time, any changes made to the database organization or structure are immediately reflected in the program.

### 3.5 Suitability for Distributed Databases/Database Machines

One of the hottest topics in database management systems technology today is distributed databases. In a distributed database environment, the data is spread across a number of different databases (often on physically separate machines). It is the DBMS software's responsibility to intercept a user's query, translate it into appropriate queries to the various constituent databases, and assemble the results of these queries for presentation.

As discussed earlier, a SQL statement is devoid of constructs associated with describing how or where the data is stored on the database. Consequently, in a distributed database environment where SQL is used as the database query language, data can be moved between machines with no change whatsoever to existing applications. SQL is therefore becoming quite popular with the developers of distributed database management systems.

For similar reasons, SQL is gaining popularity as a query language for database machines. These machines take advantage of relational (i.e. tabular) data structures' inherent regularity to partition them across a number of parallel processors. These processors have instruction sets specifically designed to perform relational operations. The lack of representational detail in SQL queries completely insulates users from an awareness of what these machines are doing behind the scene.

## 4. SUMMARY

There is no question that SQL has quickly become the pre-eminent database query language. The database management system which does not feature a SQL interface will soon be the exception. An interesting anomaly will however emerge. The user will, over time, see less and less of SQL. Rather than trying to make SQL itself a user-friendly language, effort will be focused on the devlopment of application specific tools which provide the user with an interface tailored to the task at hand. SQL will be the common interface between these tools and the various databases.

## REFERENCES

DATE, C.J. (1985). An Introduction to Database Systems. Don Mills: Addison-Wesley.
INTERNATIONAL STANDARDS ORGANIZATION (1987). Database Language SQL. International Standards Organization 9075.

# A Bibliography on Randomized Response: 1965-1987 

GAD NATHAN ${ }^{1}$


#### Abstract

A comprehensive bibliography of books, research reports and published papers, dealing with the theory, application and development of randomized response techniques, includes a subject classification.


KEY WORDS: Survey; Sensitive issues; Confidentiality.

## 1. INTRODUCTION

The recent increase in requirements for extensive data on sensitive issues, (such as the detailed information on sexual behavior, necessary to study the spread of the AIDS epidemic), has lead to renewed examination of the techniques available for obtaining answers to sensitive questions. The difficulties of applying conventional survey techniques to obtain data on sensitive issues in a large-scalesurvey are well known and several alternative techniques have been proposed - Bradburn and Sudman (1979). The most prominent of these has been the randomized response technique, originally proposed by Warner (1965). The underlying idea is that the respondent uses a random mechanism to select the question to which he answers and the interviewer knows only the response itself, without knowing which question is being answered. This is supposed to reduce biases due to non-response and to response error, by assuring the respondent that his privacy is protected by the method (in that the question he is being asked is unknown to the interviewer) and thereby convincing him to cooperate more readily and to answer more truthfully than he might by a direct question.

Since 1965 a great deal of research into various aspects of the technique has been carried out. This includes theoretical developments, development of new randomization techniques and extensions to quantitative variables, to polytomous questions and to the multivariate case. Problems of estimation, optimization of design parameters and sample design, specific to randomized response, have also been dealt with. A large number of empirical studies using randomized response have been carried out in various application areas, such as studies of drug use, abortions, drunken driving and crime, many of them with some evaluation, often by validation studies. The experience in these studies is very divergent, with some showing marked gains due to the use of randomized response and others showing no gain at all in response rates or in response reliability. Respondents' attitudes to randomized response, their comprehension of the procedure, their perceptions of confidentiality and of the protection that the procedure provides have also been investigated, in attempts to understand the reasons for the differences in the empirical results.

This large body of research is scattered among over 250 theses, research reports, published papers and books, which have appeared, (in at least seven languages), over the last 20 odd years. These include many expository and survey papers and two bibliographies - Kim and

[^19]Flueck (1976) and Daniel (1979) - the latter an annotated one. Three comprehensive books on the subject - Defaa (1982), Fox and Tracy (1986) and Chaudhuri and Mukerjee (1988) have also appeared. Unfortunately none of these include a fully comprehensive and updated bibliography and the present one is an attempt to correct this lacuna.

Although an attempt has been made to be as comprehensive as possible, by including both published and unpublished papers, the latter are obviously covered only in as far as information about them was available from various sources. In addition, an attempt was made to reduce duplication by excluding unpublished reports or papers presented at meetings whose content is substantially included in a subsequently published paper. However, Ph.D. theses are generally included, since they usually have more detail than the papers derived from them. Papers about other survey methods for dealing with sensitive issues, which can be considered as alternatives to randomized response, are included only if they relate to a comparison of the alternative to randomized response. Papers dealing with randomization techniques to ensure confidentiality of data already collected (such as random rounding or encoding) are not included, unless they also relate to the use of randomization in the collection process itself.

The bibliography is arranged as an alphabetical listing, which gives full citation details in the standard way used for reference lists. Titles are given in the language of the paper or book, if known. Otherwise, for publications not in English, the title is given in English with a designation of the original language in parentheses. Most of the non-English papers include a summary or abstract in English. A series of letter codes on the right edge of the page, opposite each reference, indicates a classification by subject. The classification categories and codes are given below. An author index and a classified listing by subject, not included due to space limitations, are available from the author.

## 2. SUBJECT CLASSIFICATION CODES

A - Applications and field experiments.
B - Bibliographies and survey papers.
C - Confidentiality, respondent comprehension, attitude and protection.
E - Evaluation of alternative techniques or estimators.
H - Hypothesis testing, estimation and analysis.
M - Multivariate case.
O - Optimization of design parameters.
P - Polytomous questions.
Q - Quantitative variables.
R - Randomization devices and techniques.
S - Sample design.
T - Theoretical developments.
V - Validation studies.
X - Expository papers.

## ACKNOWLEDGEMENTS

The author would like to thank the many colleagues and authors of papers on randomized response who reviewed an earlier draft of this bibliography and provided comments, additional references and reprints of their papers. The preparation of this bibliography was partly supported by a National Science Foundation grant (No. SES-8612320).

## REFERENCES

ABERNATHY, J.R., GREENBERG, B.G., and HORVITZ, D.G. (1970). Estimates of
AC induced abortion in urban North Carolina. Demography, 7, 19-29.
ABUL-ELA, A.A. (1966). Randomized response models for sample surveys on human APT populations. Ph.D. thesis, University of North Carolina, Chapel Hill.
ABUL-ELA, A.A., and ABDEL-HAMIED, S.M. (1984). Randomized response ratio estimates: bias and efficiency. Proceedings of the Survey Research Methods Section, American Statistical Association, 794-799.
ABUL-ELA, A.A., and ABDEL-HAMIED, S.M. (1985). A randomized response ratio estimate from quantitative data. Proceedings of the Social Statistics Section, American Statistical Association, 300-305.
ABUL-ELA, A.A., and DAKROURI, H.M. (1980). Randomized response models: a ratio estimator. Proceedings of the Survey Research Methods Section, American Statistical Association, 205-208.
ABUL-ELA, A.A., GREENBERG, B.G., and HORVITZ, D.G. (1967). A multiproportions randomized response model. Journal of the American Statistical Association, 62, 990-1008.
ADHIKARI, A.K. (1982). On randomized response surveys with sensitive quantitative characters: a case study in the Indian Statistical Institute. Technical Report ASC826, Indian Statistical Institute, Calcutta.
ADHIKARI, A.K., CHAUDHURI, A., and VIJAYAN, K. (1984). Optimum sampling strategies for randomized response trials (with discussion). International Statistical Review, 52, 115-125.
AHSANULLAH, M., and EICHHORN, B.H. (1984). On scrambled response of sensitive quantitative data. Proceedings of the Survey Research Methods Section, American Statistical Association, 800-802.
ALBERS, W. (1982). Simple randomized response procedures with bounded respondent risk for quantitative data. Kwantitatieve Methoden, 8, 35-46.
ALEXANDER, J.R. (1978). Probability as an aid in social research: the randomized response technique. Mathematical Spectrum, 11, 10-13.
ANDERSON, H. (1975). Efficiency versus protection in randomized response designs. Ph.D. thesis, University of Lund, Sweden.
ANDERSON, H. (1976). Estimation of a proportion through randomized response (with discussion). International Statistical Review, 44, 213-217.
ANDERSON, H. (1977). Efficiency versus protection in a general randomized response model. Scandinavian Journal of Statistics, 4, 11-19.
BARKSDALE, W.B. (1971). New randomized response techniques for control of nonsampling errors in surveys. Ph.D. thesis, University of North Carolina, Chapel Hill.
BARKSDALE, W.B. (1975). New randomized response techniques for control of nonsampling errors in surveys. Proceedings of the Social Statistics Section, American Statistical Association, 302-304.
BARTH, J.T., and SANDLER, H.M. (1976). Evaluation of the randomized response technique in a drinking survey. Journal of Studies in Alcoholism, 37, 690-693.
BASULTO, J. (1982). The randomized response design of Warner: a superpopulation model (In Spanish). Estadistica Española, 96, 51-61.

BEGIN, G., BOIVIN, M., and BELLEROSE, J. (1979). Sensitive data collection through the random response technique: some improvements. Journal of Psychology, 101, 53-65.
BELDT, S.F., DANIEL, W.W., and GARCHA, B.S. (1982). The Takahasi-Sakasegawa randomized response technique. A field test. Sociological Methods and Research, 11, 101-111.

BELLHOUSE, D.R. (1980). Linear models for randomized response designs. Journal of the American Statistical Association, 75, 1001-1004.

BERMAN, J., McCOMBS, H., and BORUCH, R.F. (1977). Notes on the contamination method. Sociological Methods and Research, 6, 45-62.
BLANGIARDO, G.C. (1978). I campioni in blocco con risposta casualizzata. Rivista di Statistica Applicata, 11, 89-96.
BLANGIARDO, G.C. (1979). La stratificazione nei campioni in blocco con risposta casualizzata. Rivista di Statistica Applicata, 12, 26-36.
BLOMQVIST, N., and ERIKSSON, S.A. (1974). A general theory of randomized interviews. Research Report 1974:4, Department of Statistics, University of Gothenburg.
BORUCH, R.F. (1971a). Assuring confidentiality of responses in social research: a note on strategies. American Sociologist, 6, 308-311.
BORUCH, R.F. (1971b). Maintaining confidentiality of data in educational research: a systematic analysis. American Psychologist, 26, 413-430.
BORUCH, R.F. (1972). Relations among statistical methods for assuring confidentiality of social research data. Social Science Research, 1, 403-414.
BORUCH, R.F. (1982). Methods for resolving privacy problems in social research. In Ethical Issues in Social Science Research, (Eds. R.R. Faden, R.J. Wallace, and L. Walters), Baltimore: Johns Hopkins University Press.

BORUCH, R.F., and CECIL, J.S. (1979). Assuring the confidentiality of social research data. Philadelphia: University of Pennsylvania Press.

BOURKE, P.D. (1974a). Multi-proportions randomized response using the unrelated question technique. Report No. 74, Errors in Surveys Research Project, Institute of Statistics, University of Stockholm.
BOURKE, P.D. (1974b). Symmetry of response in randomized response designs. Report No. 75, Errors in Surveys Research Project, Institute of Statistics, University of Stockholm.
BOURKE, P.D. (1974c). Vector response in randomized response designs. Report No. 76, Errors in Surveys Research Project, Institute of Statistics, University of Stockholm.
BOURKE, P.D. (1975). Randomized response designs for multivariate estimation. Report No. 6, Confidentiality in Surveys Research Project, Institute of Statistics, University of Stockholm.
BOURKE, P.D. (1978). Randomized response designs with symmetric response for multiproportions estimation. Statistical Review, 16, 197-204.

BOURKE, P.D. (1981). On the analysis of some multivariate randomized response designs for categorical data. Journal of Statistical Planning and Inference, 5, 165-170.
BOURKE, P.D. (1982). Randomized response multivariate designs for categorical data. Communications in Statistics, Ser. A, 11, 2889-2901.
BOURKE, P.D. (1983). Randomized response designs with attribute-based randomization. Statistical Review, 5, 125-132.
BOURKE, P.D. (1984). Estimation of proportions using symmetric randomized response designs. Psychological Bulletin, 96, 166-172.

## R

AERV

HOQT

CR

S

S

MT

C

C

CE

CX

CX

PR

RT

T

HM

MR

EMR

MT

MR

MR

BOURKE, P.D., and DALENIUS, T. (1973). Multi-proportions randomized response
P using a single sample. Report No. 68, Errors in Surveys Research Project, Institute of Statistics, University of Stockholm.
BOURKE, P.D., and DALENIUS, T. (1974a). A note on inadmissable estimates in randomized enquiries. Report No. 72, Errors in Surveys Research Project, Institute of Statistics, University of Stockholm.
BOURKE, P.D., and DALENIUS, T. (1974b). Randomized response models with lying. Report No. 71, Errors in Surveys Research Project, Institute of Statistics, University of Stockholm.

BOURKE, P.D., and DALENIUS, T. (1976). Some new ideas in the realm of randomized inquiries (with discussion). International Statistical Review, 44, 219-221.
BOURKE, P.D., and MORAN, M.A. (1984). Application of the EM algorithm to randomized response data. Proceedings of the Survey Research Methods Section, American Statistical Association, 788-793.
BOURKE, P.D., and MORAN, M.A. (1986). An alternative EM formulation for randomized response data. Proceedings of the Survey Research Methods Section, American Statistical Association, 444-447.

BRADBURN, N., and SUDMAN, S. (1979). Improving interview method and questionnaire design. San-Francisco: Jossey-Bass, 1-13.
BRESSAN, F. (1983). Warner's scheme of sampling with randomized responses with memory (In Italian). Rivista di Statistica Applicata, 16, 85-98.
BREWER, K.R.W. (1981). Estimating marihuana usage using randomized response - some paradoxical findings. Australian Journal of Statistics, 23, 139-148.
BROWN, G.H. (1975). Randomized inquiry vs. conventional questionnaire method in estimating drug usage rates through mail surveys. Technical Report 75-14, Human Resources Research Organization, Alexandria, Virginia.
BROWN, G.H., and HARDING, F.D. (1973). A comparison of methods of studying illicit drug usage. Technical Report 73-9, Human Resources Research Organization, Alexandria, Virginia.
BUCHMAN, T.A., and TRACY, J.A. (1982). Obtaining responses to sensitive questions: conventional questionnaire versus randomized response technique. Journal of Accounting Research, 20, 263-271.
CAMPBELL, A.A. (1987). Randomized response technique. Science, 236, 1049.
CAMPBELL, C., and JOINER, B.L. (1973). How to get the answer without being sure you've asked the question. American Statistician, 27, 229-231.
CARR, J.W., MARASCUILO, L.A., and BUSK, P. (1982). Optimal randomized response models and methods for hypothesis testing. Journal of Educational Statistics, 7, 295-310.

CHAUDHURI, A. (1983). Randomized response technique to determine input in crop estimation. Calcutta Statistical Association Bulletin, 32, 208-210.
CHAUDHURI, A. (1987). Randomized response surveys of finite populations: a unified approach with quantitative data. Journal of Statistical Planning and Inference, 15, 157-165.
CHAUDHURI, A., and ADHIKARI, A.K. (1981). Sampling strategies with randomized response trials and their properties and relative efficiencies. Technical Report ASC815, Indian Statistical Institute, Calcutta.

CHAUDHURI, A., and MUKERJEE, R. (1985). Optionally randomized response techniques. Calcutta Statistical Association Bulletin, 34, 225-229.

CHAUDHURI, A., and MUKERJEE, R. (1987). Randomized response techniques: a review.
Statistica Neerlandica, 41, 27-44.
CHAUDHURI, A., and MUKERJEE, R. (1988). Randomized response: theory and techniques. New York: Marcel Dekker.
CHEN, E., CHOW, L.P., and LIU, P.T. (1974). Field studies on the new randomized response techniques. Department of Population Dynamics, Johns Hopkins University, Baltimore.
CHEN, T.T. (1978). Log-linear models for the categorical data obtained from randomized response techniques. Proceedings of the Social Statistics Section, American Statistical Association, 284-288.
CHEN, T.T. (1979). Analysis of randomized response as purposively misclassified data. Proceedings of the Survey Research Methods Section, American Statistical Association, 158-163.
CHI, I.C., CHOW, L.P., and RIDER, R.V. (1972). The randomized response techniques as used in the Taiwan outcome and pregnancy study. Studies in Family Planning, 3, 265-269.
CHOW, L.P., and LIU, P.T. (1973). A new randomized response technique: the multiple answer model. Department of Population Dynamics, Johns Hopkins University, Baltimore.
CHOW, L.P., GRUHN, W., and CHANG, W.P. (1979). Feasibility of the randomized response technique in rural Ethiopia. American Journal of Public Health, 69, 273-276.
CLICKNER, R.P., and IGLEWICZ, B. (1980). Warner's randomized response technique: the two sensitive questions case. South African Statistical Journal, 14, 77-86.
COHEN, J.E. (1987). Sexual behavior and randomized responses. Science, 236, 1503.
COMSTOCK, G.W., CONDE, J.G., and HELSING, K.J. (1985). A simple randomized response device. American Journal of Epidemiology, 122, 187-190.
CURLETTE, W.C. (1980). The randomized response technique: using probability theory to ask sensitive questions. Mathematics Teacher, 73, 618-621.
DALENIUS, T. (1977). Privacy transformations for statistical information systems. Journal of Statistical Planning and Inference, 1, 73-86.
DALENIUS, T. (1983). Randomized response. In Solutions to Ethical and Legal Problems in Social Research, New York: Academic Press, 237-248.
DALENIUS, T., and VITALE, R.A. (1979). A new randomized response technique for estimating the mean of a distribution. In Contributions to Statistics, Jaroslav Hajek Memorial Volume, (Ed. J. Jurechkova), Dordrecht: D. Reidel, 43-47.
DANERMARK, B., and SWENSSON, B. (1987). Measuring drug use among Swedish adolescents. Journal of Official Statistics, 3, 439-448.
DANIEL, W.W. (1979). Collecting sensitive data by randomized response: anannotated bibliography. Research Monograph No. 85, College of Business Administration, Georgia State University.
DAWES, R.M., and MOORE, M. (1980). Guttman scaling orthodox and randomized responses (In German). In Attitude Measurement, (Ed. F. Peterman), Götingen: Verlag für Psychologie, 117-133.
DAWES, R.M., and SMITH, T.L. (1985). Attitude and opinion measurement. In Handbook of Social Psychology, (3rd ed.), (Eds. G. Lindzey, and E. Aronson), Hillsdale: Erlabaum, 509-566.
DEFAA, W. (1982). Anonymisierte Bafragungen mit zufallsverschlüsselten Antworten: Die Randomized-Response-Technik (RRT). Frankfurt am Main: Verlag Peter Lang.

DELACEY, P.W. (1975). Randomized conditional response. Proceedings of the Social
DEVORE, J.L. (1977). A note on the randomized response technique. Communications T in Statistics, Ser. A, 6, 1525-1534.
DEVORE, J.L. (1979). Estimating a population proportion using randomized responses. X Mathematics Magazine, 52, 38-40.
DOWLING, T.A., and SHACHTMAN, R.H. (1975). On the relative efficiency of randomized response models. Journal of the American Statistical Association, 70, 84-87.
DOWNS, T., GILILAND, D.C., and KATZ, L. (1978). Probability in a contested election. American Statistician, 32, 122-125.
DRAGO, E. (1981). Estimate of the mean and the second moment of a population through randomized response sampling (In Italian). Rivista di Matematica per le Scienze Economiche e Sociali, 4, 49-58.
DRANE, W. (1975). Randomized response to more than one question. Proceedings of the Social Statistics Section, American Statistical Association, 395-397.
DRANE, W. (1976). On the theory of randomized responses to two sensitive questions. Communications in Statistics, Ser. A, 5, 565-574; Corrigenda (1976), 5, 1552.
DUFFY, J.C., and WATERTON, J.J. (1984). Randomized response models for estimating the distribution function of a quantitative character. International Statistical Review, 52, 165-172.
DURHAM, A.M., and LICHTENSTEIN, M.J. (1983). Response bias in self-report surveys: evaluation of randomized responses. In Measurement Issues in Criminal Justice, (Ed. G.P. Waldo), Beverly Hills: Sage Publications.
EDGELL, E. (1980). Additive constants model: a randomized response technique for eliminating evasiveness to quantitative response. Psychological Bulletin, 87, 304-308.
EDGELL, S.E., HIMMELFARB, S., and CIRA, D.J. (1986). Statistical efficiency of using two quantitative randomized response techniques to estimate correlation. Psychological Bulletin, 100, 251-256.
EDGELL, S.E., HIMMELFARB, S., and DUCHAN, K.L. (1982). Validity of forced responses in a randomized response model. Sociological Methods and Research, 11, 89-100.
EICHHORN, B., and HAYRE, L.S. (1983). Scrambled randomized response methods for obtaining sensitive quantitative data. Journal of Statistical Planning and Inference, 7, 307-316.
ELLEM, D., and HOWSON, A. (1979). Randomized response: a survey technique for sensitive questions. In Interactive Statistics, (Ed. D. McNiel), New-York: Elsevier NorthHolland, 193-207.
EMRICH, L. (1983). Randomized response techniques. In Incomplete Data in Sample Surveys, Vol. 2, (Eds. W.G. Madow, I. Olkin and D.B. Rubin), New York: Academic Press, 73-80.
ERIKSSON, S.A. (1973a). A new model for randomized response. International Statistical Review, 41, 101-113.
ERIKSSON, S.A. (1973b). Randomized interviews for sensitive questions. Ph.D. thesis, University of Gothenburg, Sweden.
ERIKSSON, S.A. (1976a). Some sampling theory for surveys with randomized response interviews. Report No. 8, Confidentiality in Surveys Research Project, Institute of Statistics, University of Stockholm.

ERIKSSON, S.A. (1976b). Regression analysis of data from randomized interviews. Report
No. 17, Confidentiality in Surveys Research Project, Institute of Statistics, University of Stockholm.
FERRARI, P. (1978). II test sequenziale di Wald nel campionamento con risposte casualizzate. Statistica, 38, 481-492.
FERRARI, P. (1984). Two-stage sampling with randomized response and unknown strata sizes (In Italian). Quaderni di Statistica e Matematica Applicata, 5, 5-19.
FIDLER, D.S., and KLEINKNECHT, R.E. (1977). Randomized response versus direct questioning: two data collection methods for sensitive information. Psychological Bulletin, 84, 1045-1049.
FIERING, M.B., and HOOPER, M. (1985). Analysis of disclosure avoidance procedures. MX Civil Engineering Systems, 2, 12-19.
FLIGNER, M.A., POLICELLO, G.E., and SINGH, J. (1977). A comparison of two randomized response survey methods with consideration for the level of respondent protection. Communications in Statistics, Ser. A, 6, 1511-1524.
FLUECK, J.A., and KIM, J.J. (1976). Bibliography for randomized response. Mimeo Series No. 33, Department of Statistics, Temple University.
FOLSOM, R.E. (1974). A randomized response validation study: comparison of direct and randomized reporting of DUI arrests. Final Report 25U-807, Research Triangle Institute, Research Triangle Park, North Carolina.
FOLSOM, R.E., GREENBERG, B.G., HORVITZ, D.G., and ABERNATHY, J.R. (1973). The two alternate questions randomized response model for human surveys. Journal of the American Statistical Association, 68, 525-530.

FOX, J.A., and TRACY, P.E. (1980a). A field-validation of a quantitative randomized response model (with discussion). Proceedings of the Survey Research Methods Section, American Statistical Association, 299-304.
FOX, J.A., and TRACY, P.E. (1980b). The randomized response approach: applicability to criminal justice research and evaluation. Evaluation Review, 4, 601-622.
FOX, J.A., and TRACY, P.E. (1984). Measuring associations with randomized response. Social Science Research, 13, 188-197.
FOX, J.A., and TRACY, P.E. (1986). Randomized response: a method for sensitive surveys. Beverly Hills: Sage Publications.
GERSTEL, E.K., BRUCE, J., FOLSOM, R.E., and DURHAM, J. (1974). The effectiveness of the Mecklenburg county alcohol safety action project. Mimeo Report, Research Triangle Institute, Research Triangle Park, North Carolina.
GERSTEL, E.K., MOORE, P., FOLSOM, R.E.; and KING, D.A. (1970). Mecklenburg county drinking driving attitude survey. Mimeo Report, Research Triangle Institute, Research Triangle Park, North Carolina.
GEURTS, M.D., ANDRUS, R.R., and REINMUTH, J.E. (1975). Researching shoplifting and other deviant customer bebavior using the randomized response design. Journal of Retailing, 51, 43-48.
GODAMBE, V.P. (1980). Estimation in randomized response trials. International Statistical Review, 48, 29-32.
GOODE, T., and HEINE, W. (1978). Surveying the extent of drug use. Survey Statistician, 1, 10-12.
GOODSTADT, M.S., and GRUSON, V. (1975). The randomized response technique: a test on drug use. Journal of the American Statistical Association, 70, 814-818.

GOODSTADT, M.S., COOK, G., and GRUSON, V. (1978). The validity of reported drug
AE
HOR
GOULD, A.L., SHAH, B.V., and ABERNATHY, J.R. (1969). Unrelated question randomized response techniques with two trials per respondent. Proceedings of the Social Statistics Section, American Statistical Association, 351-359.
GREENBERG, B.G., ABERNATHY, J.R., and HORVITZ, D.G. (1969). Application of the randomized response technique in obtaining quantitative data. Proceedings of the Social Statistics Section, American Statistical Association, 40-43.
GREENBERG, B.G., ABERNATHY, J.R., and HORVITZ, D.G. (1970). A new survey technique and application to the field of public health. Millbank Memorial Fund Quarterly, 484, Pt. 2, 39-55.
GREENBERG, B.G., ABERNATHY, J.R., and HORVITZ, D.G. (1986). Randomized response. In Encyclopedia of Statistical Sciences (Vol. 7), (Eds. S. Kotz, N.L. Johnson and C.B. Read), New York: John Wiley, 540-548.
GREENBERG, B.G., ABUL-ELA, A.A., SIMMONS, W.R., and HORVITZ, D.G. (1969). The unrelated question randomized response model: theoretical framework. Journal of the American Statistical Association, 64, 520-539.
GREENBERG, B.G., HORVITZ, D.G., and ABERNATHY, J.R. (1974). A comparison of randomized response designs. In Reliability and Biometry; Statistical Analysis of Lifelength, (Eds. F. Proschan and R.J. Serfling), Philadelphia: SIAM, 787-815.
GREENBERG, B.G., KUEBLER, R.R., ABERNATHY, J.R., and HORVITZ, D.G. (1971). Application of the randomized response technique in obtaining quantitative data. Journal of the American Statistical Association, 66, 243-250.
GREENBERG, B.G., KUEBLER, R.R., ABERNATHY, J.R., and HORVITZ, D.G. (1977). Respondent hazards in the unrelated question randomized response model. Journal of Statistical Planning and Inference, 1, 53-60.
GUNEL, E. (1985a). A Bayesian comparison of randomized and voluntary response sampling models. Communications in Statistics, Ser. A, 14, 2411-2435.
GUNEL, E. (1985b). On the design of randomized response sampling plan. Proceedings of the Survey Research Methods Section, American Statistical Association, 457-459.
HAYASHI, C. (1968). Response errors and biased information. Annals of the Institute of Statistical Mathematics, 20, 211-228.
HILMAR, N.A. (1968). Anonymity, confidentiality and invasions of privacy: the responsibility of the researcher. American Journal of Public Health, 58, 324-330.
HIMMELFARB, S., and EDGELL, S.E. (1980). Additive constants model: a randomized response technique for eliminating evasiveness to quantitative response questions. Psychological Bulletin, 87, 525-530.
HIMMELFARB, S., and EDGELL, S.E. (1982). Note on "The randomized response approach." Addendum to Fox and Tracy. Evaluation Review, 6, 279-284.
HIMMELFARB, S., and LICKTEIG, C. (1982). Social desirability and the randomized response technique. Journal of Personality and Social Psychology, 43, 710-717.
HOCHBERG, Y. (1975). Two stage randomized response schemes for estimating a multinomial. Communications in Statistics, 4, 1021-1032.
HORVITZ, D.G., GREENBERG, B.G., and ABERNATHY, J.R. (1975). Recent developments in randomized response designs. In A Survey of Statistical Design and Linear Models, (Ed. J.N. Srivastava), New-York: North Holland, 271-285.
HORVITZ, D.G., GREENBERG, B.G., and ABERNATHY, J.R. (1976a). Randomized response: a data gathering device for sensitive questions (with discussion). International Statistical Review, 44, 181-196.

HORVITZ, D.G., GREENBERG, B.G., and ABERNATHY, J.R. (1976b). The randomized
BX response technique. In Perspectives on Aititude Assessment: Surveys and their Alternatives, (Eds. H.W. Sinaiko and L.A. Broedling), Champaign: Pendelton Publications.
HORVITZ, D.G., SHAH, B.V., and SIMMONS, W.R. (1967). The unrelated question randomized response model. Proceedings of the Social Statistics Section, American Statistical Association, 65-72.
IGLEWITZ, B. (1976). A coding approach to the sensitive question problem. Proceedings of the Social Statistics Section, American Statistical Association, 414-415.
IIT Research Institute and the Chicago Crime Commission (1971). A study of organized crime in Chicago. IITRI Project No. H-6031, Report prepared for the Illinois Enforcement Commission, Chicago.
KAMMERMANN, L.A., GREENBERG, B.G., and QUADE, D. (1985). Selecting optimal values for $\pi_{y}$ in the unrelated question randomized response model. Proceedings of the Survey Research Methods Section, American Statistical Association, 470-475.
KIM, J.J. (1978). Randomized response techniques for surveying human populations. Ph.D. thesis, Temple University, Philadelphia.
KIM, J.J. (1987). A further development of randomized response for masking dichotomous variables. Proceedings of the Survey Research Methods Section, American Statistical Association, 239-244.
KIM, J.J., and FLUECK, J.A. (1976). A review of randomized response designs and some new results. Proceedings of the Social Statistics Section, American Statistical Association, 477-482.
KIM, J.J., and FLUECK, J.A. (1978a). An additive randomized response model. Proceedings of the Survey Research Methods Section, American Statistical Association, 351-355.
KIM, J.J., and FLUECK, J.A. (1978b). Modifications of the randomized response technique for sampling without replacement. Proceedings of the Survey Research Methods Section, American Statistical Association, 346-350.
KOCH, G.G., ABERNATHY, J.R., and IMREY, P.B. (1975). On a method for study of family size preferences. Demography, 12, 57-66.
KOLATA, G. (1987). How to ask about sex and get honest answers. Science, 236, 382.
KRAEMER, H.C. (1980). Estimation and testing of bivariate association using data generated by the randomized response technique. Psychological Bulletin, 87, 304-308.
KROTKI, K.J., and FOX, B. (1974). The randomized response technique, the interview and the self administered questionnaire: an empirical comparison of fertility reports. Proceedings of the Social Statistics Section, American Statistical Association, 367-371.
KROTKI, K.J., and McDANIEL, S.A. (1975). Three estimates of illegal abortion in Alberta, Canada: Survey, mailback questionnaire and randomized response technique. Bulletin of the International Statistical Institute, 46, 67-70.
KROTKI, K.J., and McDANIEL, S.A. (1978). La technique de réponse rendue aléatoire; quelques résultats d'une étude à Edmonton, Alberta. Population et Famille, 41, 91-119.
LAI, C.D. (1982). A review of randomized response survey models. Occasional Publications in Mathematics, 12, Department of Mathematics, Massey University, New Zealand.
LAMB, C.W., and STEM, D.E. (1978). An empirical validation of the randomized response technique. Journal of Marketing Research, 15, 616-621.
LANDENNA, G. (1983). Sampling with randomized responses: a general view (In Italian). Rivista di Statistica Applicata, 16, 5-14.

LANKE, J. (1975). On the choice of the unrelated question in Simmon's version of randomized response. Journal of the American Statistical Association, 70, 80-83.
LANKE, J. (1976). On the degree of protection in randomized interviews (with discussion). CE International Statistical Review, 44, 197-203.
LAVIN, P. (1974). A necessary and sufficient condition for asymptotic masking of the Warner MLE estimate. Report No. 79, Institute of Statistics, University of Stockholm.
LERNER, M. (1973). The collection of data on deviant behavior: public policy issues (Abstract). Bulletin of the International Statistical Institute, 45, 150.
LEVY, K.J. (1976a). Reducing the occurrence of omitted or untruthful responses when testing hypotheses concerning proportions. Psychological Bulletin, 83, 759-761.
LEVY, K.J. (1976b). The randomized response technique and large sample pairwise comparisons among the parameters of k independent binomial populations. British Journal of Mathematical and Statistical Psychology, 29, 257-262.

LEVY, K.J. (1977a). The randomized response technique and appropriate sample sizes for selecting the largest value of $\pi$ from among $k$ binomial populations. British Journal of Mathematical and Statistical Psychology, 30, 234-236.
LEVY, K.J. (1977b). The randomized response technique and comparisons among the parameters of k independent binomial populations. Psychological Bulletin, 84, 244-246.
LEVY, K.J. (1978). Sample size comparisons involving the randomized response technique. Journal of Experimental Education, 47, 21-23.
LEVY, K.J. (1980). The randomized response technique and large sample tests concerning the parameters of a multinomial distribution. Educational and Psychological Measurement, 40, 701-708.
LEYSIEFFER, F.W. (1975). Respondent jeopardy in randomized response procedures. Technical Report M338, Florida State University, Talahassee, Florida.
LEYSIEFFER, F.W., and WARNER, S.L. (1976). Respondent jeopardy and optimal designs in randomized response models. Journal of the American Statistical Association, 71, 649-656.
LIU, P.T., and CHOW, L.P. (1976a). A new discrete quantitative randomized response model. Journal of the American Statistical Association, 71, 72-73.
LIU, P.T., and CHOW, L.P. (1976b). The efficiency of the multiple trial randomized response technique. Biometrics, 32, 607-618.
LIU, P.T., CHEN, C.N., and CHOW, L.P. (1976). A study of the feasability of Hopkins randomized response models. Proceedings of the Social Statistics Section, American Statistical Association, 561-566.
LIU, P.T., CHOW, L.P., and MOSLEY, H.W. (1975). Use of the randomized response technique with a new randomizing device. Journal of the American Statistical Association, 70, 329-332.
LOCANDER, W., SUDMAN, S., and BRADBURN, N. (1976). An investigation of interview method, threat and response distortion. Journal of the American Statistical Association, 71, 269-275.
LOYNES, R.M. (1976). Asymptotically optimal randomized response procedures. Journal of the American Statistical Association, 71, 924-928.
MADIGAN, F.C., ABERNATHY, J.R., HERRIN, A.N., and TAN, C. (1976). Purposive concealment of death in household surveys in Misamis Oriental Province. Population Studies, 30, 295-303.
MARASINI, D. (1978). La stratificazione nel campionamento con riposte casualizzate.

MARASINI, D. (1981). The randomized response in the two-stage sampling scheme (In
S Italian). Quaderni di Statistica e Matematica Applicata, 4, 81-96.
MARASINI, D., and FERRARI, P. (1983). Sampling with randomized responses: estimation and hypotheses testing in case of stratified and two-stage sampling (In Italian). Rivista di Statistica Applicata, 16, 15-41.
MARASINI, D., and OLIVIERI, D. (1983). Randomized response models and the quality of statistical data (In Italian). Societa Italiana di Statistica, Atti del Convegno, 1, 489-513.
MATLOFF, N.S. (1984). Use of covariates in randomized response settings. Statistics and Probability Letters, 2, 31-34.
MAZZALI, A. (1983). A scheme of sampling with randomized responses in case of $k$ questions (In Italian). Rivista di Statistica Applicata, 16, 99-105.
McDANIEL, S.A., and KROTKI, K.J. (1979). Estimates of the rate of illegal abortion and the effect of eliminating therapeutic abortion, Alberta 1973-74. Canadian Journal of Public Health, 70, 393-398.
MILLER, J.D. (1984). A new survey technique for studying deviant behavior. Ph.D. thesis, George Washington University, Washington, DC.
MOORS, J.J.A. (1971). Optimization of the unrelated question randomized response model. Journal of the American Statistical Association 66, 627-629.
MOORS, J.J.A. (1981). Inadmissibility of linearly invariant estimators in truncated parameter spaces. Journal of the American Statistical Association, 76, 910-915.
MOORS, J.J.A. (1985). Estimation in truncated parameter spaces. Ph.D. thesis, Tilburg University, The Netherlands.
MORIARTY, M., and WISEMAN, F. (1976). On the choice of a randomization technique with the randomized response model. Proceedings of the Social Statistics Section, American Statistical Association, 624-626.
MUKERJEE, R. (1981). Inference on confidential characters from survey data. Calcutta Statistical Association Bulletin, 30, 77-88.
MUKHOPADHYAY, P. (1980a). A survey on the socio-economic conditions of some college students of Calcutta. Project Report, Indian Statistical Institute.
MUKHOPADHYAY, P. (1980b). On estimation of some confidential characters from survey data. Calcutta Statistical Association Bulletin, 29, 133-141.
MUKHOPADHYAY, P., and HALDER, A.K. (1980). Bayesian tables for Warner's randomized response probabilities. Technical Report ASC802, Indian Statistical Institute, Calcutta.
OLIVIERI, D. (1981). Le risposte casualizzate: stime, dimensioni ed errori campionari. Rivista di Statistica Applicata, 14, 79-98.
OLIVIERI, D. (1982). La diffusione della droga nelle scuole secondarie superiori di Verona. Vicenza, Italy: Cassa di Risparmio di Verona Vicenza e Belluno.
OLIVIERI, D. (1983a). On a modification of Simmon's scheme of sampling with randomized responses, with efficiency comparisons (In Italian). Rivista di Statistica Applicata, 16, 57-75.
OLIVIERI, D. (1983b). Stratified sampling with randomized responses and fixed alternative response (In Italian). Rivista di Statistica Applicata, 16, 77-84.
OLIVIERI, D. (1984). Estimation of parameters and efficiency in the Poole's randomized response model (In Italian). Societa Italiana di Statistica, Atti dela Riunione Scientifica, 32, 463-472.

OLIVIERI, D., and BRESSAN, F. (1984). A randomized response model with fixed alternative answer (In Italian). Rivista di Statistica Applicata, 17, 165-172.
ORWIN, R.G., and BORUCH, R.F. (1982). RRT meets RDD: statistical strategies for assuring response privacy in telephone surveys. Public Opinion Quarterly, 46, 560-571.
O'BRIEN, D.M., and COCHRAN, R.S. (1977). The comprehensio.. 'actor in randomized response. Proceedings of the Social Statistics Section, American Statistical Association, 270-272.

O'BRIEN, D.M., and COCHRAN, R.S. (1978). The effect of less than complete truthfulness on a quantitative randomized response model. Proceedings of the Business and Economic Statistics Section, American Statistical Association, 743-747.
O'HAGAN, A. (1987). Bayes linear estimators for randomized response models. Journal of the American Statistical Association, 82, 584-587.
PEARL, R.L., and FEDERER, W.T. (1975). Varying levels of probability for selecting sensitive questions using a randomized response technique. Proceedings of the Social Statistics Section, American Statistical Association, 584-587.
PITZ, G.E. (1980). Bayes analysis of random response models. Psychological Bulletin, 87, 209-212.
POHL, B.B., and POHL, N.F. (1975). Random response techniques for reducing nonsampling error in interview survey research. Journal of Experimental Education, 44, 48-53.
POLLOCK, K.H., and BEK, Y. (1976). A comparison of three randomized response models for quantitative data. Journal of the American Statistical Association, 71, 884-886.
POOLE, W.K. (1974). Estimation of the distribution function of a continuous type random variable through randomized response. Journal of the American Statistical Association, 69, 1002-1005.
POOLE, W.K., and CLAYTON, C.A. (1982). Generalizations of a contamination model for continuous type random variables. Communications in Statistics, Ser. A, 11, 1733-1742.
RAGHAVARAO, D. (1978). On an estimation problem in Warners's randomized response technique. Biometrics, 34, 87-90.
RAGHAVARAO, D., and FEDERER, W.T. (1979). Block total response as an alternative to the randomized response method in surveys. Journal of the Royal Statistical Society, Ser. B, 41, 40-45.
REASER, J.M., HARTSOCK, S., and HOEHN, A.J. (1975). A test of the forced alternative random response questionnaire technique. Technical Report 75-9, Human Resources Research Organization, Alexandria, VA.
REINMUTH, J.E., and GEURTS, M.D. (1975). The collection of sensitive information using a two-stage randomized response model. Journal of Marketing Research, 12, 402-407.

ROSENBERG, M.J. (1979). Multivariable analysis of a randomized response technique for statistical disclosure control. Ph.D. thesis, University of Michigan, Ann Arbor.
ROSENBERG, M.J. (1980). Categorical data analysis by a randomized response technique for statistical disclosure control (with discussion). Proceedings of the Survey Research Methods Section, American Statistical Association, 311-316.
ROSENBERG, M.J. (1985). An application of PROC FUNCAT to randomized response data. Proceedings of the SAS Users Group International Conference, 10, 1070-1075.

ROSENBLATT, R.R., and KELLY, E.L. (1978). A comparison of the sensitivity of the
unrelated question randomized response model with three other data accumulation techniques using examination cheating as a model. Proceedings of the Survey Research Methods Section, American Statistical Association, 356-361.
SAKASEGAWA, H., and TAKAHASI, K. (1974). Effects of repetition and finite population corrections in randomized response models (In Japanese). Proceedings of the Institute of Statistical Mathematics, 22, 59-67.
SAKASEGAWA, H., TAKAHASI, K., and SUZUKI, T. (1977). An investigation of a new randomized response model (In Japanese). Proceedings of the Institute of Statistical Mathematics, 25.
SCHEERS, N.J., and DAYTON, C.M. (1982). The covariate unrelated question randomized response model. Proceedings of the Survey Research Methods Section, American Statistical Association, 407-410.
SCHEERS, N.J., and DAYTON, C.M. (1986). RRCOV: Computer program for covariate randomized response models. American Statistician, 40, 229.
SCHMEIDLER, J. (1987). Assesssing quality of randomized response: were instructions followed? Proceedings of the Survey Research Methods Section, American Statistical Association, 245-249.
SEN, P.K. (1974a). On the estimation of basic distributions in randomized response models. Communications in Statistics, 3, 1081-1092.
SEN, P.K. (1974b). On unbiased estimation for randomized response models. Journal of the American Statistical Association, 69, 997-1001.
SEN, P.K. (1976). Asymptotically optimal estimators of general parameters in randomized response models (with discussion). International Statistical Review, 44, 223-224.
SHIMIZU, I.M., and BONHAM, G.S. (1978). Randomized response technique in a national survey. Journal of the American Statistical Association, 73, 35-39.
SHOTLAND, R.L., and YANKOVSKI, L.D. (1982). The random response method: a valid and ethical indicator of the "truth" in reactive situations. Personality and Social Psychology Bulletin, 8, 174-179.
SILVA, L.C. (1983). The randomized response technique: a general model for polytomous variables (In Spanish). Investigacion Operacional, 4, 75-100.
SIMMONS, W.R. (1970). Response to randomized inquiries: a technique for reducing bias. Administrative Applications Division Conference Transaction, American Society for Quality Control, Chapter 4-B.
SINGH, J. (1976). A note on the randomized response technique. Proceedings of the Social Statistics Section, American Statistical Association, 772.
SINGH, J. (1978). A note on maximum likelihood from randomized response models. Proceedings of the Social Statistics Section, American Statistical Association, 282-283.
SMITH, E.P., and SOSNOWSKI, T.S. (1972). Faculty evaluations by randomized response sampling. Journal of Experimental Education, 41, 70-72.
SMITH, L.L., FEDERER, W.T., and RAGHAVARAO, D. (1974). A comparison of three techniques for eliciting truthful answers to sensitive questions. Proceedings of the Social Statistics Section, American Statistical Association, 447-452.
SOEKEN, K.L. (1987). Randomized response methodology in health research. Evaluation and Health Professions, 10, 58-66.
SOEKEN, K.L., and MACREADY, G.B. (1982). Respondents' perceived protection when using randomized response. Psychological Bulletin, 92, 487-489.

SOEKEN, K.L., and MACREADY, G.B. (1985). Randomized response parameters as factors in frequency estimates. Educational and Psychological Measurement, 45, 89-100.
SPURRIER, J.D., and PADGETT, W.J. (1980). The application of Bayesian techniques in randomized response. In Sociological Methodology, San-Francisco: Jossey-Bass, 533-544.

STEM, D.E., and LAMB, C.W. (1981). The marble-drop technique: a procedure for gathering sensitive information. Decision Science, 12, 702-707.
STEM, D.E., and STEINHORST, R.K. (1984). Telephone interview and mail questionnaire applications of the randomized response model. Journal of the American Statistical Association, 79, 555-564.

SUZUKI, T., TAKAHASI, K., and SAKASEGAWA, H. (1976). Some notes on randomized response techniques (In Japanese). Proceedings of the Institute of Statistical Mathematics, 24, 1-13.
SWENSSON, B. (1972). Stratified randomized response with the special case: a combined use of regular interview and randomized response interview. Report No. 45, Errors in Surveys Research Project, Institute of Statistics, University of Stockholm.
SWENSSON, B. (1976a). A note on relations among one-sample randomized response techniques for dichotomies. Report No. 12, Confidentiality in Surveys Research Project, Institute of Statistics, University of Stockholm.
SWENSSON, B. (1976b). Combined independent questions versus randomized response, efficiencies under equal degree of protection. Report No. 15, Confidentiality in Surveys Research Project, Institute of Statistics, University of Stockholm.
SWENSSON, B. (1976c). Using mixtures of techniques for estimating sensitive attributes. Report No. 13, Confidentiality in Surveys Research Project, Institute of Statistics, University of Stockholm.
SWENSSON, B. (1977). Survey measurement of sensitive attributes. Ph.D. thesis, University of Stockholm.

TAKAHASI, K., and SAKASEGAWA, H. (1977). A randomized response technique without making use of any randomizing device. Annals of the Institute of Statistical Mathematics, 29, 1-8.

TAMHANE, A.C. (1977). A randomized response technique for investigating several sensitive attributes. Proceedings of the Social Statistics Section, American Statistical Association, 273-278.

TAMHANE, A.C. (1981). Randomized response techniques for multiple sensitive attributes. Journal of the American Statistical Association, 76, 916-923.

TRACY, P.E., and FOX, J.A. (1981). The validity of randomized response for sensitive measurements. American Sociological Review, 46, 187-200.
VERDOOREN, L.R. (1976). Loten bij delicate vragen: een overzicht van "randomized response'' technieken. Statistica Neerlandica, 30, 7-24.

WARNER, S.L. (1965). Randomized response: a survey technique for eliminating evasive answer bias. Journal of the American Statistical Association, 60, 63-69.

WARNER, S.L. (1971). The linear randomized response model. Journal of the American Statistical Association, 66, 884-888.
WARNER, S.L. (1976). Optimal randomized response models (with discussion). International Statistical Review, 44, 205-212.
WINKLER, R.L., and FRANKLIN, L.A. (1979). Warner's randomized response model: a Bayesian approach. Journal of the American Statistical Association, 74, 207-214.

ACE

AEHM AEV

B

EHRT

MQRT

CEOT

WISEMAN, F., MORIARTY, M., and SCHAFER, M. (1975). Estimating public opinion with the randomized response model. Public Opinion Quarterly, 39, 507-513.
ZDEP, S.M., and RHODES, I.N. (1976). Making the randomized response technique work. AE Public Opinion Quarterly, 40, 531-537.

ZDEP, S.M., RHODES, I.N., SCHWARZ, R.M., and KILKENNY, M.J. (1979). The AV validity of the randomized response technique. Public Opinion Quarterly, 43, 544-549.
'On the Stratification of Skewed Populations' by P. Lavallée and M.A. Hidiroglou, Survey Methodology (1988), 14, 33-43.

Formula (3.10), for the computation of $b^{\prime \prime}{ }_{(h)}$ should be

$$
b^{\prime \prime}{ }_{(h)}=\frac{-\beta_{h}^{\prime}+\sqrt{\beta_{h}^{\prime 2}-4 \alpha_{h}^{\prime} \gamma_{h}^{\prime}}}{2 \alpha_{h}^{\prime}}, h=1, \ldots, L-1 .
$$

Its finite population analogue on page 39 , should also be as above.

## ACKNOWLEDGEMENTS

The Survey Methodology Journal wishes to thank the following persons who have served as referees between February 1, 1988 and December 31, 1988. An asterisk indicates that the person served more than once.
M.G. Arellano, University of California

* J. Armstrong, Statistics Canada
T.R. Balakrishnan, University of Western Ontario
* M. Bankier, Statistics Canada
* D. Bellhouse, University of Western Ontario
L. Biggeri, University of Florence
J.-R. Boudreau, Statistics Canada
G. Brackstone, Statistics Canada
R.D. Burgess, Statistics Canada
R. Butcher, U.K. Office of Population Censuses and Surveys
M. Cairns, Office of Privatization and Regulatory Affairs
* R.G. Carter, Statistics Canada
D. Chapman, U.S. Bureau of the Census
* S. Cheung, Statistics Canada
N. Chinnappa, Statistics Canada
G.H. Choudhry, Statistics Canada
C. Courchesne, Bureau de la Statistique du Québec
M.L. Cohen, University of Maryland
F.R. Cronkhite, U.S. Bureau of Labor Statistics
P. Demmons, Statistics Canada
T. Dielman, Texas Christian University
D. Drew, Statistics Canada
J.L. Eltinge, Iowa State University
E.P. Ericksen, Temple University
V.M. Estevao, Statistics Canada
R.E. Fay, U.S. Bureau of the Census
R. Fecso, U.S. Department of Agriculture
I.P. Fellegi, Statistics Canada
M. Fluet, SOM Inc.
R.E. Folsom, Research Triangle Institute
J. Gambino, Statistics Canada
J.F. Gentleman, Statistics Canada
M.-F. Germain, Statistics Canada
M. Gonzalez, U.S. Office of Management and Budget
* G.B. Gray, Statistics Canada
S.G. Heeringa, University of Michigan
* M.A. Hidiroglou, Statistics Canada
T. Hill, Statistics Canada
R.R. Hocking, Texas A \& $M$ University
H. Hogan, U.S. Bureau of the Census
D. Holt, University of Southampton
T.B. Jabine, Statistical Consultant
J. Kadane, Carnegie-Mellon University
G. Kalton, University of Michigan
B.F. King, Educational Testing Service
B.F. Klugh, Jr., U.S. Department of Agriculture
P. Krishnan, University of Alberta
* S. Kumar, Statistics Canada
N.M. Lalu, University of Alberta
P. Lavallée, Statistics Canada
B. Lefrançois, Statistics Canada
H. Lee, Statistics Canada
J. Leyes, Statistics Canada
R. Little, University of California
L. Mach, Statistics Canada
K. Namboodiri, Ohio State University
J.L. O'Brien, U.S. Bureau of the Census
S.C. Puri, Agriculture Canado
B. Quenneville, Statistics Canada
* J.N.K. Rao, Carleton University
T.J. Rao, Indian Statistical Institute
J.G. Robinson, U.S. Bureau of the Census
* D. Royce, Statistics Canada
I. Sande, Statistics Canada
C. Särndal, University of Montreal
F. Scheuren, U.S. Internal Revenue Service
V.A. Sposito, Iowa State University
T.W.F. Stroud, Queen's University
A. Stuart, London School of Economics
J. Sullivan, U.S. Bureau of the Census
J.-L. Tambay, Statistics Canada
* V. Tremblay, Statplus
K. Wachter, University of California
M.J. Wenzowski, Statistics Canada
G.S. Werking, U.S. Bureau of Labor Statistics
W.E. Winkler, U.S. Bureau of the Census

Acknowledgements are also due to those who assisted during the production of the 1987 issues: C. VanBastelaar (Photocomposition), G. Gaulin (Author Services) and M. Haight (Translation Services).

We would like to thank the staff of Social Survey Methods and Business Survey Methods Divisions who assisted in proofreading and verification. Finally we wish to acknowledge J. Clarke, E. Corriveau, J. Dufresne, M. Kent, C. Larabie, D. Lemire and N. Smalldridge for their support with coordination, typing and copy editing.

# Applied Statistics 

## CONTENTS

Page
Space-time modelling with long-memory dependence: assessing Ireland's wind power resource J. Haslett and A. E. Raftery ..... 1
Case-weighted measures of influence for proportional hazards regression
A. N. Pettitt and f. Bin Daud ..... 51
Approximating the normal tail probability and its inverse for use on a pocket calculator J.-T. Lin ..... 69
Unweighted sum of squares test for proportions J. B. Copas ..... 71
On the role of cause-of-death data in the analysis of rodent tumorigenicity experiments L. E. Archer and L. M. Ryan ..... 81
A probabilistic model of squash: strategies and applications J. Simmons ..... 95
A note on 'random' purchasing: additional insights from Dunn, Reader and Wrigley B. E. Kahn and D. G. Morrison ..... 111
The use of guided reformulations when collinearities are present in non-linear regression J. S. Simonoff and C.-L. Tsai ..... 115Multiple-spell regression models for duration dataA. Hamerle127
Rotation of ill-defined principal componentsI. T. Jolliffe 139
Construction of row and column designs with contiguous replicates
E. R. Williams and J. A. John ..... 149
Statistical Software Reviews ..... 155
Statistical Algorithms
AS 242 The exact likelihood of a vector autoregressive moving average model
B. L. Shea ..... 161
AS. 243 Cumulative distribution function of the non-central $t$ distribution
R. V. Lenth ..... 185
AS 244 Decomposability and collapsibility for log-linear models Z. Geng ..... 189
Remarks
AS R76 A remark on Algorithm AS 215: Maximum-likelihood estimation of the parameters of the generalized extreme-value distribution A. J. Macleod ..... 198
AS R77 A remark on Algorithm AS 152: Cumulative hypergeometric probabilities
B. L. Shea ..... 199

## Correction

Correction to Algorithm AS 231: The distribution of a noncentral $\chi^{2}$ variable with nonnegative degrees of freedom
R. W. Farebrother204
Published in three parts per year. Annual subscription $£ 23.00$; single issues $£ 10.00$. The Algorithm Section is available separately-annual subscription $£ 4.00$. All communications should be addressed to the Executive Secretary, Royal Statistical Society, 25 Enford Street, London WIH 2BH, UK.

## GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 10, No. 2 and onward) of Survey Methodology as a guide and note particularly the following points:

## 1. Layout

1.1 Manuscripts should be typed on white bond paper of standard size ( $81 / 2 \times 11$ inch $)$, one side only, entirely double spaced with margins of at least $11 / 2$ inches on all sides.
1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
1.4 Acknowledgements should appear at the end of the text.
1.5 Any appendix should be placed after the acknowledgements but before the list of references.
2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.
3. Style
3.1 Avoid footnotes, abbreviations, and acronyms.
3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp( $\cdot$ )" and " $\log (\cdot)$ ", etc.
3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
3.4 Write fractions in the text using a solidus.
3.5 Distinguish between ambiguous characters, (e.g., w, $\omega ; 0,0,0 ; 1,1$ ).
3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

## 4. Figures and Tables

4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

## 5. References

5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching $a, b, c$ to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.


[^0]:    ${ }^{1}$ R.D. Burgess, Social Survey Methods Division, Statistics Canada, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, K1A 0T6.

[^1]:    ${ }^{1}$ A migrant is a person who at the time of the previous Census was living outside Canada, in a different province or in a different municipality (or CSD). RRC mobility data used here are those given by the RRC sample person in the Census and not those derived within the RRC (based upon a comparison of addresses).
    2 Statistics Canada 1983a.

[^2]:    ${ }^{1}$ Anatole Romaniuc, Director, Demography Division, Statistics Canada, Ottawa, Ontario.

[^3]:    ${ }^{1}$ Figures in brackets are Standard Deviations.
    Source: Demography Division, Statistics Canada.

[^4]:    ${ }^{1}$ Population of 5 years and over.
    Source: Demography Division, Statistics Canada.

[^5]:    'C.Y. Choi, D.G. Steel and T.J. Skinner, Australian Bureau of Statistics, P.O. Box 10, Belconnen, ACT, 2616,
    Australia.

[^6]:    (a) Actual location basis.
    (b) Demographic estimates based on 1921 Population Census and post 1921 demographic events.

[^7]:    ${ }^{1}$ Noel Cressie, Department of Statistics, Iowa State University, Ames, IA 50011.

[^8]:    ${ }^{1}$ Donald B. Rubin and Joseph L. Schafer, Department of Statistics, Harvard University, Cambridge, MA 02138, USA; Nathaniel Schenker, Division of Biostatistics, UCLA School of Public Health, Los Angeles, CA 90024, USA.

[^9]:    ${ }^{1}$ David J. Fein and Kirsten K. West, Undercount Research Staff, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C. 20233. The views expressed in this paper are those of the authors and do not necessarily reflect those of the Census Bureau.

[^10]:    **: p $<.01$

    * : p < . 05
    ${ }^{a}$ Log Likelihood $X^{2}=42.2, \mathrm{df}=45, \mathrm{p}=.5922$.

[^11]:    **: p < . 01
    *: p < . 05
    ${ }^{a}$ Log Likelihood $X^{2}=103.8, \mathrm{df}=150, \mathrm{p}=.9985$.

[^12]:    ${ }^{1}$ Mary Mulry, Undercount Research Staff, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C. 20233. Bruce Spencer, Department of Statistics, Northwestern University, Evanston, IL 60201 and NORC. The views expressed are attributed to the authors and do not necessarily reflect those of the Census Bureau.

[^13]:    ' Alan M. Zaslavsky, Statistics Center, Massachusetts Institute of Technology, Room E40-111, Cambridge, MA 02139, U.S.A. and Harvard University Department of Statistics, Cambridge, MA 02138, U.S.A.

[^14]:    I Jacques Lorigny, Administrateur à l'Institut National de la Statistique et des Études Économiques 18, Bld Adolphe Pinard 75675 PARIS CEDEX 14 (France).

[^15]:    ${ }^{1}$ M.J. Wenzowski, Research and General Systems, Statistics Canada, Room 2306, Main Building, Ottawa, Ontario, K1A 0T6.

[^16]:    ${ }^{1}$ This is a revised version of the paper presented at the Fourth Annual Research Conference, Bureau of the Census, Arlingion, Virginia, USA, March 1988.
    ${ }^{2}$ W.V. Mudryk, Business Survey Methods Division, Informatics and Methodology Branch, Statistics Canada, 10-J, Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6.

[^17]:    ${ }^{1}$ Yves DeGuire, Research and General Systems, Statistics Canada, room 2405, Main Building, Tunney's Pasture, Ottawa, Ontario, K1A 0T6.

[^18]:    ${ }^{1}$ David N. Emery, Statistics Canada, Research and General Systems Subdivision, Room 2405, Main Building, Tunney's Pasture, Ottawa, Ontario K1A 0T6.

[^19]:    ${ }^{1}$ Gad Nathan, Department of Statistics, Hebrew University, Mt. Scopus, 91905 Jerusalem, Israel. This bibliography was prepared while serving as Service Fellow at the U.S. National Center for Health Statistics, 3700 East-West Highway, Hyattsville, MD 20782.

