Statistics Canada Statistique Canada

SURVEY METHODOLOGY

A JOURNAL
OF
STATISTICS CANADA

VOLUME 11, NUMBER 1
JUNE 1985

Canada

Statistics Canada

SURVEY METHODOLOGY

A JOURNAL OF STATISTICS CANADA

June 1985

Published under the authority of
the Minister of Supply and
Services Canada

December 1985
8-3200-501

Price: Canada, $10.00, $20.00 a year
Other Countries, $11.50, $23.00 a year

Payment to be made in Canadian funds or equivalent

Catalogue 12-001, Vol. 11, No. 1

ISSN 0714-0045

Ottawa

SURVEY METHODOLOGY

A Journal of Statistics Canada

EDITORIAL POLICY

The Survey Methodology Journal will publish 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, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, smoothing and extrapolation methods, demographic studies, data integration and analysis and related computer systems development and applications. The emphasis will be on the development and evaluation of specific methodologies as applied to actual 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, Census and Household 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.

SURVEY METHODOLOGY

A Journal of Statistics Canada
Volume 11, Number 1, June 1985

CONTENTS

Survey Methodology, June 1985
Vol. 11, No. 1, pp. 1–14
Statistics Canada

Some Aspects of Nonresponse Adjustments

R. PLATEK and G.B. GRAY[1]

ABSTRACT

Unit and item nonresponse almost always occur in surveys and censuses. The larger its size the larger its potential effect will be on survey estimates. It is, therefore, important to cope with it at every stage where they can be affected. At varying degrees the size of nonresponse can be coped with at design, field and processing stages. The nonresponse problems have an impact on estimation formulas for various statistics as a result of imputations and weight adjustments along with survey weights in the estimates of means, totals, or other statistics. The formulas may be decomposed into components that include response errors, the effect of weight adjustment for unit nonresponse, and the effect of substitution for nonresponse. The impacts of the design, field, and processing stages on the components of the estimates are examined.

KEY WORDS: Nonresponse; Imputation; Estimation.

1. INTRODUCTION

As survey data are gathered from sampled unit, unit and item nonresponse will occur for at least some units despite all efforts to avoid it. The problem of dealing with nonresponse and the resultant missing data is two-fold. First, the effort through callbacks, repeated mailings etc. must be determined to the extent that it is cost-effective in reducing the mean square error of survey data and second, for the remaining nonresponse, the adjustments for the missing data must be obtained in order to reduce the nonresponse bias.

The field or survey centre effort to reduce or minimize unit nonresponse often means repeated attempts to contact selected units until a responsible person is available to reply to the survey questionnaire. The attempts pertain either to personal or telephone interview. In the case of mail surveys, repeated attempts mean successive mailings of a survey questionnaire to nonresponding units. In some cases, the repeated attempts may result in telephone or personal follow-ups. Some nonresponse is inevitable although every reasonable attempt should be made to minimize its levels. Thus, there will always remain some nonrespondents for whom all the efforts to convert them seem insufficient or inappropriate. The result is some imputation procedure to account for the missing data. This paper addresses the problems of controlling nonresponse at the design and field stage, followed by an examination of nonresponse adjustments at the processing stage. The examination will consider the feasibility and the practical as well as the methodological issues pertaining to the nonresponse adjustments.

Item nonresponse is often a more complex problem to deal with than unit nonresponse which is the type mostly referred to above. The most important factors which may reduce item nonresponse are good questionnaire design and a high quality of interviewers through proper hiring and training. A poorly designed questionnaire may also result in problems of following or completing the proper sequence of questions, whether by an interviewer or in a self-interview situation. Consequently, item nonresponse may occur in a questionnaire without the interviewer or respondent being aware of it. In addition, respondents may be willing to answer some but not all questions in a survey. Whatever the reason for missing items, the problems of substituting for them remains. Usually, a survey organization is unwilling to throw out whatever information

[1] R. Platek and G.B. Gray, Census and Household Survey Methods Division, Statistics Canada, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario K1A 0T6.

has been obtained unless of course the responses to major items appear very faulty or illogical. Thus, other means of imputing for missing items while maintaining the partial information on the records are usually undertaken.

Various statistics are required from a survey or census to explain social phenomena, determine socio-economic policies, etc. These include means, totals, ratios, distributions, percentiles and graphs. The statistics are assumed to be based on a universe of N units that belong to the target population; where N may or may not be known.

It may be demonstrated that all of the statistics mentioned above may be expressed in terms of totals or counts. Consequently, the remainder of the article will deal with missing data as they affect estimates of totals and counts in surveys. Some references to censuses will also be made.

2. ESTIMATION FORMULA

In the presence of unit and item nonresponse, the estimate of the total of characteristic y may be given by the general expression as in (2.1) below.

$$\tilde{Y} = \sum_{i=1}^{N} t_i \pi_i^{-1} \left\{ \delta_i \left[\delta_{iy} y_i + (1 - \delta_{iy}) z_{iy} \right] + (1 - \delta_i) z_i \right\}, \text{ where} \qquad (2.1)$$

t_i = 1 or 0 according as unit i is selected or not,
π_i = probability that unit i is selected.
δ_i = 1 or 0 according as unit i responds or not,
δ_{iy} = 1 or 0 according as responding unit i responds to item or characteristic y or not,
y_i = observed response for characteristic y when $\delta_{iy} = \delta_i = 1$; y_i may or may not $= Y_i$, the true value,
z_{iy} = imputed value for item nonresponse, when $\delta_i = 1$, $\delta_{iy} = 0$.
z_i = imputed value for unit nonresponse when $\delta_i = 0$.

The above estimate may pertain to a class a of units, when one inserts the indicators variable β_{ia} equal to 1 or 0 after π_i^{-1} to indicate whether or not unit i belongs to class a (e.g., age-sex class a).

In the case of item nonresponse, z_{iy} is nearly always an explicit *imputed value* for the missing information. The imputed value may be obtained by (i) a hot deck procedure i.e., substitution of an available response of characteristic y from the survey questionnaire of another unit that responded with respect to the characteristic and that is as similar as possible to unit i according to a decision table, (ii) substitution from other sources of data from the same unit such as an earlier survey, census, or administratrive data if such data are available, (iii) by regression methods or (iv) by logical deduction and the list is by no means exhaustive. In some cases, systematic errors may occur from, for example, faulty coders or keypunchers. In such cases one attempts to change the codes to logical values relative to other information on the questionnaire in place of imputation. In any case, one hopes to achieve an imputed value or altered code as close to the true value Y_i as possible. In the case of continuous surveys, with characteristics that are stable over a long period of time (such as employment in some industries and occupations), the response or earlier survey data may be considered almost as good as that of current survey data for the same unit. This would be especially when the reference periods of the current and earlier survey data are not too far apart in time. This may be also true in the case of survey data one year apart in the case of seasonal characteristics such as, for example, those related to the fishing industry. Sometimes the imputation of earlier survey data may be used also for unit nonrespondents that were respondents previously and with stable characteristics.

Usually, in the case of unit nonresponse, the imputation is undertaken by weight adjustment by the inverse response rate in a cell or area. The estimate of total is then given by:

$$\tilde{Y} = \sum_{i=1}^{N} t_i\pi_i^{-1}(wa)_i\delta_i\left[\delta_{iy}y_i + (1 - \delta_{iy})z_{iy}\right] \tag{2.2}$$

where $(wa)_i$ = weight adjustment for unit i to compensate for the deficient sample due to unit nonresponse. In the above expression, it is assumed that all item nonresponse has already been imputed for by z_{iy} in the case of responding unit i when $\delta_{iy} = 0$.

The estimates of the cumulative distribution function from the sample in the context of potential missing data may be obtained by replacing the observed value y_i by the indicator variable $c(y_i, Y) = 1$ or 0 according as $y_i \leq$ or $> Y$ and similarly for z_{iy} and z_i. The estimated c.d.f.'s corresponding to (2.1) and (2.2) are respectively given by (2.3) and (2.4) below.

$$\tilde{F}(Y) = \frac{1}{\hat{N}} \sum_{i=1}^{N} t_i\pi_i^{-1}\left\{\delta_i\left[\delta_{iy}c(y_i, Y) + (1 - \delta_{iy})c(z_{iy}, Y)\right] + (1 - \delta_i)c(z_i, Y)\right\} \tag{2.3}$$

where $\hat{N} = \sum_{i=1}^{N} t_i\pi_i^{-1}$ denotes the estimated or the true count of units in the universe. Thus, depending upon the frame, sample design, and listings of units, $\hat{N}$ may or may not $= N$.

$$\tilde{F}(Y) = \frac{1}{\hat{N}} \sum_{i=1}^{N} t_i\pi_i^{-1}(wa)_i\delta_i\left[\delta_{iy}c(y_i, Y) + (1 - \delta_{iy})c(z_{iy}, Y)\right] \tag{2.4}$$

While $\tilde{Y}$, as defined in (2.1) and (2.2), is identical according as to whether imputation for unit nonresponse is regarded as a substitution of mean values of respondents or as a weight adjustment, the c.d.f. estimates, $\tilde{F}(Y)$ as defined in (2.3) and (2.4), are not identical. When the mean of respondents, either overall or in adjustment cells defined for compensation of nonresponse, is substituted for each missing value as in (2.1) or (2.3), there results a spiking of such mean values in the estimated c.d.f., not reflecting the real shape of the c.d.f. in the population. The use of the weight adjustment $(wa)_i$, to inflate the sample weight π_i^{-1} in (2.4) avoids this spiking effect, yielding a different but more realistic estimate of the c.d.f.

Under full unit and item response, the estimates (2.1) and (2.2) simplify to the Horvitz-Thompson (1952) estimate of the total, which is unbiased apart from response errors. In the presence of missing data and imputation for them, the estimates (2.1) and (2.2) however are likely to be biased for reasons other than response errors unless z_{iy}'s and z_i's tend to equal y_i's when imputation for either item or unit nonresponse is required.

In the next section, the estimates (2.1) and (2.2) are decomposed into various components due to response error, imputation error due to item nonresponse, imputation error due to unit nonresponse and the effect of weight adjustments exceeding one.

3. Components of the Estimate

The estimate $\tilde{Y}$ given by (2.1) or (2.2) may be split up into 5 components, beginning with the Horvitz-Thompson estimate using the true values of the characteristic as in Table 1. The estimated c.d.f. $\tilde{F}(Y)$ as in (2.4) may be similarly split up but will be omitted in this paper.

When the weight adjustment $(wa)_i = 1$, the last line cancels out and the first 4 lines (3.1) to (3.4) total the estimate as given by (2.1). When the unit nonresponse is compensated for by a weight adjustment $(wa)_i > 1$, there is no direct substitution z_i for the missing value

Table 1:
Components of the Estimate $\tilde{Y}$

$\tilde{Y} = \sum_{i=1}^{N} t_i \pi_i^{-1} Y_i$	..	unbiased estimate based on full response, with true values	(3.1)
$+ \sum_{i=1}^{N} t_i \pi_i^{-1} (y_i - Y_i)$	..	effect of response error	(3.2)
$+ \sum_{i=1}^{N} t_i \pi_i^{-1} \delta_i (1 - \delta_{iy})(z_{iy} - y_i)$	..	effect of item nonresponse	(3.3)
$+ \sum_{i=1}^{N} t_i \pi_i^{-1} (1 - \delta_i)(z_i - y_i)$	..	effect of unit nonresponse	(3.4)
$+ \sum_{i=1}^{N} t_i \pi_i^{-1} [(wa)_i - 1] \delta_i [\delta_{iy} y_i + (1 - \delta_{iy}) z_{iy}]$	..	effect of weight adjustment for unit nonresponse	(3.5)

and z_i is taken to be 0 in (3.4). In that case, the 5 lines total the estimate as given by (2.2) and the negative effect of unit nonresponse in (3.4) is compensated for by the positive effect of weight adjustment in (3.5).

(a) Response error

The sum of the 1st and 2nd lines of the estimate $\tilde{Y}$ (See 3.1 and 3.2) equal the desired Horvitz-Thompson estimate of total under full response. The observed response y_i for unit i may not equal the true value Y_i so that a response error at unit i level may result. The response error, which is not the real subject of this paper, can only be reduced, though not likely eliminated, by proper interviewer training, good questionnaire design with unambiguous definitions of characteristics and questions and without cluster that would confuse the interviewer and/or respondent.

When the sampled weighted response errors of (3.2) do not cancel out, the estimate of the total $\tilde{Y}$ under full response, contains response error and upon taking expected value over all possible samples and response E_1 and E_3 (See Platek and Gray 1983), it may be found to be subject to response bias B_r and response variance in addition to sampling variance (SV). The response variance may be decomposed into simple (SRV) and correlated response variance (CRV) components.

The response bias, and all of the variance components (SV), (SRV) and (CRV) for the above estimate are derived in Platek and Gray (1983), subsection 2.2, pp. 257-8.

Response errors are usually studied by means of a reconciled reinterview program, whereby a subsample of responding units are reinterviewed and any observed differences between the original and reinterview data pertaining to the sample reference period are reconciled to determine which of the original or reinterview is the correct response. Reconciled reinterview surveys are undertaken in both the Canadian Labour Force Survey and the U.S. Current Population Surveys (CPS), two similar monthly surveys to measure unemployment,employment. etc.

For example, Poterba and Summers (1984), present in Table 2 some CPS results for a reconciled Reinterview Survey of May, 1976, based on a subsample of 3,329 men and 3,750 women. By means of reconciliation of a reinterviewed subsample, the *true* status of an individual is obtained so that it can be determined whether or not that individual responded correctly or not in the original survey, which in this case is CPS. Thus,the number of individuals with the true characteristics *Employed* in the reconciled interview sample who were actually reported as Employed, Unemployed, or Not in the LF in the original survey may be determined. From the three numbers, the proportion (or the probability) of correct and incorrect responses by true LF status may be estimated as in the table below.

Thus, for all of the men who were actually unemployed, 0.8720 is the estimated proportion of such men according to the reconciled reinterview study, who were accurately reported as unemployed while (0.0474 + 0.0806) or 0.1280 of the unemployed men were incorrectly reported as either *Employed* or *not in the Labour Force*. Thus, if y denotes characteristic *unemployed* i.e. $Y_i = 1$ when individual no. i is actually unemployed and a male then $y_i = 1$ correctly with probability 0.8720 while $y_i = 0$, incorrectly with probability 0.1280.

In the Canadian Labour Force Survey, the reconciled reinterview study sample during Jan.-Nov., 1984 covered 7,148 individuals and the corresponding probabilities of reporting labour force status as employed, unemployed or NILF in the regular LFS by *true* status as determined by the reinterview during 1984 are given in Table 3 below.

Thus the probability of correctly labelling an individual as unemployed, given that he/she actually unemployed is estimated to be .8691 in LFS compared with .8602 in CPS, almost

Table 2

Probabilities of Reporting Labour Force Status as Employed, Unemployed, or NILF in the Regular CPS, by *True* Status as Determined by the Reinterview Survey, May 1976.

	Status as Reported in the Regular CPS		
True Status	Employed	Unemployed	NILF
Total[1]			
Employed	0.9905	0.0016	0.0079
Unemployed	0.0356	0.8602	0.1041
NILF	0.0053	0.0025	0.9923
Men[2]			
Employed	0.9922	0.0013	0.0065
Unemployed	0.0474	0.8720	0.0806
NILF	0.0062	0.0048	0.9890
Women[3]			
Employed	0.9892	0.0019	0.0089
Unemployed	0.0194	0.8442	0.1363
NILF	0.0049	0.0015	0.9936

[1] Sampling size = 7,079

[2] Sampling size = 3,329

[3] Sampling size = 3,750

Source: Tables were computed from "General Labour Force Status in the CPS Reinterview by Labour Force Status in the Original interview.
Both Sexes. Total. After Reconciliation.
May 1976, Bureau of the Census (unpublished)

Table 3

Number of Individual and Probabilities of Reporting LF Status (in brackets) by *True* Characteristic. Jan.-Nov. 1984

True LF Characteristic (Reconciled reinterview)	Regular LFS			
	Employed	Unemployed	NILF	Total
Employed	4,082 (0.9831)	19 (0.0046)	51 (0.0123)	4,152
Unemployed	8 (0.0122)	571 (0.8691)	78 (0.1187)	657
NILF	28 (0.0120)	30 (0.0128)	2,281 (0.9752)	2,339
Total	4,118	620	2,410	7,148

the same. The corresponding probabilities for *Employed* and *Not in the Labour Force* in LFS are estimated during 1984 to be .9831 and .9752 compared with .9905 and .9923 for CPS, both somewhat lower in LFS. The reason for the difference cannot be determined at this stage. In any case, the response errors are likely more serious at national than at small area levels. For example, at national levels the response biases may be larger in magnitude relative to their sampling errors while a small area level estimate may be subject to response biases of about the same percent as at national level, but which may be much smaller than the sampling errors.

(b) Item Nonresponse and Imputation Error

The third line (3.3) of the estimate $\tilde{Y}$ in Table 1 showed the deviation from the desired estimate $\hat{Y}$ as a result of imputation for item nonresponse when the imputed value $z_{iy} \neq y_i$ and when the sampled weighted differences $(z_{iy} - y_i)$ over the sampled units with imputations for item nonresponse do not cancel out. Item nonresponse results from a respondent refusing to answer certain questions on the questionnaire may have been inadvertently left incompleted by either the respondent (in the case of self-enumeration) or by the interviewer. The second of the two causes of item nonresponse may result from similar causes as for response errors; i.e. complex questions with ambiguous definitions and/or an involved or cluttered questionnaire with a tendency for potential errors in following the proper path, depending upon replies to filter questions.

When item nonresponse does occur, an imputation strategy as described earlier may be undertaken, which almost always results in an explicit substitution. Crucial to data analysis at micro-levels is the need to obtain a value z_{iy} as close to the true value Y_i or at least as close to what would be the observed y_i, if the unit had responded to the question(s) that determine(s) characteristic y. There is unfortunately no way of knowing how close z_{iy} agrees with y_i except through re-enumeration of the unit, or a review and study of external sources or earlier survey data (which may not be available). The further danger of item nonresponse and the imputation for it may be the false sense of security to the data user who may not be aware or who may not be informed of the substituted value z_{iy} in place of a bonafide response at the micro-data level. The imputed value z_{iy} will tend to deviate in either direction from the true value Y_i to a greater extent than the potential response error y_i if that unit responds to the characteristic. This may not always be the case. Unfortunately, it usually cannot be determined at the micro-level whether or not z_{iy} is less accurate than y_i would be. Even if the imputation error may sometimes be lower than the potential response error, it may further deteriorate the quality of the published statistics because of the presence of additional variance components.

Item nonresponse and response errors are often detected in the LFS by a monthly project Field Edit Module which analyzes questionnaires that failed edit for one or more questions. The distinction between response errors and item nonresponse however is often quite blurred in the analysis without probing into the individual questionnaires in detail. The common type of discrepancy is a miscoding of a question rather than item nonresponse per se. Many questions are split up into 5 or 6 different sub-categories and a miscoding may be interpreted as an item nonresponse for one sub-category and a response error for another sub-category pertaining to the same question. The analysis of the Field Edit Module deals with items (questions) but not sub-categories of the questions. The item discrepancy rate is thus difficult to define unambiguously. It pertains to a subset of questionnaires for which a specific question, say, No. q is relevant according to filter questions and decision tables. Let us suppose that out of a responding sample size of m questionnaires, question No. q is relevant for $m_q \leq m$ questionnaires. Then the discrepancy rate is the proportion of m_q questionnaires that failed edit, whether by item nonresponse or faulty coding. The ambiguity in the definition lies in whether the subset m_q should include those questionnaires with the question completed in error, those with the question left blank in error or merely those questionnaires with the question coded correctly or incorrectly. Notwithstanding the possible ambiguity in the definition, the item discrepancy rates for about 50 items as analysed for calendar year 1984 should indicate an upper bound to the fractional error in the estimates of statistics based on the items. A sample of item (defined in Table 4a) discrepancy rates for 1984 is given in Table 4 below.

Thus, for a straightforward item like (10) "Did the respondent do any work last week? Yes or No," the discrepancy rate is only 0.2%, much lower than even the national standard error. For more complex items likes Nos. 12, 36, 41, 54 and 77 the discrepancy rate averages more than 10% with ranges 2 to 6% in either direction from the mean over the year. The discrepancies are corrected for, by hot deck procedures, use of last survey's responses (if available) or by logical deduction from other questionnaire data. Thus, in many instances an item discrepancy may be altered to a response subject to response rather than imputation error so that the discrepancy rates should be construed as an upper bound to the overall imputation error rates for the items.

(c) Unit Nonresponse and Weight Adjustment

In the case of unit nonresponse the two components of $\tilde{Y}$ given by (3.4) and (3.5) must be studied together since unit nonresponse is generally compensated for by a weight adjustment $(wa)_i$ rather than direct substitution z_i for a missing unit value. Weight adjustments are usually calculated by inverse rates in adjustment cells of which there are two basic types, balancing areas and weighting classes. Balancing areas are frequently design-dependent geographic areas such as a stratum, primary sampling unit, cluster, or a groups of strata or even the entire sample. Weighting classes are defined by post-strata (strata defined after sampling) formed on the basis of information available to both respondents and nonrespondents in the sample. The nonrespondent's information may be obtained from partial nonrespondents with some known characteristics even though the particular characteristic being estimated is not known for the partial nonrespondents. Alternatively, the information may be derived from external sources pertaining to the nonrespondents. Inverse response rates may be calculated for either balancing areas or weighting classes and used as weight adjustments to compensate for missing data in the cells.

Table 4
Average Discrepancy Rate by Item (defined in Table 4a)

Item	Average Discrepancy Rate	Range of Rates in 1984 (Min. to Max.)
10	0.2%	0.2% Every month
12	12.3%	10.4% to 14.3%
14	6.7%	5.7% to 8.4%
16	0.4%	0.3% to 0.5%
17	6.6%	2.0% to 9.9%
30	0.4%	0.3% to 0.5%
32	7.0%	3.0% to 11.6%
33	4.3%	1.8% to 6.0%
36	10.6%	8.1% to 12.7%
40	4.1%	1.5% to 6.8%
41	12.1%	6.2% to 19.7%
54	10.1%	7.9% to 12.1%
76	<0.1%	0.0% to 0.1%
77	15.0%	11.8% to 17.3%

Source: Internal report by Karen Switzer to P.D. Ghangurde March 4, 1985 "Some Findings on the Field Edit Module (FEM) Reports from 1984".

Table 4a
Definition of Items

(10) Last week did (respondent) do any work at a job or business? Yes or No.

(12) If yes to 11, "Did... have more than one job last week, was this a result of changing employers?" Yes or No.

(14) What is the reason... usually works less than 30 hours per week, if actual response to (13) no. of hrs. worked 30.

(16) Last week, how many hours was ... away from work for any reason whatsoever (holidays, vacations, illness, labour dispute, etc.) "00" should be filled in

(17) What was the main reason for being away from work? (10 possible codes)

(30) Last week did ... have a job or business at which he/she did not work? Yes or No.

(32) Counting from the end of last week, in how many weeks will ... start to work at his/her new job? (Reply to Yes in (31), "Last week did ... have a job to start at a definite date in the future?")

(33) Why was ... absent from work last week? (8 possible codes)

(36) Identical to (14) but pertaining to *Unemployed* instead of *Employed* individuals.

(40) Inthe past 4 weeks has ... looked for another job? Yes or No.

(41) What has ... done in the past 4 weeks to find another job? (8 possible codes, 1 to 3 different codes in 1, 2, or 3 spaces).

(54) What was the main reason why ... left that job? (9 possible codes) in response to yes to (50) has ... ever worked at a job or business (pert. to individuals permanently unable to work) and questions (51) to (53) dealing with date of last job and part/full time status. (54) is slipped if date of last job not too recent according to a pre-printed date in (52).

(76/77) Class of worker and whether or not same as previous month, with respect to main job (76) and other job (77)

There are several types of weight adjustments available for inflation of the sample to compensate for unit nonresponse, the most common being the inverse response rate defined by the ratio of the sample size to the responding sample size in an adjustment cell. Thus, if the cell contains N_b units in its population and is represented by n_b selected units, where:

$n_b = \sum_{i\epsilon b} t_i$ the sample size in cell b which may or may not be a constant; depending on the definition of the cell,

$\hat{N}_b = \sum_{i\epsilon b} \pi_i t_i$, an estimate of the size of cell b in the population, usually N_b would not be known except in a census.

$m_b = \sum_{i\epsilon b} t_i \delta_i$ = no. of responding units in cell b, i.e., the responding sample size,

then, $(wa)_i = n_b/m_b$ when i lies in adjustment cell b. (3.6)

Before defining other possible weight adjustments, we will concentrate on the frequently applied inverse unweighted response rate in a cell as in (3.6). The estimate of the total defined by (2.2) with $(wa)_i = n_b/m_b$ may be rewritten as a special case of (2.1), with z_i given by:

$$z_i = \hat{T}_b / \pi_i^{-1} m_b, \tag{3.7}$$

where $\hat{T}_b = \sum_{i\epsilon b} \pi_i^{-1} t_i \delta_i [\delta_{iy} y_i + (1 - \delta_{iy}) z_{iy}]$, sample weighted total of responding units in cell b. In the case of equal sample weights in a cell, the imputed value z_i simplifies to the mean value of m_b respondents in the cell. By substituting z_i given by (3.7), into (2.1), it may be shown that the estimate is identical to (2.2) with $(wa)_i = (n_b/m_b)$. Thus, one may regard imputation for unit nonresponse as a substitution of $z_i = \hat{T}_b/(\pi_i^{-1} m_b)$ in (2.1) or as a weight adjustment to the sample weights by $(wa)_i = n_b/m_b$ in (2.2). In the case of the weight adjustment, one would set $z_i = 0$ in (3.4) in $\tilde{Y}$ as split up into 5 components. Alternatively, one may employ the imputed value z_i as defined in (2.1) and in that case, one would set $(wa)_i = 1$ in (3.5) resulting in that component of $\tilde{Y} = 0$. Thus in order to consider the effect of weight adjustment $(wa)_i > 1$, both the negative component (3.4) and positive component (3.5) must be studied together; but to consider the effect of the implicit imputed value z_i , given by (3.7), one needs only to consider (3.4).

The weight adjustment (n_b/m_b) is used in LFS, where the adjustment cells are design-dependent psu's in non-self representing areas (NSR) and strata (subunits) of contiguous city blocks in self-representing areas (SR). In Table 5, the number of cells, the unweighted average of the weight adjustments and the frequency distribution of the weight adjustment in intervals 1-1.01, 1.01-1.02. ..., 1.10 and over are given by region/type of area for the survey, Jan. 1983.

The average weight adjustment of 1.0348 at Canada level is less than what one would expect with a nonresponse rate of about 5%. The reason for the apparent low average weight adjustment is that, for purposes of calculations of the inverse response rate, some unit nonrespondents with available responding data of the previous month for imputation purposes are treated like respondents. This applies to about 20 to 30% of the nonrespondents every month.

Table 5

Number of Adjustment Cells, Average and Frequency Distribution of the Weight Adjustments by Region/Type of Area. January, 1983

				No. of cells in intervals of $(wa)_i$										
Region	Type of Area	No. Cells	Aver. $(wa)_i$	1- 1.01	1.01- 1.02	1.02- 1.03	1.03- 1.04	1.04- 1.05	1.05- 1.06	1.06- 1.07	1.07- 1.08	1.08- 1.09	1.09- 1.10	1.10+
Atl.	NSR	254	1.0250	143	6	22	21	13	13	9	7	8	2	10
Atl.	SR	123	1.0246	58	5	11	15	14	4	3	6	4	1	2
Que.	NSR	126	1.0550	72	2	8	10	10	6	8	6	0	1	3
Que.	SR	185	1.0265	106	0	7	8	23	11	4	5	7	3	11
Ont.	NSR	120	1.0333	58	1	10	11	11	8	4	2	2	2	11
Ont.	SR	252	1.0416	116	1	13	24	21	16	9	9	8	10	25
Pr.	NSR	328	1.0348	167	5	17	22	23	24	15	12	10	8	25
Pr.	SR	149	1.0306	40	23	23	20	13	8	7	3	5	4	3
BC	NSR	85	1.0468	38	3	7	8	8	2	5	1	1	1	11
BC.	SR	119	1.0412	46	4	7	15	10	7	7	7	3	3	10
Can.	NSR	913	1.0358	478	17	64	72	65	53	41	28	21	14	60
Can.	SR	828	1.0337	366	33	61	82	81	46	30	30	27	21	51
Canada		1,741	1.0348	844	50	125	154	146	99	71	58	48	35	111

Without a knowledge of the nonrespondents' characteristics, it cannot be determined precisely the threshold level beyond which the weight adjustment would become critical to result in an unacceptable bias along with an increase in the variance due to a smaller effective sample size. If the threshold is arbitrarily set for LFS at 1.05 (a level sometimes assumed by survey practitioners) then about 1/4 of the balancing units (441 out of 1,741) across Canada had critical weight adjustments of 1.05 or more in Jan. 1983. In many other surveys such as those dealing with income and expenditure, the nonresponse rate is higher overall and would likely be critical in nearly all cells if the same threshold of 1.05 is assumed.

There are other types of weight adjustments in cells. For example, one could exclude from cell b as defined above, those units that contain item nonresponse for at least one question. Let us suppose there are m_{bQ} units in cell b free of item nonresponse for the whole set of questions on the questionnaire. For $(m_b - m_{bQ})$ responding units in the cell with some item nonresponse the weight $(wa)_i = 1$, and for the remaining m_{bQ} responding units, free of item nonresponse, the weight adjustment is given by:

$$(wa)_i = [n_b - (m_b - m_{bQ})] / m_{bQ}, \text{ which exceeds } n_b/m_b. \qquad (3.7a)$$

The following is the justification for applying no weight adjustment i.e., $(wa)_i = 1$, for those units in the cell with some item nonresponse but a larger weight adjustment (3.7a) than (n_b/m_b), for those units free of item nonresponse Records with item nonresponse likely contain response and imputation errors while records free of item nonresponse contain only response errors and with the large weight applied to records free of item nonresponse, it may be possible to obtain estimates with lower mean square error than by using the same weight adjustment for all m_b responding units in the cell. To our knowledge, weight adjustments such as described above have not been applied but they may be worthy of study if the decrease in the bias offsets the increase in the variance that would occur with the different weights.

In the case of units with unequal probability sampling, there exists a weight adjustment based on the weighted sample and responding units in a cell instead of the unweighted ones. In such as case,

$$(wa)_i = \hat{N}_b / \hat{M}_b, \tag{3.8}$$

where $\hat{M}_b = \sum_{i \epsilon b} \pi_i^{-1} t_i \delta_i$ is the sample weighted count of responding units in cell b. For the analogous case to the weight adjustment $(wa)_i$ in (3.7a) applied only to responding units free of item nonresponse,

$$(wa)_i = [\hat{N}_b - (\hat{M}_b - \hat{M}_{bQ})] / \hat{M}_{bQ} \tag{3.9}$$

where $\hat{M}_{bQ} = \sum_{i \epsilon b} \pi_i^{-1} t_i \delta_i \Pi_{q=1}^{Q} \delta_{iq}$, the weighted count of responding units in cell b, free of item nonresponse.

$\delta_{iq} = 1$ or 0 according as unit i responded or did not respond to question no. q of the survey questionnaire containing Q questions; thus, $\Pi_{q=1}^{Q} \delta_{iq} = 1$ only if responding unit i is free of item nonresponse.

The justification for using (3.9) in lieu of (3.8) may be similar to that for using (3.7a) instead of (3.6). The justification for using weighted in place of unweighted response rates needs explanation and is provided after Table (6).

One could derive separate $(wa)_i$ expressions as of (3.7a) or (3.9) for each question q or for each characteristic y, defined by a set of one or more questions. Unfortunately, one would be faced with different weight adjustments in an adjustment cell for different questions or characteristics resulting in inconsistencies among different characteristics in published tables. In order to ensure uniform survey weights and weight adjustments, $(wa)_i$ should depend only on the unit and not on the question or characteristic though one may permit imputations for some items while excluding them for other items such as major ones in the weight adjustments (3.7a) or (3.9) as long as the inclusions and exclusions are consistent in the adjustment cell. For example, one may consider an imputation for missing item by logical deduction rather than by hot decking as pertaining to a record free of item nonresponse for weight adjustment purposes.

For each of the above weight adjustments as in (3.6) to (3.9), it can be shown that (2.2) is a particular case of (2.1) with z_i given by a weighted or unweighted mean of respondents. Thus, the implicit imputed value z_i for nonresponding unit i for each of the four cases of weight adjustments cited above is given by the expressions in Table (6). Additional notation is required for the expressions as given below:

$$\hat{T}_b = \sum_{i=1}^{N_b} t_i \pi_i^{-1} \delta_i [\delta_{iy} y_i + (1 - \delta_{iy}) z_{iy}] = \text{sample weighted total of unit respondents including imputations for item nonresponse but excluding weight adjustments by inverse unit response rate.} \tag{3.10}$$

$\hat{T}_{by} = \sum_{i=1}^{N_b} t_i\pi_i^{-1}\delta_i\delta_{iy}y_i$ = sample weighted total of unit and item respondents with respect to characteristic y, (3.11)

$\hat{T}_{bQy} = \sum_{i=1}^{N_b} t_i\pi_i^{-1}\delta_i \prod_{q=1}^{Q} \delta_{iq}y_i$ = sample weighted total of unit and item respondents with respect to characteristic y, but excluding those records in the cell with imputation for any item nonresponse (3.12)

Thus, $\hat{T}_{bQy} \leq \hat{T}_{by} \leq \hat{T}_b$.

The weight adjustment $(n_b - m_b + m_{bQ})/m_{bQ} = 1 + (n_b - m_{bQ})/m_{bQ}$ of (c) $\geq$ the weight adjustment of (n_b/m_b) of (a) since $m_{bQ} \leq m_b$(see Table 6). Hence, for a given response rate m_b/n_b in a cell, one may anticipate a larger variance of an estimate using (c) than one using (a). The larger variance may or may not counteract a potentially smaller imputation bias in the overall mean square error. The same holds true in the case of applying weighted response rates $(\hat{N}_b - \hat{M}_b + \hat{M}_{bQ})/\hat{M}_{bQ}$ in (d) as opposed to $\hat{N}_b/\hat{M}_b$ in (b) since $\hat{M}_{bQ} \leq \hat{M}_b$. When pps sampling is applied, the use of weighted vs. unweighted response rates leads to another interesting result. It is shown in Platek and Gray (1983), p. 264-265 that, when the response and selection probabilities, i.e., α_i and π_i, are positvely correlated, the weight adjustments with weighted response rates will tend to be higher than those with unweighted rates. Thus under the condition of positive c orrelation between α_i and π_i, $E(\hat{N}_b/\hat{M}_b) > E(n_b/m_b)$ and similarly, $E[(\hat{N}_b - \hat{M}_b + \hat{M}_{bQ})/\hat{M}_{bQ}] > E[(n_b - m_b + m_{bQ})/m_{bQ}]$, where $E = E_1E_2$, the expected value overall possible samples of units and subsamples of responding units as described by Platek and Gray (1983), p. 251.

Table 6
Implicit Imputed Value for Unit Nonrespondent by
Weight Adjustment (Cell Level)

	Weight Adjustment	Reference in text	Implicit Imputed value when $i=0$	Description
(a)	n_b/m_b	(3.6)	$\hat{T}_b/(\pi_i^{-1}m_b)$	Unweighted unit response rate
(b)	$\hat{N}_b/\hat{M}_b$	(3.8)	$\hat{T}_b/\hat{M}_b$	Weighted unit response rates
(c)	$\dfrac{n_b - m_b + m_{bQ}}{m_{bQ}}$	(3.7a)	$\hat{T}_{bQy}/\pi_i^{-1}m_{bQ}$	Unweighted unit response rates among units free of item nonresponse
(d)	$\dfrac{\hat{N}_b - \hat{M}_b + \hat{M}_{bQ}}{\hat{M}_{bQ}}$	(3.9)	$\hat{T}_{bQy}/\hat{M}_{bQ}$	Weighted unit response rates among units free of items nonresponse

Note: In the case of self-weighting sample (srswor as a particular case), the implicit imputed value z_i becomes the simple mean of respondents for both cases (a) and (b), and the simple mean of respondents (excluding those with some item nonresponse) in the cases of (c) and (d).

* See appendix I for derivation.

Whatever the weight adjustment used to compensate for unit nonresponse, it is doubtful that the individual values z_i *implicit imputed* would be close to the individual true values Y_i or even to the potential observed responses y_i. The best that can be achieved with the weight adjustment is to hope that adjustment cells formed to compensate for missing data due to unit nonresponse will ensure minimum differences between the characteristics of respondents and nonrespondents in the cells. Thus, the formation and delineation of adjustment cell is most crucial for compensation regardless of the type of weight adjustment that is applied.

7. FINAL REMARKS

As seen in the sections above, there is no ready-made solution to the missing data, whatever the types that occur. The initial strategy is to minimize the occurance of missing data to the extent possible, without incurring great cost or sacrificing the timeliness of the survey data. Every attempt should be made at the onset to prepare for some nonresponse and set up imputation strategies. If missing data occur in about the manner anticipated, then the survey data processing ought to proceed on schedule, with the appropriate substitutions or weight adjustments. Clearly, the scheduling of survey data collection, publishing, etc. can proceed in a more orderly fashion in continuous or repeated surveys than in ad hoc one-time surveys for which the survey designer may not realize, until after the fact, all the things that can go wrong such as unexpected refusals or lack of interest on the part of both interviewers and respondents.

In order to deal with the nonresponse problems it is essential to maintain a continuous study of nonresponse rates by the survey characteristic (in the case of item nonresponse), reason for nonresponse, and if possible, to extend the study to an analysis of item and unit response probabilities so that imputation biases may be estimated from the survey itself. Alternatively, model-based estimates may continue to be explored to examine the imputation bias and, furthermore, to strengthen the estimates by employing additional information.

APPENDIX

Derivation of Implicit Value z_i for Unit Nonresponse imputation

In the case of (c) and (d) of Table 6, the estimate of cell b level is given by:

$$\tilde{Y}_b = \hat{T}_{bQy}\,(wa)_i + (\hat{T}_b - \hat{T}_{bQy}) \tag{A.1}$$

$$= \hat{T}_b + [(wa)_i - 1]\,\hat{T}_{bQy}$$

In case (c), $(wa)_i - 1 = (n_b - m_b)/m_{bQ}$

$$= \sum_i t_i(1 - \delta_i)/m_{bQ}$$

or $\tilde{Y}_b = \hat{T}_b + \sum_i t_i\pi_i^{-1}(1 - \delta_i)\hat{T}_{bQy}/\pi_i^{-1}m_{bQ}$ (A.2) or by equating (A.2) to (A.1), noting the definitions of $\hat{T}_b$ in (3.10) and $\bar{Y}$ in (2.1), one may see that the imputed value z_i is given by $\hat{T}_{bQy}/\pi_i^{-1}m_{bQ}$ as stated in (c) of Table (6).

Similarly, when weighted response rates are employed, the implicit imputed value z_i may be found to be $\hat{T}_{bQy}/\hat{M}_{bQ}$ as in (d) of Table (6). The results for (a) and (b) of Table (6) follow by setting $m_{bQ} = m_b$ and $\hat{M}_{bQ} = \hat{M}_b$.

REFERENCES

HORVITZ, D.G. and THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

LESSLER, J.T. (1979). An expanded survey error model. In *Incomplete Data in Sample Surveys*, Volume 3 - Proceedings of the Symposium (eds. W.G. Madow, I. Olkin, and B.D. Rubin), San Diego: Academic Press, 259-270.

PLATEK, R. (1977). Some factors affecting nonresponse. *Survey Methodology*, 3, 191-214.

PLATEK, R. (1980). Causes of incomplete data, adjustments and effects. *Survey Methodology*, 6, 93-132.

PLATEK, R., and GRAY, G.B. (1978). Nonresponse and imputation. *Survey Methodology*, 4, 144-177.

PLATEK, R., and GRAY, G.B. (1979). Methodology and application of adjustments for nonresponse. Presented at the 42nd Session of International Statistical Institute, Manila, Philippines.

PLATEK R., and GRAY, G.B. (1983). Part V - Imputation Methodology: Total Survey Error. In *Incomplete Data in Sample Surveys*, Volume 2 - Theory and Bibliographies (eds. W.G. Madow, I. Olkin, and D.B. Rubin), San Diego: Academic Press, 249-333.

POTERBA, J.M., and SUMMERS, L.H. (1984). Response variation in the CPS: Caveats for the unemployment analyst. *Monthly Labour Review*, March 1984. Research Summaries, 37-43.

Survey Methodology, June 1985
Vol. 11, No. 1, pp. 15–31
Statistics Canada

Conditional Inference in Survey Sampling

J.N.K. RAO[1]

ABSTRACT

Conventional methods of inference in survey sampling are critically examined. The need for conditioning the inference on recognizable subsets of the population is emphasized. A number of real examples involving random sample sizes are presented to illustrate inferences conditional on the realized sample configuration and associated difficulties. The examples include the following: estimation of (a) population mean under simple random sampling; (b) population mean in the presence of outliers; (c) domain total and domain mean; (d) population mean with two-way stratification; (e) population mean in the presence of non-responses; (f) population mean under general designs. The conditional bias and the conditional variance of estimators of a population mean (or a domain mean or total), and the associated confidence intervals, are examined.

KEY WORDS: Conditional inference; Conditional bias; Conditional variance; Population mean; Random sample sizes

1. INTRODUCTION

In the conventional set-up for inference in survey sampling the sample design defines the sample space S (set of possible samples s) and the associated probabilities of selection, $p(s)$. The choice of an estimator is based on the criterion of consistency or unbiasedness and on the comparison of mean square errors (MSE), under repeated sampling with probabilities $p(s)$, using the sample space S as the reference set. Thus, an estimator $\hat{\bar{Y}}$ of a population mean $\bar{Y}$ is unbiased if $E(\hat{\bar{Y}}) = \sum_{s\epsilon S} p(s)\hat{\bar{Y}}_s = \bar{Y}$, where $\hat{\bar{Y}}_s$ is the value of $\hat{\bar{Y}}$ for the sample s. The MSE of the estimator $\hat{\bar{Y}}$ is given by $\text{MSE}(\hat{\bar{Y}}) = \sum_{s\epsilon S} p(s)(\hat{\bar{Y}}_s - \bar{Y})^2$, and $\hat{\bar{Y}}$ is consistent if its MSE approaches zero as the sample size increases. A consistent or unbiased estimator of $\text{MSE}(\hat{\bar{Y}})$, denoted as $\text{mse}(\hat{\bar{Y}})$, provides a measure of uncertainty in $\hat{\bar{Y}}$. If $\hat{\bar{Y}}$ is unbiased or consistent, then the observed values $\hat{\bar{Y}}_s$ and $\text{mse}(\hat{\bar{Y}}_s)$ provide a large sample, $(1 - \alpha)$-level, confidence interval given by

$$I_s = \hat{\bar{Y}}_s \pm z_{\alpha/2}\sqrt{\text{mse}(\hat{\bar{Y}}_s)}, \tag{1}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$-point of a $N(0, 1)$ variable. The interpretation of (1) is that in repeated sampling with S as the reference set, approximately $100\,(1 - \alpha)\%$ of the intervals, I_s, will contain the true value $\bar{Y}$.

The comparison of unconditional mean square errors, $\text{MSE}(\hat{\bar{Y}})$, is appropriate at the design stage, but the sample space S may not be the relevant reference set for inference after the sample s has been drawn, if the sample contains "recognizable subsets". The concept of recognizable subsets will be illustrated in subsequent sections through examples involving random sample sizes. The choice of relevant reference set, however, is not unique. In fact, the surveyed sample s can be viewed as unique in a real sense, but then no inference under a repeated sampling set-up can be made since the relevant reference set would contain a singleton (Holt and Smith 1979).

[1] J.N.K. Rao, Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada K1S 5B6.

Conditional inference has attracted considerable attention and controversy in classical statistics since Fisher (1925). For instance, in testing for independence in a 2 × 2 table of counts, Fisher argued that the inference should be conditional on the observed row and column marginal totals even if the margins are not fixed by the design. Yates (1984) revived this problem. The choice of relevant reference set is not always clear-cut, but the following guidelines look reasonable: (1) A conditional procedure should be chosen *before* observing the data, especially in the public domain. (2) A conditioning partition of S should be chosen in such a way that the partition contains no (or little) information on the parameters of interest, i.e. the statistic indexing the partition should be an ancillary statistic (Cox and Hinkley 1974, p. 38). (3) If the sample sizes are random (e.g., domain sample sizes) and their population distribution is completely known (or at least partially known), then the inferences should be conditional on the observed sample sizes. In this context, Durbin (1969, p. 643) says "If the sample size is determined by a random mechanism and one happens to get a large sample one knows perfectly well that the quantities of interest are measured more accurately than they would have been if the sample size had happened to be small. It seems self-evident that one should use the information available on sample size in the interpretation of the result. To average over variations in sample size which might have occurred but did not occur, when in fact the sample size is exactly known, seems quite wrong from the standpoint of the analysis of the data actually observed".

The discussion throughout the paper will be confined to conditional inference in the presence of random sample sizes, as in guideline (3) above. Even with this restriction, it will be shown that conditional inferences are not always easy to implement in practice. We begin our discussion with simple examples and then extend it to more complex problems. In the context of sample surveys, Holt and Smith (1979) provide the most compelling arguments in favour of conditional inference, although their discussion was restricted to poststratification of a simple random sample (SRS); see Section 3.1.

Lahiri (1969) pointed out the "difficulties of conveying convincingly the real import of the sample survey estimates to intelligent but lay users of statistical data"; in particular, "the fallacy in implicitly using the (sampling) standard error as a measure of precision of the *observed* (sample) estimate, illustrating this point with a number of examples drawn from the current theory".

2. SIMPLE RANDOM SAMPLING WITH REPLACEMENT

Simple random sampling (SRS) with replacement is seldom used in practice, but it provides a simple introduction to conditional inference.

Suppose a simple random sample, s, of size n is selected fom a population of size N with replacement so that S contains N^n samples s. Let ν denote the number of distinct units in s. Then ν is a random variable with possible values $1, \ldots, n$. Let t_i denote the number of times the i-th population unit is included in s. Then two well-known estimators of the population mean $\bar{Y}$ are given by

$$\bar{y}_n = \frac{1}{n} \sum_{i \in s} t_i y_i, \tag{2.1}$$

the sample mean based on all the n draws, and

$$\bar{y}_\nu = \frac{1}{\nu} \sum_{i \in s} y_i, \tag{2.2}$$

the mean based on the distinct units in s. Both $\bar{y}_n$ and $\bar{y}_\nu$ are unconditionally unbiased under the reference set S, and the unconditional variance of $\bar{y}_\nu$ is always smaller than that of $\bar{y}_n$. Hence, from efficiency considerations $\bar{y}_\nu$ should be preferred over $\bar{y}_n$. The Horvitz-Thompson estimator

$$\bar{y}_{HT} = \frac{1}{N} \sum_{i \epsilon s} \frac{y_i}{\pi_i} = \frac{\nu}{E(\nu)} \bar{y}_\nu \tag{2.3}$$

is also unconditionally unbiased, where π_i is the probability that unit i is included at least once in the sample:

$$\pi_i = \frac{E(\nu)}{N} = 1 - \left(1 - \frac{1}{N}\right)^n.$$

The comparison of variances of $\bar{y}_\nu$ and $\bar{y}_{HT}$ shows that $\bar{y}_\nu$ is not always better than $\bar{y}_{HT}$.

Following Durbin's (1969) argument, it is clear that for the purpose of inference one should condition on the observed value of ν, i.e., the relevant reference set is the set S_ν of $\binom{N}{\nu}$ samples of effective size ν, and not S. Fortunately, it is easy to implement conditional inference in this case since $P(s_\nu | \nu) = \binom{N}{\nu}^{-1}$, i.e. conditionally, the observed sample, s_ν, of distinct units is a simple random sample of size ν drawn without replacement. It follows that $\bar{y}_\nu$ is conditionally unbiased, i.e. $E_2(\bar{y}_\nu) = \bar{Y}$ where E_2 denotes conditional expectation, whereas $E_2(\bar{y}_{HT}) = [\nu/E(\nu)]\bar{Y} \neq \bar{Y}$ so that $\bar{y}_{HT}$ is conditionally biased. Hence, $\bar{y}_\nu$ should be preferred over $\bar{y}_{HT}$, despite the inconclusive comparison of unconditional variances. Note that $\bar{y}_{HT}$ would be a serious underestimate if the observed ν is much smaller that $E(\nu)$.

A relevant measure of uncertainty is the conditional variance, $V_2(\bar{y}_\nu)$, which is estimated unbiasedly by

$$v(\bar{y}_\nu) = \left(\frac{1}{\nu} - \frac{1}{N}\right) s^2_{\nu y}, \tag{2.4}$$

where $(\nu - 1)s^2_{\nu y} = \sum_{i \epsilon s} (y_i - \bar{y}_\nu)^2$ and V_2 denotes the conditional variance. The appropriate confidence interval for $\bar{Y}$ is given by

$$I_\nu = \bar{y}_\nu \pm z_{\alpha/2} \sqrt{v(\bar{y}_\nu)}. \tag{2.5}$$

Conditionally, the confidence level of I_ν is $1 - \alpha$ approximately if ν is not small. Another variance estimator

$$v^*(\bar{y}_\nu) = \left[E\left(\frac{1}{\nu}\right) - \frac{1}{N}\right] s^2_{\nu y} \tag{2.6}$$

is conditionally biased, although unbiased when averaged over the whole sample space, S. It follows from (2.4) and (2.6) that $v(\bar{y}_\nu) < v^*(\bar{y}_\nu)$ if $1/\nu < E(1/\nu)$ and vice versa if $1/\nu > E(1/\nu)$. Thus, the confidence interval based on (2.6) would be too narrow if $E(1/\nu) < 1/\nu$ and hence yield a confidence level less than $1 - \alpha$, and too wide if $E(1/\nu) > 1/\nu$ leading to a confidence level greater than $1 - \alpha$. It may be noted that confidence intervals that are conditionally correct are automatically correct in the unconditional framework.

3. SIMPLE RANDOM SAMPLING WITHOUT REPLACEMENT

Suppose a simple random sample of fixed size n is drawn without replacement. In the absence of recognizable subsets, the relevant reference set is the set S of $\binom{N}{n}$ samples s, each of size n, and the sample mean $\bar{y}_n$ is unbiased and its variance is estimated unbiasedly by

$$v(\bar{y}_n) = \left(\frac{1}{n} - \frac{1}{N}\right)s_{ny}^2 \tag{3.1}$$

where $(n - 1)s_{ny}^2 = \sum_{i\epsilon s}(y_i - \bar{y}_n)^2$. The resulting confidence interval is given by I_s: $\bar{y}_n \pm z_{\alpha/2}\sqrt{v(\bar{y}_n)}$ with confidence level $1 - \alpha$ approximately if n is not small.

Suppose now that recognizable subsets exist in the sense that we observe the sample configuration $\underset{\sim}{n} = (n_1, \ldots, n_k)$ belonging to k post-strata with known weights $W_i = N_i/N$. Ideally, stratified sampling should have been used but the strata frames were not available. The relevant reference set now is the set $S_{\underset{\sim}{n}}$ of $\prod\binom{N_i}{n_i}$ samples having the realized configuration $\underset{\sim}{n}$ since the distribution of $\underset{\sim}{n}$ is completely known.

3.1 All $n_i \geq 1$

If *all* the observed $n_i \geq 1$, then the customary post-stratified estimator

$$\bar{y}_{pst} = \sum W_i\bar{y}_i \tag{3.2}$$

is conditionally unbiased given $\underset{\sim}{n}$ since $P(s|\underset{\sim}{n}) = \prod\binom{N_i}{n_i}^{-1}$, i.e., conditionally the observed sample s is a stratified random sample $(s_1, \ldots, s_k)$ with strata sample sizes n_i. Here $\bar{y}_i$ denotes the sample mean in the i-th stratum. A relevant measure of uncertainty is the conditional variance, $V_2(\bar{y}_{pst})$, which is estimated unbiasedly by

$$v(\bar{y}_{pst}) = \sum W_i^2\left(\frac{1}{n_i} - \frac{1}{N_i}\right)s_{iy}^2, \tag{3.3}$$

provided *all* $n_i \geq 2$, where $(n_i - 1)s_{iy}^2 = \sum_{j\epsilon s_i}(y_{ij} - \bar{y}_i)^2$ (Holt and Smith 1979). The resulting confidence interval, I_{pst}: $\bar{y}_{pst} \pm z_{\alpha/2}\sqrt{v(\bar{y}_{pst})}$, is conditionally correct. Another variance estimator

$$v^*(\bar{y}_{pst}) = \sum W_i^2\left[E\left(\frac{1}{n_i}\right) - \frac{1}{N_i}\right]s_{iy}^2 \tag{3.4}$$

$$\doteq \left(\frac{1}{n} - \frac{1}{N}\right)\sum W_i s_{iy}^2$$

is conditionally biased, although unbiased when averaged over the whole sample space, S (assuming that $P(n_i \leq 1)$ is negligible). The conditional performance of confidence interval based on (3.4) evidently depends on the extent of divergence of the observed values $1/n_i$ from their expectataions $E(1/n_i)$. It may be noted that the interval I_{pst} is also correct in the unconditional framework, provided $P(n_i \leq 1)$ is negligible for all i.

If $n_i = 1$ for some i, no conditionally unbiased variance estimator can be obtained, but it might be satisfactory to use a collapsed strata method or use the model-based solution of Hartley *et al.* (1969) originally proposed for variance estimation in stratified random sampling with one unit per stratum. Empirical studies might throw some light on the applicability of the latter methods.

The customary justification for preferring $\bar{y}_{pst}$ over $\bar{y}$ is that the unconditional variance of $\bar{y}_{pst}$ is approximately equal to the variance under proportional allocation and hence smaller than the unconditional variance of $\bar{y}$. We are also reminded that gains in efficiency under proportional allocation are likely to be modest. It is more important, however, to note that the sample mean $\bar{y}$ is conditionally biased:

$$E_2(\bar{y}) = \sum w_i \bar{Y}_i \neq \sum W_i \bar{Y}_i = \bar{Y}, \; w_i = \frac{n_i}{n}, \tag{3.5}$$

and hence the resulting inferences could be conditionally incorrect.

Example 1. Suppose $k = 2$ (say, male, female strata with known projected census weights W_1 and $W_2 = 1 - W_1$, or small and big hospitals (Royall 1970)). Royall used a super-population model

$$E_m(y_i) = \beta x_i, \; i = 1, \ldots, N, \; \beta > 0, \; x_i > 0 \tag{3.6}$$

to demonstrate that $\bar{y}$ is model-biased conditionally, where E_m denotes the model expectation, i.e.,

$$E_m(\bar{y}) = \beta \bar{x} \neq E_m(\bar{Y}) = \beta \bar{X} \tag{3.7}$$

unless the sample mean $\bar{x}$ coincides with the population mean $\bar{X}$. In his example, x_i = number of beds in the i-th hospital, y_i = number of occupied beds in the i-th hospital, and $x_1, \ldots, x_N$ are known. Royall argues that $\bar{y}$ leads to serious underestimation if the observed sample contains all (or mostly) small hospitals since $B_m(\bar{y}) = E_m(\bar{y}) - E_m(\bar{Y}) = \beta(\bar{x} - \bar{X})$ and $\bar{x} \ll \bar{X}$. This point can also be illustrated in our conditional framework without assuming a model. The ratio of the conditional bias of $\bar{y}$ to the population of large hospitals, $\bar{Y}_2$, may be expressed as

$$\frac{B_2(\bar{y})}{\bar{Y}_2} = (W_1 - w_1)\delta = (w_2 - W_2)\delta, \tag{3.8}$$

where $B_2(\bar{y}) = E_2(\bar{y}) - \bar{Y}$ denotes the conditional bias of $\bar{y}$, $\delta = (\bar{Y}_2 - \bar{Y}_1)/\bar{Y}_2$ and $0 < \delta < 1$ since the population mean, $\bar{Y}_1$, of small hospitals is smaller than $\bar{Y}_2$. If $w_1 = 1$ (i.e., all small hospitals observed in the sample), then $E_2(\bar{y}) = \bar{Y}_1 \ll \bar{Y}$ and hence $\bar{y}$ is a serious underestimate. Similarly, if $w_1 \gg W_1$ (i.e., mostly small hospitals observed), then it follows from (3.8) that $\bar{y}$ would lead to serious underestimation.

In this example, one should use the post-stratified estimator $\bar{y}_{pst} = W_1\bar{y}_1 + W_2\bar{y}_2$ which is conditionally unbiased unless $n_1 = 0$ or $n_2 = 0$. It might be preferable, in fact, to use a post-stratified ratio estimator

$$\bar{y}_{pst,r} = \frac{\bar{y}_{pst}}{\bar{x}_{pst}} \bar{X}, \tag{3.9}$$

where $\bar{x}_{pst} = W_1\bar{x}_1 + W_2\bar{x}_2$ and $\bar{x}_i$ is the sample mean of x in the i-th stratum. The estimator (3.9) is approximately unbiased conditionally and more efficient than $\bar{y}_{pst}$ if n is large.

Remark 1. In Royall's example, one should, in fact, use a more efficient design than simple random sampling since all the population x-values are known, e.g., stratified random sampling under x-stratification and, perhaps, optimal allocation based on the x-values.

Remark 2. Royall justifies the use of the customary ratio estimator $\bar{y}_r = (\bar{y}/\bar{x})\bar{X}$ under his model (3.6), but it cannot be justified in the conditional (repeated sampling) framework since $\bar{y}_r$ is conditionally biased:

$$B_2(\bar{y}_r) \doteq \bar{X}\left[\frac{w_2\bar{Y}_1 + w_2\bar{Y}_2}{w_1\bar{X}_1 + w_2\bar{X}_2} - R\right], R = \frac{\bar{Y}}{\bar{X}} \tag{3.10}$$

$$\neq 0$$

unless $\bar{y}_1/\bar{x}_1 = \bar{y}_2/\bar{x}_2 = R$. In the extreme case of $w_1 = 1$, $B_2(\bar{y}_r) = \bar{X}(R_1 - R)$ where $R_1 = \bar{Y}_1/\bar{X}_1$. Hence, $B_2(\bar{y}_r) \lesseqgtr 0$ according as $R_1 \lesseqgtr R$.

Remark 3. If the weight W_1 is unknown but $\bar{X}$ is known, we cannot implement either $\bar{y}_{pst}$ or $\bar{y}_{pst,r}$. Royall suggests the use of $\bar{y}_r$ with inference conditional on the observed mean $\bar{x}$. However, the choice $\bar{x}$ is somewhat arbitrary, and the conditional bias of $\bar{y}_r$ could be quite large unless the model (3.6) is true, at least approximately.

If good prior information on W_1 is available, say $W_1^* \leq W_1 \leq W_1^{**}$ where W_1^* and W_1^{**} are known, then one could use the following "pseudo" post-stratified estimator of $\bar{Y}$:

$$\bar{y}_{pst}^* = \tilde{W}_1\bar{y}_1 + \tilde{W}_2\bar{y}_2, \tag{3.11}$$

where $\tilde{W}_1 = w_1$ if $W_1^* \leq w_1 \leq W_1^{**}$, $= W_1^*$ if $w_1 < W_1{}^*$, $= W_1^{**}$ if $w_1 > W_1^{**}$ and $\tilde{W}_2 = 1 - \tilde{W}_1$. The estimator $\bar{y}_{pst}^*$ and its ratio analogue should perform better conditionally given (n_1, n_2) than $\bar{y}$ and $\bar{y}_r$, although biased. Unconditionally, the MSE of $\bar{y}_{pst}^*$ should be smaller than the MSE of $\bar{y}$, provided $W_1^* \leq W_1 \leq W_1^{**}$. One could also utilize a formal Bayesian approach to estimate W_1 by specifying a prior distribution on W_1.

Example 2 (outliers). The problem of estimating a population mean $\bar{Y}$ in the presence of outliers is similar to the hospital example above. Suppose the population is known to contain a small fraction, W_2, of outliers (large observations) but W_2 is unknown, i.e. $W_1 >> W_2$ and $\bar{Y}_2 >> \bar{Y}_1$. Then, if the observed sample contains no outliers (i.e., $w_2 = 0$), we would say that $\bar{y}$ is "far from the true value $\bar{Y}$" (Chinnappa 1976) and yet $\bar{y}$ is (unconditionally) unbiased. The meaning of this statement follows from the fact that $E_2(\bar{y}) = \bar{Y}_1 << \bar{Y}$, where E_2 is the conditional expectation as before.

On the other hand, we would say that $\bar{y}$ is a serious overestimate if the sample contains outliers. This follows from (3.8) noting that $w_2 >> W_2$ (since W_2 is very small). For instance, if $N_2 = 1$ then $w_2 = 1/n >> W_2 = 1/N$. In this situation, we are told to modify the estimate $\bar{y}$ by reducing the weight attached to outliers in the sample. One suggestion is to modify $\bar{y}$ by reducing the weight attached to outliers from $1/n$ to $1/N$ and adjusting the weights for non-outliers such that the n weights sum to 1:

$$\bar{y}^* = \frac{N - n_2}{N}\bar{y}_1 + \frac{n_2}{N}\bar{y}_2. \tag{3.12}$$

The conditional relative bias of $\bar{y}^*$ is given by

$$\frac{B_2(\bar{y}^*)}{\bar{Y}_2} = \left(w_2\frac{n}{N} - W_2\right)\delta, \tag{3.13}$$

whereas $B_2(\bar{y})/\bar{Y}_2 = (w_2 - W_2)\delta$. If $w_2 \frac{n}{N} - W_2 < 0$, then

$$\left|w_2 \frac{n}{N} - W_2\right| = W_2 - w_2 \frac{n}{N} < w_2 - W_2 \text{ if } 2W_2 < w_2\left(1 + \frac{n}{N}\right).$$

The inequality $2W_2 < w_2(1 + n/N)$ should be satisfied since $w_2 \gg W_2$. If $w_2 n/N - W_2 > 0$, then

$$\left|w_2 \frac{n}{N} - W_2\right| = w_2 \frac{n}{N} - W_2 < w_2 - W_2.$$

Hence, the estimator $\bar{y}^*$ should have a smaller absolute value of conditional bias than $\bar{y}$.

The estimator $\bar{y}^*$ is essentially obtained from the post-stratified estimator $\bar{y}_{pst}$ by pretending that $N_2 = n_2$. A more satisfactory solution can be obtained by gathering good prior information on $W_1(= 1 - W_2)$, say from census data, and then using the estimator $\bar{y}^*_{pst}$ or the estimator based on a Bayes estimator of W_1.

Hidiroglou and Srinath (1981) derived the conditional bias and conditional and unconditional MSE of $\bar{y}$, $\bar{y}^*$ and some other modifications of $\bar{y}$, but they did not compare the conditional biases of $\bar{y}$ and $\bar{y}^*$ as above.

3.2 $n_i = 0$ for Some i

If the total sample size, n, is small or if too many post-strata chosen, then n_i could be zero for some i. The post-stratified estimator (3.2) in this case reduces to

$$\bar{y}_{pst} = \sum{}' W_i \bar{y}_i, \tag{3.14}$$

where $\sum{}'$ denotes summation over strata with nonzero n_i. The estimator (3.14) is conditionally biased:

$$E_2(\bar{y}_{pst}) = \sum{}' W_i \bar{Y}_i \neq \sum W_i \bar{Y}_i. \tag{3.15}$$

It remains conditionally biased even under the strong assumption $\bar{Y}_i = \bar{Y}$ for all i, which incidentally shows that $\bar{y}_{pst}$ could lead to serious underestimation. It is also unconditionally biased. One commonly used method to overcome these difficulties is to collapse similar strata to ensure that $n_i > 0$ for all i in the reduced set of strata. Fuller (1966) proposed a more efficient solution for the special case of $k = 2$ post-strata, but his framework is unconditional in the sense that the probability, P_1^*, of $n_1 = 0$ given that either $n_1 = 0$ or $n_2 = 0$, is brought into the picture. His estimator is given by

$$\begin{aligned} \bar{y}_F &= \frac{W_1}{P_1^*}\bar{y}_1 \text{ if } n_2 = 0 \\ &= \frac{W_2}{P_2^*}\bar{y}_2 \text{ if } n_1 = 0, \end{aligned} \tag{3.16}$$

where $P_2^* = 1 - P_1^*$. The estimator $\bar{y}_F$ is conditionally unbiased given that either $n_1 = 0$ or $n_2 = 0$, but is conditionally biased given (n_1, n_2), even in the case $\bar{Y}_1 = \bar{Y}_2 = \bar{Y}$.

An unconditionally unbiased estimator is given by

$$\bar{y}_D = \sum \frac{a_i}{E(a_i)} W_i \bar{y}_i, \tag{3.17}$$

(Doss *et al.*, 1979), where $a_i = 1$ if at least one unit from stratum i in the sample, $= 0$ otherwise, and $\bar{y}_i$ is defined as $\bar{Y}_i$ if $n_i = 0$ (note that $a_i\bar{y}_i = 0$ if $n_i = 0$ even though $\bar{Y}_i$ is unknown). The estimator $\bar{y}_D$, however, is conditionally biased since

$$E_2(\bar{y}_D) = \Sigma' \frac{W_i\bar{Y}_i}{E(a_i)} \neq \Sigma W_i\bar{Y}_i = \bar{Y}.$$

It remains conditionally biased even if $\bar{Y}_i = \bar{Y}$ for all i.

Doss *et al.* criticized $\bar{y}_D$ on the grounds that it is not translation-invariant (i.e., $\bar{y}_D$ does not change to $\bar{y}_D + c$ when each y_i is changed to $y_i + c$, where c is an arbitrary constant), and hence that the variance of $\bar{y}_D$, when y_i is changed to $y_i + c$, can be made arbitrarily large by increasing c sufficiently. On the other hand, the ratio estimator

$$\bar{y}_{rD} = \frac{\Sigma \frac{a_i}{E(a_i)} W_i\bar{y}_i}{\Sigma \frac{a_i}{E(a_i)} W_i}, \tag{3.18}$$

proposed by Doss *et al.*, is translation-invariant. It is conditionally biased, but the conditional bias is approximately zero if $\bar{Y}_i = \bar{Y}$ for all i, unlike the conditional bias of $\bar{y}_D$. Another ratio estimator which is similar to $\bar{y}_{rD}$ conditionally is given by

$$\bar{y}_{r(pst)} = \frac{\Sigma' W_i\bar{y}_i}{\Sigma' W_i}, \tag{3.19}$$

but it is inconsistent unconditionally, unlike $\bar{y}_{rD}$. Hence, $\bar{y}_{rD}$ may be preferred to $\bar{y}_{r(pst)}$ or $\bar{y}_D$.

If concomitant information on *all* strata is available, then one could fit a model to the observed strata means $\bar{y}_i$ and predict the population means of strata with $n_i = 0$. For example, if the population means $\bar{X}_i$ of a concomitant variable are linearly related to the corresponding $\bar{Y}_i$, then the predicted value of a $\bar{Y}_i$ is given by $\hat{\alpha} + \hat{\beta}\bar{X}_i = \bar{y}_i^*$ (say), where $\hat{\alpha}$ and $\hat{\beta}$ are the least squares estimators obtained by minimising $\Sigma'(\bar{y}_i - \alpha - \beta\bar{X}_i)^2$. The resulting estimator of $\bar{Y}$ is given by

$$\bar{y}^*_{pst} = \Sigma' W_i\bar{y}_i + \Sigma'' W_i\bar{y}_i^*, \tag{3.20}$$

where Σ'' denotes summation over strata with $n_i = 0$. This estimator should have good conditional properties if the fitted model is adequate. It should be clear from this discussion that there is no simple solution if $n_i = 0$ for some of the strata.

4. TWO-WAY STRATIFICATION

Ingenious designs to improve the efficiency of estimators have been proposed in the literature. Bryant *et al.* (1960) proposed a design involving two-way stratification in which the sample sizes n_{ij} are zero for some strata (cells). Their method is supposed to permit estimation of the population mean even when the total sample size n is less than the total number of strata. Using proportional allocation for the marginal sample sizes $(n_{i.}, n_{.j})$, they obtained a random allocation n_{ij} such that $E(n_{ij}) = (n_{i.}n_{.j})/n = nW_{i.}W_{.j}$, where $W_{i.}$ and $W_{.j}$ are the row and column marginal totals of cell weights W_{ij}.

Bryant *et al.* proposed the estimator

$$\bar{y}_U = \frac{1}{n}\Sigma\Sigma n_{ij}G_{ij}\bar{y}_{ij}, \tag{4.1}$$

where $G_{ij} = n^2 W_{ij}/(n_{i.}n_{.j})$ and $\bar{y}_{ij}$ may be taken as $\bar{Y}_{ij}$ if $n_{ij} = 0$. The estimator $\bar{y}_U$ is unconditionally unbiased. However, the distribution of n_{ij} is completely known (since all W_{ij} are known) and hence the relevant reference set is the set of samples having the observed configuration $\{n_{ij}\}$, i.e., one should treat the design as stratified simple random sampling for inference purposes. The estimator $\bar{y}_U$ is conditionally biased:

$$E_2(\bar{y}_U) = \sum\sum(\frac{n_{ij}G_{ij}}{n})\bar{Y}_{ij} \neq \sum\sum W_{ij}\bar{Y}_{ij} = \bar{Y},$$

noting that $E_2(\bar{y}_{ij}) = \bar{Y}_{ij}$ if $n_{ij} > 0$. It also has the defects of $\bar{y}_D$ in the previous section which can be circumvented by using the ratio estimator

$$\bar{y}_r = \frac{\bar{y}_U}{\bar{a}_U} = \frac{\sum\sum n_{ij}G_{ij}\bar{y}_{ij}}{\sum\sum n_{ij}G_{ij}} \qquad (4.2)$$

where $\bar{a}_U = \sum\sum n_{ij}G_{ij}/n$. $\bar{y}_r$ is also conditionally biased, but the conditional bias is approximately zero if $\bar{Y}_{ij} = \bar{Y}$ for all (i, j). The latter condition, however, may be unrealistic in the present context since the strata are different by design.

As in Section 3.1, it seems necessary to use a model connecting the sampled and nonsampled strata. A reasonable model, in the absence of concomitant information, is to assume that

$$y_{ijk} = \mu + \beta_j + \tau_i + \varepsilon_{ijk} \qquad (4.3)$$

where y_{ijk} is the k-th observation in the (i, j)-th cell, β_j and τ_i are fixed effects and ε_{ijk} are independent errors with zero mean and common variance σ^2. Unfortunately, the linear combination $\mu + \beta_j + \tau_i$ for nonsampled strata is not estimable from sample data and hence the corresponding $\bar{Y}_{ij}$ cannot be predicted. This difficulty can be avoided by assuming that β_j and τ_i are random variables and then obtaining a predictor $\hat{\mu} + \hat{\beta}_j + \hat{\tau}_i$, but the random effects model may be less realistic than (4.3) in the present context.

Motivated by the above-mentioned difficulty, Bankier (1985) discussed a raking procedure in the context of independent stratified samples according to two different criteria of stratification. His estimator is approximately model-unbiased under the fixed effects model (4.3), while the usual Horvitz-Thompson estimator and its ratio extension are model-biased.

Bankier's method can be adapted to the two-way stratification problem. The raking ratio estimator of $\bar{Y}$ is given by

$$\bar{y}(p) = \sum\sum\frac{G_{ij}(p)}{n}y_{ij} \qquad (4.4)$$

where y_{ij} is the sample total in the (i, j)-th cell ($y_{ij} = 0$ of $n_{ij} = 0$) and $G_{ij}(p)$ are the values obtained in the p-th iteration of the raking procedure such that

$$\sum_j\frac{G_{ij}(p)}{n}n_{ij} \doteq W_{i.} = \sum_j W_{ij} \qquad (4.5)$$

and

$$\sum_i \frac{G_{ij}(p)}{n}\, n_{ij} \doteq W_{.j} = \sum_i W_{ij}.$$

The $G_{ij}(p)$ are obtained as follows: Let $G_{ij}(0) = G_{ij} > 0 \;\forall(i, j)$, and

$$\begin{aligned} G_{ij}(p) &= G_{ij}(p-1)\,\frac{W_{i.}}{\sum_j \frac{G_{ij}(p-1)}{n}\, n_{ij}} \quad \textit{if } p \text{ is odd} \\ &= G_{ij}(p-1)\,\frac{W_{.j}}{\sum_i \frac{G_{ij}(p-1)}{n}\, n_{ij}} \quad \textit{if } p \text{ is even.} \end{aligned} \tag{4.6}$$

Under the fixed effects model (4.3), we have

$$\begin{aligned} E_m[\bar{y}(p)] &\doteq \mu + \sum_i W_{i.}\tau_i + \sum_j W_{.j}\beta_j = E_m(\sum \sum W_{ij}\bar{Y}_{ij}) \\ &= E_m(\bar{Y}), \end{aligned}$$

i.e. $\bar{y}(p)$ is approximately model-unbiased. Since $E(G_{ij}(0)n_{ij}/n) = W_{ij}$ for the choice $G_{ij}(0) = G_{ij}$, these starting values should be good. However, we may encounter convergence problems with the raking process because of the many empty cells ($n_{ij} = 0$) resulting from the Bryant *et al.* design. We hope to investigate these convergence problems as well as the conditional properties of the raking ratio estimator (4.4) in a separate paper.

If the population means $\bar{X}_{ij}$ of a concomitant variable x are known for *all* strata, then one could fit a model to the observed strata means $\bar{y}_{ij}$, as in Section 3.1. For example, the model $\bar{y}_{ij} = \beta\bar{x}_{ij} + b_j + t_i + \bar{\varepsilon}_{ij}$ with random effects b_j and t_i might be reasonable, where $\bar{\varepsilon}_{ij}$ is the sample mean of errors ε_{ijk} in the (i, j)-th cell. A predictor $\hat{\beta}\bar{x}_{ij} + \hat{b}_j + \hat{t}_i$ of $\bar{Y}_{ij}$ for nonsampled strata may be used in conjunction with the observed means $\bar{y}_{ij}$ to arrive at an estimator of $\bar{Y}$. This approach is similar to modelling for small area estimates, except that the parameter of interest here is the overall mean $\bar{Y}$ rather than the individual cell means $\bar{Y}_{ij}$. We hope to investigate the conditional properties of alternative estimators of $\bar{Y}$ in a separate paper.

5. NONRESPONSE

5.1 A Simple Model

Suppose m responses are obtained in a simple random sample of size n. Let W_1 denote the proportion in the response stratum and $\bar{Y} = W_1\bar{Y}_1 + W_2\bar{Y}_2$ the population mean, where $\bar{Y}_1$ and $\bar{Y}_2$ are the means of response and nonresponse strata respectively, and $W_2 = 1 - W_1$. In this situation, conditioning on the observed value of m can be questioned since the distribution of m depends on the unknown W_1 which is involved in the parameter of interest. Also, the sample mean $\bar{y}_m$ of respondents is unconditionally biased because $E(\bar{y}_m) = \bar{Y}_2 \neq \bar{Y}$. Hence, it is necessary to assume a model for response mechanism even in the unconditional framework, unless a subsample of nonrespondents is also sampled.

A simple model assumes that the probability of response if contacted is the same for all units, say p^*, i.e., data are missing at random. Under this model, the distribution of m depends only on p^*, and hence we should condition on m if p^* is assumed known (or at least partially known or unrelated to $\bar{Y}$). Oh and Scheuren (1983) have shown that conditionally given m the sample s_m of respondents is like a simple random sample of size m from the *whole* population. Hence, $\bar{y}_m$ is conditionally unbiased, and its conditional variance is unbiasedly estimated by

$$v_2(\bar{y}_m) = (m^{-1} - N^{-1})s^2_{my}, \tag{5.1}$$

where $(m - 1)s^2_{my} = \sum_{i\epsilon s_m}(y_i - \bar{y}_m)^2$. The resulting confidence interval $\bar{y}_m \pm z_{\alpha/2}\sqrt{v_2(\bar{y}_m)}$ is conditionally correct, at least approximately, if m is not small.

On the other hand, the Horvitz-Thompson estimator (p^* known):

$$\bar{y}_{HT} = \frac{m}{E(m)}\,\bar{y}_m = \sum_{i\epsilon s_m} \frac{y_i}{np^*} \tag{5.2}$$

is conditionally biased, as in Section 2, although unbiased when averaged over the distribution of m. For general designs, the ratio estimator

$$\hat{\bar{Y}}_{HT,r} = \frac{\sum_{s_m} \frac{y_i}{\pi_i p_i^*}}{\sum_{s_m} \frac{1}{\pi_i p_i^*}} \tag{5.3}$$

is often used on grounds of efficiency, where π_i is the probability of inclusion and p_i^* is the probability of response if contacted (assumed known) for the i-th unit. In the simple case of $p_i^* = p^*$ and simple random sampling, it is interesting to note that $\hat{\bar{Y}}_{HT,r}$ reduces to $\bar{y}_m$. Hence, the ratio estimator might perform well in a conditional framework, for general designs.

5.2 A More Realistic Model

A more realistic model assumes that data are missing at random within post-strata with known weights W_i. Let n_i and m_i respectively denote the sample size and the respondent sample size in the i-th post-stratum. Then the joint distribution of (n_i, m_i) depends only on the W_i and the response probabilities within post-strata. Hence, we should condition on the observed value of (n_i, m_i) provided the post-stratum response probabilities are either known or unrelated to the parameters of interest, viz., the post-strata means. Conditionally, the observed sample is like a stratified simple random sample with fixed strata sizes m_i (Oh and Scheuren 1983). Hence, the estimator

$$\bar{y}_{pst,m} = \sum W_i\bar{y}_{mi} \tag{5.4}$$

is conditionally unbiased, and its conditional variance is unbiasedly estimated by

$$v_2(\bar{y}_{pst,m}) = \sum W_i^2\left(\frac{1}{m_i} - \frac{1}{M_i}\right)s^2_{miy} \tag{5.5}$$

where $\bar{y}_{mi}$ and s^2_{miy} are the mean and variance of sample respondents in the i-th post-stratum, respectively.

If the W_i are unknown, it is a common practice to replace W_i in (5.4) by its estimate $w_i = n_i/n$. In this case, conditional inference can be questioned since the distribution of (n_i, m_i) depends on the unknown weights W_i and since W_i are involved in the parameter $\bar{Y} = \sum W_i \bar{Y}_i$. If partial information on W_i, in the form of bounds on W_i, is available, we can proceed with conditional inference as in Example 1, Remark 3, although the resulting estimator is still conditionally biased (but likely to be better than (5.4) with W_i replaced by w_i).

6. DOMAIN ESTIMATION (SRS)

6.1 Domain mean

Under simple random sampling (SRS), the usual estimator of a subpopulation (domain) mean, $\bar{Y}_i$, is given by the sample mean

$$\bar{y}_i = \sum_{j \epsilon s_i} \frac{y_j}{n_i}, \; n_i > 0 \tag{6.1}$$

where s_i is the sample falling in the domain and n_i is the corresponding size.

If the domain size, N_i, is known, then one should condition on the observed value, n_i. The estimator $\bar{y}_i$ is conditionally unbiased if $n_i > 0$ since conditionally s_i is a SRS sample of fixed size n_i from the domain. An unbiased estimate of the conditional variance is

$$v(\bar{y}_i) = \left(\frac{1}{n_i} - \frac{1}{N_i}\right) s^2_{iy}, \; n_i > 0 \tag{6.2}$$

and the resulting confidence interval $\bar{y}_i \pm z_{\alpha/2} \sqrt{v(\bar{y}_i)}$ is conditionally correct.

The estimator $\bar{y}_i$, however, is unstable for small domains (small areas) with small n_i. Also $\bar{y}_i$ is not defined if $n_i = 0$. One solution to the latter problem, suggested in the literature, is to use a modified estimator.

$$\bar{y}_i' = \frac{a_i}{E(a_i)} \bar{y}_i, \; n_i \geq 0 \tag{6.3}$$

where $a_i = 1$ if $n_i \geq 1$; $= 0$ if $n_i = 0$ and $\bar{y}_i$ is taken as $\bar{Y}_i$ if $n_i = 0$. The estimator $\bar{y}_i'$, however, is conditionally biased:

$$E_2(\bar{y}_i') = \frac{a_i}{E(a_i)} \bar{Y}_i.$$

It is an underestimate if $n_i = 0$, and an overestimate if $n_i \geq 0$, although unconditionally unbiased. The extent of overestimation depends on the magnitude of $E(a_i) = P(n_i \geq 1)$. If, for example, $P(n_i \geq 1) = 0.75$, then $E_2(\bar{y}_i') = (4/3)\bar{Y}_i$ if $n_i \geq 1$.

Sarndal (1984) proposed the following estimator in the context of small area estimation:

$$\bar{y}_{iS} = \bar{y} + \frac{w_i}{W_i}(\bar{y}_i - \bar{y}), \; n_i \geq 0, \tag{6.4}$$

where $\bar{y} = \sum w_j \bar{y}_j$ is the overall sample mean and $w_i = n_i/n$. The estimator is approximately unconditionally unbiased, but conditionally biased unless $w_i = W_i$:

$$B_2(\bar{y}_{iS}) = (\frac{w_i}{W_i} - 1)(\bar{Y}_i - \bar{Y}'), \tag{6.5}$$

where $\bar{Y}' = \sum w_i \bar{Y}_i$. If $n_i = 0$, the estimator $\bar{y}_{iS}$ reduces to the "synthetic" estimator $\bar{y}$. The extent of under- (or over-) estimation of $\bar{y}_{iS}$ depends on both $w_i/W_i - 1$ and $\bar{Y}_i - \bar{Y}'$ and hence more complex to analyse than the bias of $\bar{y}_i'$. However, $\bar{y}_{iS}$ would have a larger absolute conditional bias* than $\bar{y}$ if $w_i > 2W_i$ (and hence a larger conditional MSE). Also, the conditionally unbiased estimator $\bar{y}_i$ has a smaller conditional variance than $\bar{y}_{iS}$ if $w_i > W_i$ (neglecting the variance of $\bar{y}$ relative to that of $\bar{y}_i$) and hence smaller conditional MSE.

Hidiroglou and Sarndal (1985) proposed a modification of $\bar{y}_{iS}$:

$$\bar{y}_{iS}^{**} = \begin{cases} \bar{y}_i \text{ if } w_i \geq W_i \\ \bar{y}_{iS}^{*} = \bar{y} + \left(\frac{w_i}{W_i}\right)^2 (\bar{y}_i - \bar{y}) \text{ if } w_i < W_i. \end{cases} \tag{6.6}$$

The estimator $\bar{y}_{iS}^{**}$ is conditionally unbiased if $w_i \geq W_i$, while its conditional absolute bias is smaller than that of $\bar{y}$ if $w_i < W_i$. A motivation for $\bar{y}_{iS}^{**}$ is that the conditional variance of $\bar{y}_{iS}^{*}$ (or $\bar{y}_{iS}$) is larger than that of $\bar{y}_i$ (neglecting the variance of $\bar{y}$ relative to that of $\bar{y}_i$) if $w_i > W_i$, while the conditional variance of $\bar{y}_{iS}^{*}$ is smaller than that of $\bar{y}_{iS}$ if $w_i < W_i$. However, the absolute conditional bias of $\bar{y}_{iS}^{*}$ is larger than that of $\bar{y}_{iS}$ if $w_i < W_i$. Hence, the choice between $\bar{y}_{iS}^{*}$ and $\bar{y}_{iS}$ in the case $w_i < W_i$ is not clear-cut and no simple recipe seems to exist.

Drew *et al.* (1982) proposed another sample size dependent estimator which depends on a parameter K_0. In the SRS case and the choice $K_0 = 1$, their estimator reduces to

$$\bar{y}_{iD} = \begin{cases} \bar{y}_i \text{ if } w_i \geq W_i \\ \bar{y}_{iS} \text{ if } w_i < W_i. \end{cases} \tag{6.7}$$

As noted above, the choice between $\bar{y}_{iS}$ and $\bar{y}_{iS}^{*}$ in the case $w_i < W_i$ is not clear-cut. Consequently, the choice between $\bar{y}_{iD}$ and $\bar{y}_{iS}^{**}$ is also not clear-cut.

If N_i is unknown, the conditional argument may still be relevant provided N_i is unrelated to the parameter of interest $\bar{Y}_i$. It is also relevant when partial information on N_i is available, such as bounds on N_i.

If a concomitant variable x with known domain mean $\bar{X}_i$ is available, the ratio estimator

$$\bar{y}_{ir} = \frac{\bar{y}_i}{\bar{x}_i} \bar{X}_i \tag{6.8}$$

*Sarndal's estimator, however, should perform better in the case of a one-way model. The estimator is obtained by pooling estimators of the form (6.4) over two or more groups.

and a regression-type estimator (Battese and Fuller 1981)

$$\bar{y}^*_{ir} = \bar{y}_i + \frac{\bar{y}}{\bar{x}}(\bar{X}_i - \bar{x}_i) \tag{6.9}$$

are both conditionally unbiased (approximately), but $\bar{y}^*_{ir}$ is likely to be more efficient if a regression model (through the origin) with a common slope holds true, at least approximately, for the small areas. If the slopes are varying, then an empirical Bayes estimator, which is more complex, might be more relevant (Dempster *et al.* 1981).

6.2 Domain Total

If N_i is known, then an estimate of domain total $Y_i = N_i\bar{Y}_i$ is simply obtained by multiplying a chosen estimator of $\bar{Y}_i$ by N_i. On the other hand, the usual unbiased estimator

$$\hat{Y}_i = \hat{N}_i\,\bar{y}_i = \frac{N}{n}\sum_{j\epsilon s_i} Y_j,\ n_i \geq 1 \tag{6.9}$$

is used if N_i is unknown, where $\hat{N}_i = Nw_i$ is the unbiased estimator of N_i and $P(n_i = 0)$ is assumed to be negligible.

Suppose now that we have prior information, say $N_i^* \leq N_i \leq N_i^{**}$. Then the conditional argument may be relevant. The conditional bias of $\hat{Y}_i$ is

$$B_2(\hat{Y}_i) = (\hat{N}_i - N_i)\bar{Y}_i. \tag{6.10}$$

It follows from (6.10) (assuming $\bar{Y}_i > 0$) that $B_2(\hat{Y}_i) > 0$, i.e., overestimation, if $\hat{N}_i > N_i$ and that $B_2(\hat{Y}_i)$ increases as the domain sample size n_i increases. Similarly, $B_2(\hat{Y}_i) < 0$, i.e., underestimation, if $\hat{N}_i < N_i$ and $|B_2(\hat{Y}_i)|$ increases as n_i decreases; the conditional bias is zero if $\hat{N}_i = N_i$.

Utilizing the prior information, we can modify $\hat{Y}_i$ as

$$\hat{Y}_i^* = \begin{cases} N_i^*\bar{y}_i & \text{if } \hat{N}_i < N_i^* \\ \hat{N}_i\,\bar{y}_i & \text{if } N_i^* \leq \hat{N}_i \leq N_i^{**} \\ N_i^{**}\bar{y}_i & \text{if } \hat{N}_i > N_i^{**}. \end{cases} \tag{6.11}$$

The absolute conditional bias of $\hat{Y}_i^*$ is smaller than that of $\hat{Y}_i$ if either $\hat{N}_i < N_i^*$ or $\hat{N}_i > N_i^{**}$, while $\hat{Y}_i^* = \hat{Y}_i$ in the interval $N_i^* \leq \hat{N}_i \leq N_i^{**}$. Hence, $\hat{Y}_i^*$ is conditionally better than the unbiased estimator $\hat{Y}_i$. Also the unconditional MSE of $\hat{Y}_i^*$is smaller than that of $\hat{Y}_i$, although $\hat{Y}_i^*$ is unconditionally biased. Unfortunately, there is no simple way to improve upon $\hat{Y}_i^*$ in the range $N_i^* \leq \hat{N}_i \leq N_i^{**}$. In any case, $\hat{Y}_i^*$ should be preferred over $\hat{Y}_i$. Good supplementary information on the domain size is necessary in estimating a domain total efficiently.

7. GENERAL DESIGNS

Post-stratification adjustment is commonly employed in complex large-scale surveys, mainly to increase the efficiency of estimators, e.g., the age-sex adjustment in the Canadian Labour Force Survey (LFS). A general theory of unconditional inference is also available.

The estimator of total Y is given by

$$\hat{Y}_{pst} = \sum M_i \frac{\hat{Y}_i}{\hat{M}_i} \tag{7.1}$$

where $\hat{Y}_i$ and $\hat{M}_i$ are the usual unbiased domain estimators of the i-th post-stratum total Y_i and size M_i respectively. In the LFS, projected census counts are used for the M_i. The estimator $\hat{Y}_{pst}$ reduces to $\sum N_i \bar{y}_i$ in the SRS case (see (3.2)) and we have already seen that $\sum N_i \bar{y}_i$ is conditionally unbiased in the SRS case (assuming all $n_i \geq 1$). However, for complex designs it seems difficult to investigate the conditional properties of (7.1); even the choice of reference set is not so clear-cut. To illustrate this difficulty, consider stratified SRS with $L = 2$ strata and $k = 2$ post-strata. If we condition on the observed post-strata sample sizes (n_{h1}, n_{h2}) in each stratum h, the theory is straightforward provided the post-strata sizes N_{hi} in each stratum are known. However, in practice we will run into problems with zero sample sizes n_{hi} and also the sizes N_{hi} in each stratum may not be available or the projections inaccurate, although $N_{.i} = \sum_h N_{hi} = M_i$ are available. Hence, we may prefer to condition on the observed total sample sizes $(n_{.1}, n_{.2})$, where $n_{.i} = \sum_h n_{hi}$.

The estimator $\hat{Y}_{pst}$ in this special case of stratified SRS ($L = 2$, $k = 2$) reduces to

$$\hat{Y}_{pst} = N_{.1} \frac{N_{1.} \frac{y_{11}}{n_{1.}} + N_{2.} \frac{y_{21}}{n_{2.}}}{N_{1.} \frac{n_{11}}{n_{1.}} + N_{2.} \frac{n_{21}}{n_{2.}}} + N_{.2} \frac{N_{1.} \frac{y_{12}}{n_{1.}} + N_{2.} \frac{y_{22}}{n_{2.}}}{N_{1.} \frac{n_{12}}{n_{1.}} + N_{2.} \frac{n_{22}}{n_{2.}}} \tag{7.2}$$

where $N_{h.} = N_{h1} + N_{h2}$ and $n_{h.} = n_{h1} + n_{h2}$ are the strata population and sample sizes respectively, and y_{hi} are the sample totals in the (h, i)-th cell. The conditional expectation of (7.2) given $(n_{.1}, n_{.2})$ is not tractable since one has to evaluate the sum

$$E_2(\hat{Y}_{pst}) = \sum_t p(s_t | n_{.1}, n_{.2}) \hat{Y}_{pst}(t) \tag{7.3}$$

where s_t is a possible sample such that the observed sample sizes $\tilde{n}_{hi}$ satisfy $\tilde{n}_{1i} + \tilde{n}_{2i} = n_{.i}$ ($i = 1, 2$), and $\hat{Y}_{pst}(t)$ is the value of (7.2) for the sample s_t, and $p(s_t | n_{.1}, n_{.2})$ is the conditional probability of observing s_t given $(n_{.1}, n_{.2})$:

$$p(s_t | n_{.1}, n_{.2}) = \left[\sum_{n_{11}=0}^{n_1} \binom{N_{11}}{n_{11}} \binom{N_{12}}{n_{1.} - n_{11}} \binom{N_{21}}{n_{.1} - n_{11}} \binom{N_{22}}{n_{2.} - n_{.1} + n_{11}} \right]^{-1}. \tag{7.4}$$

It is clear from (7.3) and (7.4), however, that $E_2(\hat{Y}_{pst}) \neq Y$ since $\hat{Y}_{pst}$ does not depend on the cell totals N_{hi} unlike $p(s_t | n_{.1}, n_{.2})$.

Turning to variance estimation, the usual formula for general designs is given by

$$v^*(\hat{Y}_{pst}) = v(z_t^*) \tag{7.5}$$

where $v(y_t) = v(\hat{Y})$ is the usual variance estimator of the estimated total $\hat{Y}$, and $v(z_t^*)$ is obtained from $v(\hat{Y})$ by replacing y_t by

$$z_t^* = y_t - \sum_i \frac{\hat{Y}_i}{M_i} a_t(i) \tag{7.6}$$

where $a_t(i) = 1$ if the t-th element belongs to the i-th post-stratum and $a_t(i) = 0$ otherwise (Williams 1962). In the SRS case, (7.5) reduces to

$$v^*(\hat{Y}_{pst}) \doteq N^2 \left(\frac{1}{n} - \frac{1}{N}\right) \sum n_i s_{iy}^2 \tag{7.7}$$

(assuming $(n_i - 1)/(n-1) \doteq n_i/n$) which is not equal to (3.3) when multiplied by N^2. Hence, (7.5) does not behave well in the conditional framework, even in the SRS case. On the other hand, a new variance estimator

$$v(\hat{Y}_{pst}) = v(z_t), \tag{7.8}$$

where

$$z_t = \sum_i \frac{M_i}{\hat{M}_i} (y_t(i) - \frac{\hat{Y}_i}{\hat{M}_i} a_t(i)) \tag{7.9}$$

and $y_t(i) = y_t$ if the t-th element belongs to the i-th post-stratum and $y_t(i) = 0$ otherwise, might be preferable over $v^*(\hat{Y}_{pst})$ since in the SRS case it reduces to (3.3) when multiplied by N^2 and the finite population correction is ignored:

$$v(\hat{Y}_{pst}) = \sum_i \frac{N_i^2}{n_i} s_{iy}^2. \tag{7.10}$$

Some theory for ratio estimators under models also suggests that $v(\hat{Y}_{pst})$ might perform better conditionally than $v^*(\hat{Y}_{pst})$. In any case, there is no harm in switching to (7.8) since it is asymptotically equivalent to the customary variance estimator (7.5), unconditionally.

8. DISCUSSION

Our study clearly shows that conditional inference for complex designs involves formidable difficulties. Nevertheless, we should not use conventional procedures blindly. In those cases where conditional inference is feasible, as in the SRS case, we should certainly employ conditionally relevant methods as elaborated in Sections 2 - 6, while in the more complex cases we should at least make simple modifications to conventional methods, as in (7.8), so that they agree with known, conditionally correct results in special cases. Clearly, we need more research in this area.

ACKNOWLEDGEMENTS

This paper is based on my lectures given at a workshop on conditional inference in sample surveys. I am thankful to Mrs. Nanjamma Chinnappa for organizing the workshop at Statistics Canada. Constructive comments from colleagues at Statistics Canada and Professor D. Holt were helpful in preparing this paper.

REFERENCES

BANKIER, M. (1985). Conditionally unbiased estimators based on any number of independent stratified samples. Memorandum, Business Survey Methods Division, Statistics Canada.

BATTESE, G.E., and FULLER, W.A. (1981). Prediction of county crop areas using survey and satellite data. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 500-505.

BRYANT, E.C., HARTLEY, H.O., and JESSEN, R.J. (1960). Design and estimation in two-way stratification. *Journal of the American Statistical Association,* 55, 105-124.

CHINNAPPA, B.N. (1976). A preliminary note on methods of dealing with unusually large units in sampling from skew populations. Unpublished Technical Report, Institution and Agriculture Survey Methods Division, Statistics Canada.

COX, D.R., and HINKLEY, D.V. (1974). *Theoretical Statistics.* London: Chapman and Hall.

DEMPSTER, A.P., RUBIN, D.B., and TSUTAKAWA, R.K. (1981). Estimation in covariance component models. *Journal of the American Statistical Association,* 76, 341-353.

DOSS, D.C., HARTLEY, H.O., and SOMAYAJULU, G.R. (1979). An exact small sample theory for post-stratification. *Journal of Statistical Planning and Inference,* 3, 235-248.

DREW, J.H., SINGH, M.P., and CHOUDHRY, H. (1982). Evaluation of small area estimation techniques for the Canadian Labour Force Survey. *Survey Methodology,* 8, 17-47.

DURBIN, J. (1969). Inferential aspects of the randomness of sample size in survey sampling. In *New Developments in Survey Sampling* (Eds. N.L. Johnson and H. Smith), New York: Wiley - Interscience.

FISHER, R.A. (1925). *Statistical Methods for Research Workers.* Edinburgh: Oliver and Boyd (5th Ed., 1934).

FULLER, W.A. (1966). Estimation employing post strata. *Journal of the American Statistical Association,* 61, 1172-1183.

HARTLEY, H.O., RAO, J.N.K., and KIEFER, G. (1969). Variance estimation with one unit per stratum. *Journal of the American Statistical Association,* 64, 841-851.

HIDIROGLOU, M.H., and SÄRNDAL, C.E. (1985). An empirical study of some regression estimators for small domains. *Survey Methodology,* 11, 65-77.

HIDIROGLOU, M.H., and SRINATH, K.P. (1981). Some estimators of the population total from simple random samples containing large units. *Journal of the American Statistical Association,* 76, 690-695.

HOLT, D., and SMITH, T.M.F. (1979). Post-stratification. *Journal of the Royal Statistical Society,* Ser. A, 142, 33-46.

LAHIRI, D.B. (1969). On the unique sample, the surveyed one. Unpublished Technical Report, Indian Statistical Institute.

OH, H.L., and SCHEUREN, F.J. (1983). Weighting adjustment for unit nonresponse. In *Incomplete Data in Sample Surveys,* Vol. 2, Academic Press, 142-184.

ROYALL, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika,* 57, 377-387.

SARNDAL, C.E. (1984). Design-consistent versus model-dependent estimators for small domains. *Journal of the American Statistical Association,* 79, 624-631.

WILLIAMS, W.H. (1962). The variance of an estimator with post-stratified weighting. *Journal of the American Statistical Association,* 57, 622-627.

YATES, F. (1984). Tests of significance for 2 × 2 contingency tables. *Journal of the Royal Statistical Society,* Ser. A, 147, 426-463.

Survey Methodology, June 1985
Vol. 11, No. 1, pp. 33–50
Statistics Canada

Cost-Variance Optimization for the Canadian Labour Force Survey

G.H. CHOUDHRY, H. LEE, and J.D. DREW[1]

ABSTRACT

The cost-variance optimization of the design of the Canadian Labour Force Survey was carried out in two steps. First, the sample designs were optimized for each of the two major area types, the Self-Representing (SR) and the Non-Self-Representing (NSR) areas. Cost models were developed and parameters estimated from a detailed field study and by simulation, while variances were estimated using data from the Census of Population. The scope of the optimization included the allocation of sample to the two stages in the SR design, and the consideration of two alternatives to the old design in NSR areas. The second stage of optimization was the allocation of sample to SR and NSR areas.

KEY WORDS: Multi-stage designs; Sample allocation; Linear cost function; Components of variance.

1. INTRODUCTION

The Canadian Labour Force Survey (LFS) is a monthly household survey conducted by Statistics Canada to produce estimates for various labour force characteristics. It follows a stratified multi-stage rotating sample design with six rotation groups. Since its inception in 1945, the survey has undergone a sample redesign following each decennial census of population. These redesigns serve to update the sample to reflect population changes. They also provide the opportunity to introduce improved sampling and estimation methodologies, and to respond to shifts in information needs to be satisfied by the survey.

The 1981 post censal redesign effort included a research phase as outlined in an earlier paper (Singh and Drew 1981) in which all aspects of the survey design were examined in an effort to improve the cost efficiency of the survey vehicle. Highlights of the research program were presented by Singh, Drew, and Choudhry (1984). This report deals with the research aimed at cost-variance optimization of the sample design.

The two important factors in the choice of a sample design are the total cost and the reliability of the resulting estimates. The optimum solution can be obtained by minimizing either total cost or total variance when the other is fixed. Equivalently, the approach we have followed is one of minimizing the product of variance and cost for fixed sample size.

The cost-variance optimization was carried out in two steps. We first consider the optimization of the sample designs followed in each of the two major area types identified in the LFS design; i.e., the SR Areas or major cities, and NSR Areas which are the smaller urban and rural areas. The scope of the optimization includes the allocation of sample to the two stages of the SR design (Section 2), and the consideration of alternatives to the old design in NSR areas (Section 3). For NSR areas the old design is first evaluated empirically via a components of variance approach, and one stage of sampling in rural areas is identified for elimination. Subsequently the modified old design is compared to an alternative design featuring explicit rural/urban stratification from an overall cost-variance perspective. For both types of areas variances are obtained empirically using data from the 1971 and 1976 Censuses, while cost models are developed using data from a time and cost study, and by means of a simulation study.

[1] G.H. Choudhry, H. Lee, and J.D. Drew, Census and Household Survey Methods Division, Statistics Canada, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario K1A 0T6.

In Section 4, we consider the second stage of optimization, the allocation of sample to NSR and SR areas, taking into account the design improvements identified for each type of area. Finally, Section 5 summarizes the improvements identified, and their implications on the redesigned sample.

2. SR DESIGN

The old SR design is a stratified two-stage design (Platek and Singh 1976). Each Self-Representing Unit (SRU) is stratified into a number of contiguous strata called subunits and each subunit is subdivided into clusters which are the primary sampling units (PSU's). The PSU's are selected using the random group method due to Rao, Hartley, and Cochran (1962) and at the second stage of sampling, a systematic sample of dwellings is taken in such a manner that the design becomes self-weighting. Let $1/W$ be the sampling rate in the stratum and n be the number of PSU's to be selected from the stratum. The N PSU's in the stratum are randomly partitioned into n groups so that the i-th random group contains N_i PSU's and $\sum_{i=1}^{n} N_i = N$. Let x_j and M_j, $j = 1, 2, \ldots, N$, respectively be the size measure and dwelling count for the j-th PSU in the stratum.

Define

$$\lambda_j = \frac{x_j}{\sum_{t=1}^{N} x_t}$$

and

$$\delta_{ij} = 1 \text{ if } j\text{-th PSU is in } i\text{-th group}$$
$$= 0 \text{ otherwise.}$$

Then $\pi_i = \sum_{j=1}^{n} \delta_{ij}\lambda_j$ is the relative size of the i-th group. Now define W_{ij}'s as

$$W_{ij} = \delta_{ij}\left[W\frac{\lambda_j}{\pi_i}\right] \text{ or } \delta_{ij}\left[W\frac{\lambda_j}{\pi_i} + 1\right] \tag{2.1}$$

such that $\sum_{j=1}^{N} W_{ij} = W$ for $i = 1, 2, \ldots, n$, where $[a]$ is the greatest integer less than or equal to a. Now select one PSU from each of the n random groups independently with probability proportional to W_{ij}'s and sub-sample the selected PSU j from the i-th group at the rate $1/W_{ij}$. Then the overall sampling rate within each of the random groups is $1/W$ so that the design becomes self-weighting with a design weight equal to W. The average sample size for the stratum is given by

$$m = \frac{1}{W}\sum_{j=1}^{N} M_j \tag{2.2}$$

$$= M_0/W$$

where M_0 is the total number of dwellings in the stratum. Let M_{ij} be the number of dwellings in the selected PSU j in the i-th group, then $m_i = M_{ij}/W_{ij}$ dwellings will be selected from the i-th group. The average number of dwellings selected from the i-th group for a given random grouping is $1/W \sum_j \delta_{ij} M_j$ and the average over all possible random groupings is m N_i/N since the expected value of δ_{ij} is N_i/N. If $N_i/N = 1/n$, i.e., the number of psu's in each of the random groups is the same, then the average sample per selected PSU is $m/n = d$(say), where d will be called the average density for the stratum. Since m is fixed, the sample of m dwellings can be elected by varying n and d such that the product *(nd)* remains equal to

m, the total sample size for the stratum. Our objective here is to obtain d which for a fixed sample size minimizes the product of variance and cost. For the optimization we obtain the total variance via the components of variance approach and consider a linear cost function as described in the following section.

2.1 Variance Function

Suppose that we are interested in the total of a characteristic y for the subunit. Let y_{jh} be the y-value for the h-th household in PSU j where $h = 1, 2, \ldots, N$, then the total $Y = \sum_{j=1}^{N} \sum_{h=1}^{M_j} y_{jh}$ is estimated by

$$\hat{Y} = W \sum_{i=1}^{n} y_i \tag{2.3}$$

where y_i is the sum of the y-values for the m_i selected households from the PSU selected from the i-th group, $i = 1, 2, \ldots, n$. Ignoring the effect due to rounding involved in defining W_{ij}, the variance of $\hat{Y}$ is given by (Rao et al. 1962)

$$\operatorname{Var}(\hat{Y}) = A \left[\sum_{j=1}^{N} \frac{Y_j^2}{\lambda_j} - Y^2 \right] + \sum_{j=1}^{N} M_j S_j^2 \left[W - 1 - A\left(\frac{1}{\lambda_j} - 1\right)\right]. \tag{2.4}$$

where
$$Y_j = \sum_{h=1}^{M_j} y_{jh},$$

$$S_j^2 = \frac{1}{M_j - 1} \sum_{h=1}^{M_j} \left(y_{jh} - \frac{Y_j}{M_j} \right)^2,$$

$$A = \frac{\sum_{1}^{n} N_i^2 - N}{N(N-1)}.$$

If $N_i = N/n$, i.e., all random groups have equal number of PSU's, then

$$A = \frac{N-n}{n(N-1)}.$$

Relative variance of $\hat{Y}$ defined by $\operatorname{Var}(\hat{Y})/Y^2$ will be

$$\begin{aligned}
\text{Rel. Var}(\hat{Y}) &= A \left[\frac{1}{Y^2} \sum_{j=1}^{N} \frac{Y_j^2}{\lambda_j} - 1 \right] + \frac{1}{Y^2} \sum_{j=1}^{N} M_j S_j^2 \left[W - 1 - A\left(\frac{1}{\lambda_j} - 1\right)\right]. \\
&= A\mu_1 + (W-1)\mu_2 + A\mu_2 - A\mu_3 \\
&= (W-1)\mu_2 + A(\mu_1 + \mu_2 - \mu_3)
\end{aligned} \tag{2.5}$$

where
$$\mu_1 = \frac{1}{Y^2} \sum_{j=1}^{N} \frac{Y_j^2}{\lambda_j} - 1$$

$$\mu_2 = \frac{1}{Y^2} \sum_{j=1}^{N} M_j S_j^2 \, ,$$

$$\mu_3 = \frac{1}{Y^2} \sum_j M_j \frac{S_j^2}{\lambda_j} \, .$$

μ_1, μ_2, and μ_3 are the population prameters and are fixed for a particular characteristic. Since $m = nd$ and if we assume that $N_i = N/n$ then we can write A as

$$A = \frac{1}{N-1} (N \frac{d}{m} - 1)$$

and

$$\text{Rel. Var}(\hat{Y}) = (W - 1)\, \mu_2 + (N \frac{d}{m} - 1) \frac{(\mu_1 + \mu_2 - \mu_3)}{(N - 1)}$$

$$= \alpha_0 + \alpha_1 d \qquad (2.6)$$

where

$$\alpha_0 = (W - 1)\, \mu_2 - \frac{(\mu_1 + \mu_2 - \mu_3)}{(N - 1)}$$

$$\alpha_1 = \frac{N}{m} \frac{(\mu_1 + \mu_2 - \mu_3)}{(N - 1)} \, .$$

From (2.6), we observe that from reliability point of view, the value $d = 1$ (i.e., one dwelling per PSU) is optimum. But this will have impact on the cost as discussed in the next section. The values of α_0 and α_1 for unemployed for Halifax SRU were obtained from 1981 census data and these are

$$\alpha_0 = 0.019005, \qquad \alpha_1 = 0.0007972.$$

Since α_1 is very small as compared to α_0, the increase in the variance with the corresponding increase in d will be very small. Next we examine the effect on the cost due to varying the value of the average density d.

2.2 Cost Model

A simple cost model has been considered to investigate the impact on the cost as the density is varied. Due to telephone interviewing in the SR areas, personal visits are only required to a PSU during the rotation month and in cases where some households were without a telephone or did not agree to telephone interviewing.

A breakdown of the interviewing cost by telephone and personal visit is available for individual interviewers from field operations, but further breakdown of the personal visit component of the cost was required to construct the cost model. For this purpose a special time and cost study was carried out in the field for a period of six months (February-July 1982) on a random sample of interviewers. The results from the analysis of time and cost data are documented in a report by Lemaitre (1983). For the purpose of our cost model, we define the following set of parameters

c_0 = Fixed costs
c_1 = Average cost of dwelling-to-dwelling travel within the same PSU
c_2 = Average cost of PSU-to-PSU travel
γ = Number of PSU-to-PSU moves per selectd PSU.

The fixed cost c_0 includes the time spent actually conducting interviews whether by telephone or in person and the travel cost from home to area and back. The fixed cost c_0 depends only on the total sample size m and not on n, the number of selected PSU's. Suppose that there are g_1 dwelling-to-dwelling moves and g_2 PSU-to-PSU moves made, then the total cost for m dwellings will be

$$T = c_0 + g_1 c_1 + g_2 c_2. \tag{2.7}$$

If n is increased then g_2 will also increase and g_1 will decrease and vice-versa but $(g_1 + g_2)$ should remain constant because the number of moves depends on the sample size m and the proportion of households interviewed by personal visit. Then we may write

$$g_1 + g_2 = \theta m. \tag{2.8}$$

From (2.8) we substitute g_1 in equation (2.7) and obtain

$$\begin{aligned} T &= c_0 + \theta m c_1 + g_2(c_2 - c_1) \\ &= c_0 + \theta m c_1 + n\gamma(c_2 - c_1). \end{aligned}$$

Now replacing n by m/d we have

$$T = c_0 + \theta m c_1 + \frac{m\gamma}{d}(c_2 - c_1)$$

and cost per dwelling C as a function of average ensity d is given by

$$C = \frac{c_0}{m} + \theta c_1 + \frac{\gamma}{d}(c_2 - c_1). \tag{2.9}$$

From Time and Cost Study the parameters c_1 and c_2 for Halifax were 0.78 and 2.51 respectively. These parameters were observed with average density equal to 5 but c_2 increases with d and c_1 decreases with d. Assuming that the average distance between the units is inversely proportional to the square root of the number of units in an area, we can replace c_1 by $c_1(5/d)^{1/2}$ and c_2 by $c_2(d/5)^{1/2}$ in our model so that the modified model becomes

$$C = \frac{c_0}{m} + \theta c_1\left(\frac{5}{d}\right)^{1/2} + \frac{\gamma}{d}\left\{c_2\left(\frac{d}{5}\right)^{1/2} - c_1\left(\frac{5}{d}\right)^{1/2}\right\}. \tag{2.10}$$

c_0/m is fixed per dwelling cost and does not depend on density and its value was 3.28 from Time and Cost Study. The parameter θ does not depend on the density either and was equal to 0.356 from Time and Cost Study. The parameter γ increases with density because the average number of visits to a PSU will increase due to higher density. We have approximated γ by

$$\frac{1}{6} + \frac{5}{6}(1 - p^d)$$

where p is the probability of telephone interview for a household in a non rotate-in PSU and the value of p was 0.85 as obtained from interviewers' data. From the cost model (2.10), the values of per dwelling cost for $d = 2, 3, \ldots, 10$ are given in Table 1 along with the relative variances and the products of these two which are the values of the objective function to be minimized.

Table 1

Value of Relative Variance, Cost per Dwellings and Objective Function for Various Densities (Unemployed)

Density	Relative Variance	Cost per Dwelling	Objective Function
2	0.0206	3.79	0.078
3	0.0214	3.79	0.081
4	0.0222	3.79	0.084
5	0.0230	3.78	0.087
6	0.0238	3.77	0.090
7	0.0246	3.76	0.092
8	0.0254	3.75	0.095
9	0.0262	3.74	0.098
10	0.0270	3.73	0.101

As expected, we observe that under the model considered here, the cost per dwelling decreases very slowly as the density increases since the fixed per dwelling cost (c_0/m) dominates in (2.10) due to telephone intervewing. From the previous section we had found that the increase in the relative variance is very small as the density increases. As a result our objective function is monotonically increasing but the loss in the cost-variance efficiency with increase in d is small. However it was decided to retain the old density of 5 for the redesigned sample on the grounds that lower density would have resulted in more selected PSU's with higher implementation and maintenance costs.

3. NSR DESIGN

3.1 NSR Design Alternatives

Design Alternative D_0: Old NSR Design (see Figure 1)

Key features of the old NSR design (Platek and Singh 1976) were:

i) **Stratification:** Economic Regions (ER's) whose numbers varied from 1-10 per province served as major strata. Within ER's, from 1-5 geographicaly contiguous strata were formed, using industry data from the 1971 Census.

ii) **Primary Sampling Units (PSU's):** These were delineated within strata, to be geographically compact areas similar to the stratum with respect to stratification variables, and with respect to the ratio of rural to urban population. PSU populations ranged from 3,000 to 5,000. In the first stage PSU's were selected following the randomized probability proportional to size systematic (RPPSS) method of Hartley and Rao (1962). Within PSU's urban and rural parts were sampled separately.

iii) **Within PSU Sampling: Urbans** All urban centers assigned in whole or in part to selected PSU's were included in the sample. The second stage of sampling was a sample of blocks, following the RPPSS method. The third and final stage of sampling was a systematic sample of dwellings.

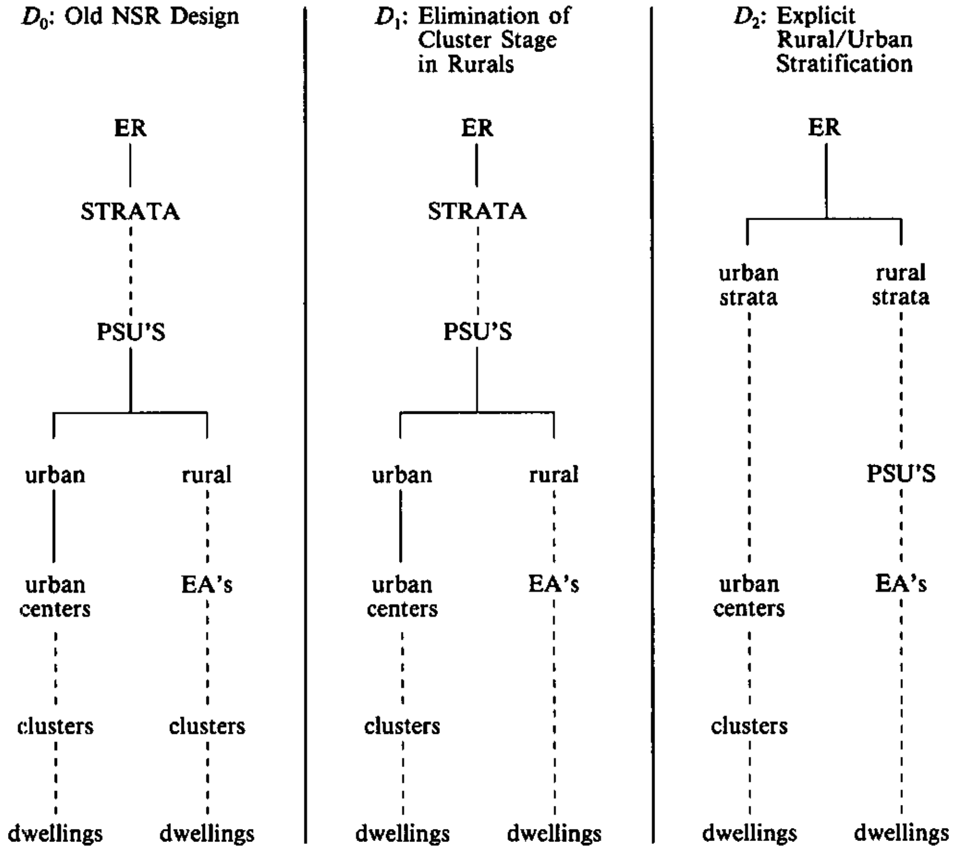


Figure 1. Representation of NSR Design Alternatives. (—— stratification, ----- stage of sampling)

iv) **Within PSU Sampling: Rurals** The second stage of sampling was a RPPSS sample of EA's. EA's were then field counted for the purposes of delineating clusters having from 3-20 dwellings. The third and fourth stages of sampling corresponded to an RPPSS sample of clusters and a systematic sample of dwellings.

Design Alternative D_1: Elimination of Cluster Stage of Sampling in Rurals

i) It would permit shortening of the lead time to select independent samples from the LFS frame to 7 months from 13 months, by eliminating the need for counting of EA's.

ii) Elimination of the clustering step would reduce sample maintenance costs.

iii) A priori, the reduction in the stages of sampling from 4 to 3 stages would translate into a reduced variance. it was expected that costs, on the other hand, would not be very much affected, particularly with the shift to telephone interviewing.

iv) At an early juncture in the redesign research program a field study was carried out on the operational implications of eliminating the cluster stage. Verification of EA listings a year later revealed no problems with the quality of listings, and analysis revealed no discernable impact on data collection costs.

Design Alternative D_2: Explicit Urban/Rural Stratification

The old design with its separate sampling of urban and rural portions of PSU's featured an implicit urban/rural stratification. A drawback of the approach however was that maintenance of the stratum urban to rural population ratio at the PSU level required frequent discontiguity between rural and urban portions of PSU's, leading in turn to increased travelling costs.

In view of this problem with the old design, design alternative D_2 was formulated as follows:

i) **Stratification:** Rural and urban portions of ER's would constitute primary strata, which would be optimally sub-stratified to the point of having strata yields of 100-150 dwellings (i.e., 2-3 PSU's each corresponding to an interviewer's assignment). ER's not able to support at least one such urban and one such rural stratum (roughly ⅓ of ER's) were considered ineligible for D_2.

   Secondary rural strata would be contiguous, while secondary urban strata would be formed without geographic constraints.

ii) **Sampling Within Rural Strata:** PSU's similar to the stratum with respect to stratification variables would be formed by grouping geographically contiguous EA's and will be selected by the RPPSS method. Second and third stages of sampling would be an RPPSS sample of EA's and systematic sample of dwellings.

iii) **Sampling Within Urban Strata:** Sampling would proceed in three stages as follows: RPPSS sample of PSU's (individual or combined urban centers), RPPSS sample of clusters, and systematic sample of dwellings.

3.2 Variance Components Model

Design alternative D_0, D_1 and D_2 were simulated using census data. Expressions for the variance components are given below:

Stage of Sampling	**Variance Expression**	
1st	$V_{(1)} = V^{\text{RPPSS}}_{(1)}$	(3.1)
2nd	$V_{(2)} = W \sum_{i=1}^{N} \frac{V^{\text{RPPSS}}_{(2)i}}{W_i}$	(3.2)
3rd	$V_{(3)} = W \sum_i \sum_j \frac{V^{\text{SRS}}_{(3)ij}}{W_{ij}}$ if last stage,	(3.3)
	$= W \sum_i \sum_j \frac{V^{\text{RPPSS}}_{(3)ij}}{W_{ij}}$ otherwise	
4th (where applicable)	$V_{(4)} = W \sum_i \sum_j \sum_k \frac{V^{\text{SRS}}_{(4)ijk}}{W_{ijk}}$	(3.4)

The variance formula and its computation method for the RPPSS sampling are described in Appendix A.

3.3 Cost Model

Whereas the cost model for the SR areas dealt with allocation of samples to 2 stages of sampling, here a cost model is needed to compare alternative NSR designs.

The cost model for design D_1 under personal interviewing was formulated as

$$C_{D_1} = F_0 + F_1 + F_2 + E_1 + E_2$$

where F_0 = fixed fee for interviewing,
F_1 = fee for home to area, between PSU, and between secondary travel,
F_2 = fee for within secondary (dwelling to dwelling) travel,
E_1 = expenses associated with home to area, between PSU, and between secondary travel,
E_2 = expenses associated with dwelling to dwelling travel.

Fees are compensation for the time spent and expenses for the distance covered. All Parameters are expressed in terms of per dwelling costs.

Under telephone interviewing, this was modified to

$$C_{D_1}^T = F_0 + \alpha(F_1 + F_2 + E_1 + E_2),$$

where α is the factor by which time and mileage would be decreased under telephoning.

Now, under the assumption that D_2 would affect F_1 and E_1, say by a factor r, but would not affect other components we have,

$$C_{D_2}^T = F_0 + \alpha r(F_1 + E_1) + \alpha(F_2 + E_2).$$

Parameters of $C_{D_1}^T$ and $C_{D_2}^T$ were estimated as follows:

F_0, F_1, F_2, E_1, E_2: These were estimated under D_0 from a special Time and Cost study (Lemaitre 1983), carried out as part of the redesign research program. Since the field test of D_1 revealed no discernable differences in data collection costs between D_0 and D_1, these parameters were assumed unchanged under D_1.

α: Field testing of telephone interviewing carried out as part of the redesign research program did not have as an objective the estimation of cost savings. An estimated 10% reduction in total data collection costs was made by Regional Operations staff, which permitted calculation of α.

r: This parameter could not be estimated based on available data, rather a Monte Carlo simulation study was needed, which is described in Appendix B.

3.4 Results of Cost-Variance Analyses

Variance Analysis: D_1 vs. D_0

Components of variance for 6 labour force characteristics were obtained for designs D_0 and D_1 using 1971 Census data for 5 ER's across Canada. Table 2 gives the % contribution from each stage of sampling to the total variance under D_0. It can be observed that 30-40% of the total variance under D_0 was due to the rural cluster (3rd) stage of sampling, and that under design D_1 20-30% variance reductions could be obtained.

Table 2

Percent Contributions to the Total Variance from Stages of Sampling for the Current Design and Percent Reduction in the Total Variance Due to Eliminating Cluster Stage of Sampling in Rural Areas; $100\ (1 - \frac{V_{D_1}}{V_{D_0}})$

Characteristic	Percent Contribution to Total Variance from						Percent Variance Reduction;
		Urban		Rural			$100\ (1 - \frac{V_{D_1}}{V_{D_0}})$
	1st stage	2nd stage	3rd stage	2nd stage	3rd stage	4th stage	
LF Population	14.5	12.9	10.8	5.8	40.5	15.5	30.5
Employed	21.2	11.2	10.4	6.3	35.0	15.8	27.1
Unemployed	12.6	15.8	16.6	4.8	33.0	17.2	24.8
Not in LF	24.7	11.9	10.7	4.8	32.9	15.1	22.9
Employed Agr.	42.4	1.0	0.8	12.3	30.8	12.6	20.4
Employed Non-Agr.	23.3	12.7	11.9	5.6	31.7	14.8	21.8

The gains might be less since for the study, the variables being estimated and the size measures referred to the same point in time whereas this would not be true in practice. No attempt was made to discount the gains, however, since the choice between D_1 and D_0 was clear both in terms of variances, and on operational grounds (as discussed in Subsection 3.1). Further efforts were devoted hence to the choice between D_1 and D_2.

Variance Analysis: D_2 vs. D_1

In this study the number of ER's was expanded to 11, and study variables (employed and unemployed) were based on the 1976 Census, whereas size measures were based on the 1971 Census. Also variances were computed with ratio estimation based on total population.

The average variance efficiency of D_2 with respect to D_1 was 1.16 for employed and 0.97 for unemployed (Table 4).

Cost Analysis: D_2 vs. D_1

Values of all the parameters in the cost model are presented in Table 3 along with $C^T_{D_1}$ and $C^T_{D_2}$ and their ratio.

As expected the between PSU and between secondary component of interviewer fees and expenses are higher under D_1 due to the frequent lack of contiguity between rural and urban portions of PSU's. The average reduction factor r in these components under D_2 was estimated as in Table 3 leading to an overall cost efficiency for D_2 vs. D_1 of 1.08 (Table 4).

Combined Cost Variance Analysis: D_2 vs. D_1

Table 4 gives the relative cost-variance efficiencies of D_2 vs. D_1 under telephone interviewing. In terms of overall efficiency, D_2 is 25% and 5% more efficient than D_1 for employed and unemployed respectively.

Based on these findings it was decided to adopt D_2 in the 2/3 of ER's capable of supporting both urban and rural strata, and design D_1 was adopted in the remaining cases.

Table 3

Values of Parameters in the NSR Cost Model and Relative Cost Efficiencies of D_1 vs. D_2 with Telephone Interviewing

ER	F_0	F_1	F_2	E_1	E_2	α	r	$C^T_{D_1}$	$C^T_{D_2}$	$C^T_{D_1}/C^T_{D_2}$
22	2.05	0.74	1.31	0.95	0.92	0.85	0.93	5.38	5.28	1.02
32	2.13	0.86	1.11	0.90	0.97	0.84	0.88	5.35	5.17	1.03
41	2.04	0.94	0.94	0.96	0.69	0.84	0.42	5.01	4.08	1.23
44	2.04	0.94	0.94	0.96	0.69	0.84	0.50	5.01	4.21	1.19
51	1.94	0.80	1.07	0.81	0.75	0.84	0.89	4.82	4.67	1.03
56	1.94	0.80	1.07	0.81	0.75	0.84	0.68	4.82	4.39	1.10
63	2.07	1.03	1.03	1.19	0.97	0.75	0.87	5.66	5.41	1.05
72	1.92	0.96	1.13	1.05	1.09	0.85	0.82	5.52	5.21	1.06
82	1.88	1.12	1.01	1.20	0.94	0.86	0.57	5.55	4.69	1.18
86	1.88	1.12	1.01	1.20	0.94	0.86	0.90	5.55	5.35	1.04
96	2.03	0.81	1.22	0.75	0.85	0.84	0.75	5.07	4.74	1.07

Table 4

Relative Cost-Variance Efficiencies of D_1 vs. D_2

	Variance Efficiency V_{D_1}/D_{D_2}		Cost Efficiency $C^T_{D_1}/C^T_{D_2}$	Relative Cost-Variance Efficiency $V_{D_1}C^T_{D_1}/V_{D_2}C^T_{D_2}$	
ER	Employed	Unemployed		Employed	Unemployed
22	1.09	0.93	1.02	1.11	0.95
32	0.91	0.72	1.03	0.94	0.74
41	1.14	0.86	1.23	1.40	1.06
44	1.39	1.14	1.19	1.65	1.37
51	0.96	1.01	1.03	0.99	1.04
56	1.12	1.51	1.10	1.23	1.66
63	1.35	1.06	1.05	1.41	1.11
72	1.00	0.91	1.06	1.06	0.96
82	1.09	1.01	1.18	1.27	1.19
86	1.20	1.05	1.04	1.25	1.09
96	1.38	1.05	1.07	1.48	1.12
All*	1.16	0.97	1.08	1.25	1.05

* Weighted average by population size.

3.5 Special 2-Stage Design for Prince Edward Island

For Canada's smallest province, Prince Edward Island, where sampling rates of 4% are required in order to produce reliable provincial data, design alternative D_3, a stratified sample of EA's and dwellings, was considered as an alternative to D_2.

D_3 did not feature any clustering of the sample into geographically contiguous primaries designed to correspond to interviewers assignments, as it was hypothesized that given the high sampling rates, the increase in data collection costs might be more than offset by variance reductions due to elimination of a stage of sampling, and due to stratification gains resulting from having more strata (i.e., up to 4 times as many as under D_2).

Cost-variance study results showed the variance efficiency of D_3 vs. D_1 to be 2.39 for employed and 1.20 for unemployed, while costs under D_3 were only 8% greater. Hence, based on overall cost-variance efficiencies of 2.21 for employed and 1.11 for unemployed, D_3 was opted for.

3.6 Number of PSU's Selected Per Stratum

Under both designs D_1 and D_2, the sample yield per PSU was fixed at 55-60 dwellings to correspond to an interviewer's assignment. In about half of the ER's, there was only enough sample for 2 or 3 PSU's to be selected. Further stratification in these cases was ruled out on the grounds that there should be at least 2 PSU's per stratum to permit unbiased estimation of variance.

For the remaining ER's, some consideration was given to having 4-5 PSU's per stratum, as this would permit greater flexibility to reduce the size of the area sample, for example, if a portion of the area sample at some time in the future were to be converted to a telephone sample under a dual frame set-up. However, stratification to the point of 2-3 PSU's per stratum was adopted, based on variance reductions of 14.8% for employed and 5.4% for unemployed for these ER's. A detailed description of the stratification procedures followed can be found in Drew, Bélanger, and Foy (1985).

4. COST-VARIANCE OPTIMIZATION BETWEEN SR and NSR AREAS

The next step in the cost-variance optimization of the LFS design was the optimization of the allocation of sample between SR and NSR areas. We used the simple cost and variance models considered by Fellegi, Gray, and Platek, (1967), i.e.,

cost: $$C = \sum_{j=1}^{2} C_j \frac{P_j}{W_j}, \tag{4.1}$$

variance: $$V = \sum_{j=1}^{2} W_j P_j \sigma_j^2, \tag{4.2}$$

where j = area type (= 1 for SR; = 2 for NSR),
C_j = unit (i.e., per person) cost,
P_j = population,
$1/W_j$ = sampling rate,
σ_j^2 = unit variance.

Fellegi et al. showed that if C is minimized with V fixed the ratio of the sampling rates is

$$\frac{W_1}{W_2} = \frac{\sigma_2}{\sigma_1}\left(\frac{C_1}{C_2}\right)^{1/2} \tag{4.3}$$

The other optimization criteria described in Section 1 also give the same ratio as above.

Parameters were estimated as follows:

(i) **Unit costs:** Historical per dwelling costs by type of area were available. These were decreased by 10% for NSR areas, to take account of the estimated effect of a shift to telephone interviewing of all rotation groups except the rotate-in group for the redesigned sample.

(ii) **Unit variances:** Optimization was carried out with respect to the characteristic unemployed, for which variances were given by:

$$\sigma_j^2 = \beta_j \frac{u_j}{P_j}\left(1 - \frac{u_j}{P_j}\right); j = 1, 2 \tag{4.4}$$

where β_j = design effect for unemployed, and u_j = unemployed.

Historical design effects by type of area were available, and were reduced to take into account of structural improvements in the respective NSR and SR designs as described in Sections 2 and 3. Unemployment levels were based on 1980-82 average LFS data, which seemed appropriate in light of medium term forecasts which were not calling for a return to pre-1982 recession levels of unemployment, and population counts were based on the 1981 Census.

Table 5 presents the percent of sample in SR areas under the following allocations: (i) old design, (ii) proportional allocation, (iii) optimum allocation under the assumed cost and variance model, and (iv) the allocation adopted for the redesigned sample. The optimum allocation could not be adopted because of subprovincial data reliability constraints. In most cases, the differences between the optimum allocation and the one adopted are small. The optimal allocation turned out to be quite close to proportional, and quite different from the allocation under the old design.

Table 5

Percent of Sample in SR Areas within Provinces for (1) Old Sample, (2) Proportional Allocation, (3) Optimum Allocation, and (4) Redesigned Sample

Province	Old Sample	Proportional Allocation	Optimum Allocation	Redesigned Sample
Newfoundland	41.8	51.3	42.6	44.6
Prince Edward Island	26.6	32.8	32.8	28.9
Nova Scotia	37.3	57.4	58.8	51.9
New Brunswick	49.5	52.5	47.4	53.6
Quebec	56.8	74.8	71.6	68.9
Ontario	62.5	79.1	78.8	75.0
Manitoba	54.1	71.0	76.4	56.4
Saskatchewan	44.7	51.8	62.1	56.8
Alberta	60.0	68.6	72.6	62.3
British Columbia	58.0	78.0	74.6	69.7
Canada	53.2	67.1	67.4	62.3

Table 6

Relative Efficiency of the Redesigned Sample Allocation with Respect to the Old by Province (Unemployed)

Province	Cost Ratio $(= \frac{C^{(O)}}{C^{(N)}})$	Variance Ratio $(= \frac{V^{(O)}}{V^{(N)}})$	Rel. Eff. $(= \frac{C^{(O)}V^{(O)}}{C^{(N)}V^{(N)}})$
Newfoundland	1.00	1.00	1.00
Prince Edward Island	1.01	1.02	1.03
Nova Scotia	1.04	1.14	1.18
New Brunswick	1.01	0.98	0.99
Quebec	1.03	1.06	1.09
Ontario	1.04	1.08	1.12
Manitoba	1.01	1.03	1.04
Saskatchewan	1.05	1.06	1.12
Alberta	1.01	1.01	1.02
British Columbia	1.02	1.09	1.11
Canada	1.03	1.07	1.10

The projected gains resulting solely from the re-allocation process under the assumption of fixed (old) provincial sample sizes and uniform sampling rates within the two area types are presented in Table 6. For this table, the unit costs and variances described above were used in determining the total costs and variances, $C^{(O)}$, $C^{(N)}$, $V^{(O)}$, $V^{(N)}$, under the old and new allocations respectively. The new allocation would have resulted in a 3% decrease in total cost and a 7% decrease in total variance of unemployed and for a combined relative efficiency (as defined in Table 6) of 1.10. Had it not been for the subprovincial data requirements, an efficiency gain of 1.12 could have been achieved under the optimal allocation.

The actual efficiency gains for the redesigned sample vs. the old sample are considered in the following section.

5. CONCLUSIONS

The changes in the LFS design taken as a result of the cost-variance studies are the following: elimination of a stage of sampling in NSR rural areas, adoption of a design featuring rural/urban stratification, adoption of a 2-stage NSR design in Prince Edward Island, increase in the number of NSR strata to the extent that only 2 or 3 PSU's per stratum will be selected, and re-optimization of the allocation of sample between NSR and SR areas. The near optimality of other design parameters established earlier by Fellegi, Gray and Platek (1967) was found to have remained unchanged, for example the number of dwellings to select per PSU in SR Areas.

The efficiency gains resulting from the changes permitted a 7% reduction in the overall LFS sample size and achieved the required reliability of subprovincial data (Singh et al. 1984) without impacting on the reliability of provincial and national estimates. The only exceptions were the provinces of Quebec and Manitoba, where greater subprovincial data demands

Table 7
Relative Efficiency of the Redesigned
vs. the Old Sample for Unemployed

Province	Cost Ratio* $(= \frac{C^{(O)}}{C^{(N)}})$	Variance Ratio $(= \frac{V^{(O)}}{V^{(N)}})$	Rel. Eff. $(= \frac{C^{(O)}V^{(O)}}{C^{(N)}V^{(N)}})$
Newfoundland	1.19	1.00	1.19
Prince Edward Island	1.10	1.13	1.24
Nova Scotia	1.22	1.04	1.27
New Brunswick	1.17	0.99	1.16
Quebec	1.15	0.95	1.09
Ontario	1.13	1.03	1.16
Manitoba	1.17	0.96	1.12
Saskatchewan	1.23	1.02	1.25
Alberta**	1.15	1.00	1.15
British Columbia	1.15	1.01	1.16
Canada	1.17	0.99	1.16

* Based on the redesigned sample with telephone interviewing and the old sample with personal visit interviewing in NSR areas.

** Supplementary sample not included.

necessitated a slight loss in provincial data reliability. Table 7 gives the cost, variance and combined cost-variance ratios for the old sample (old design with 55,500 hhlds/month and no telephone interviewing in NSR's) vs. the redesigned sample (new design with 51,600 hhlds/month and telephone interviewing). The significant cost reductions are due to the shift to telephone interviewing in months 2-6 in NSR areas, and the sample size reduction. The overall cost-variance efficiency of the redesigned sample relative to the old sample was 1.16 (Table 7).

APPENDIX A

Variance Formula and Computation Method for RPPSS Sampling

Suppose that a sample of size n is selected by the randomized PPS systematic sampling from N units. Let p_i be the normalized size measure of the i-th unit such that $\sum_{i=1}^{N} p_i = 1$. The Horvitz-Thomson estimator of the total Y for a characteristics y is given by (Horvitz and Thomson 1952):

$$\hat{Y}_{HT} = \sum_{i \epsilon S} \frac{y_i}{\pi_i},$$

Where S = the selected sample of size n

y_i = y-values of i-th unit

π_i = np_i, the probability that the i-th unit is in S.

and its variance is

$$V(\hat{Y}_{HT}) = \sum_{i=1}^{N} \sum_{i<j} (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 ,$$

where π_{ij} is the joint probability that both the i-th and j-th units are in S. Hartley and Rao (1962) gave an asymptotic formula for π_{ij}'s.

An exact formula by Connor (1966) is also available but quite involved. Recently Hidiroglou and Gray (1980) developed a computer algorithm using a modification of Connor's formula due to Gray (1971), which was used in our study and compared with the Hartley-Rao approximation. It was found that the Hartley-Rao approximations are very close to the exact values for $N \geq 16$. We decided to use the Hidiroglou-Gray algorithm for $N < 16$ and the Hartley-Rao approximation for $N \geq 16$ considering exponential increase in computation with the algorithm as N increases.

APPENDIX B

Cost Simulation of D_2 vs. D_1

In order to estimate r, the ratio of fees and expenses for travel from home to area, between PSU's, and between secondaries under NSR design alternatives D_2 and D_1, a Monte Carlo study was carried out. The sample frames under D_1 and D_2 were simulated to the level of secondaries using Census data for each of the 11 study ER's. Fifty samples were drawn following each design, and the selected secondaries for each sample were grouped into geographically optimal assignments. If $\bar{M}^{(1)}$ and $\bar{M}^{(2)}$ are the average measures of within assignment geographic dispersion under designs D_1 and D_2, then r was estimated by

$$\bar{M}^{(2)}/\bar{M}^{(1)} .$$

The M-measure for a given sample was defined in the following manner. Suppose that k interviewers cover an ER and $G_i = \{U_{ij}; j = 1, 2, \ldots, n_i\}$ is the i-th interviewer's assignment, with n_i second stage sampling units. Let (x_{ij}, y_{ij}) be the population centroid of U_{ij} defined in Euclidean coordinates. The M-measure for the ER is defined as

$$M = \sum_{i=1}^{k} M_i ,$$

$$M_i = \sum_{j=1}^{n_i} \{(x_{ij} - \bar{x}_i)^2 + (y_{ij} - \bar{y}_i)^2\}^{1/2} ,$$

where $(\bar{x}_i, \bar{y}_i)$ is the center of G_i, i.e., $\bar{x}_i = 1/n_i \sum_{i=1}^{n_i} x_{ij}$; $\bar{y}_i = 1/n_i \sum_{j=1}^{n_i} y_{ij}$.

The determination of optimum interviewer assignments, that is the minimization of the M-measure, reduces to a classification or clustering problem. The following clustering algorithms were investigated:

i) **Friedman-Rubin (1967) Transfer Algorithm**

This non-hierarchical algorithm which was adopted for stratification of the LFS sample (Drew et al. 1985), starts with a random partitioning of units and proceeds towards a local optimum by moving one unit at a time from one cluster to another if the move reduces M. It also checks that size constraints are not violated before moving a unit. An approximation to the global optimum is achieved by taking several initial random starts. A disadvantage of the Friedman-Rubin algorithm in this case was that the strict size constraints required in order to have approximately equi-sized assignments, restricted the movement of units between clusters.

ii) Dahmström-Hagnell (1975) Exchange Algorithm

This algorithm is similar to the Friedman-Rubin algorithm, except that it is based on exchanging pairs of units between clusters as opposed to transfering individual units. Hence it works better under strict size constraints.

iii) Combined Algorithms

Define a cycle of a combined algorithm as application of the exchange algorithm, followed by the transfer algorithm. Then we considered both single and two cycle combined algorithms.

The combined two cycle algorithm worked best, requiring the smallest number of random starts and the least computing cost to achieve the same level of optimality as the other algorithms. Performance of the 1 and 2 cycle combined algorithms based on 21 replicates is summarized below.

	One Cycle				Two Cycle		
	No. of Random Starts				No. of Ramdon Starts		
	1	2	4	10	1	2	4
M-measure*	336.18	329.19	325.65	325.51	327.55	325.69	325.51
Standard Deviation	15.84	15.45	15.67	15.69	16.10	15.67	15.69
Computing Cost ($)	5.94	11.24	21.67	53.90	8.17	15.12	29.38

* Average over 21 replicates.

REFERENCES

CONNOR, W.S. (1966). An exact formula for the probability that two specified sampling units will occur in a sample drawn with unequal probabilities and without replacement. *Journal of the American Statistical Association*, 61, 384-390.

DAHMSTRÖM, P., and HAGNELL, M. (1975). Multivariate stratification of primary sampling units in multi-stage sampling with an application to SCB's general purpose sample. Research Report, University of Lund.

DREW, J.D., BÉLANGER, Y., FOY, P. (1985). Multivariate clustering algorithm for stratification and its application to the Canadian Labour Force Survey. Technical Report, Census and Household Survey Methods Division, Statistics Canada (in preparation).

FELLEGI, I.P., GRAY, G.B., and PLATEK, R. (1967). The new design of the Canadian Labour Force Survey. *Journal of the American Statistical Association*, 62, 421-453.

FRIEDMAN, H.P., and RUBIN, J. (1967). On some invariant criteria for grouping data. *Journal of the American Statistical Association*, 62, 1159-1178.

GRAY, G.B. (1971). Joint probability of selection of units in systematic samples. *Proceedings of American Statistical Association*, 271-276.

HARTLEY, H.O., and RAO, J.N.K. (1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics*, 33, 350-374.

HIDIROGLOU, M.A., and GRAY, G.B. (1980). Construction of joint probability of selection for systematic PPS sampling. *Journal of Royal Statistical Society*, C29, 107-112.

HORVITZ, D.G., and THOMSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of American Statistical Association*, 47, 663-685.

LEMAITRE, G. (1983). Some results from Time and Cost Study. Technical Report, Census and Household Survey Methods Division, Statistics Canada.

PLATEK, R., and SINGH, M.P. (1976). *Methodology of the Canadian Labour Force Survey*. Catalogue No. 71-526, Statistics Canada.

RAO, J.N.K., HARTLEY, H.O., and COCHRAN, W.G. (1962). A simple procedure of unequal probability sampling without replacement. *Journal of Royal Statistical Society*, B24, 482-491.

SINGH, M.P., and DREW, J.D. (1981). Research plans for the redesign of the Canadian Labour Force Survey. *Proceedings of the Section of Survey Research Methods, American Statistical Association Meetings*.

SINGH, M.P., DREW, J.D., and CHOUDHRY, G.H. (1984). Post '81 Censal redesign of the Canadian Labour Force Survey. *Survey Methodology*, 10, 127-140.

Survey Methodology, June 1985
Vol. 11, No. 1, pp. 51–64
Statistics Canada

Performance of ARIMA Models in Time Series[1]

KIM CHIU, JOHN HIGGINSON, and GUY HUOT[2]

ABSTRACT

This study is mainly concerned with an evaluation of the forecasting performance of a set of the most often applied ARIMA models. These models were fitted to a sample of two hundred seasonal time series chosen from eleven sectors of the Canadian economy. The performance of the models was judged according to eight variable criteria, namely: average forecast error for the last three years, the chi-square statistic for the randomness of the residuals, the presence of small parameters, overdifferencing, underdifferencing, correlation between the parameters, stationarity and invertibility. Overall and conditional rankings of the models are obtained and graphs are presented.

KEY WORDS: X11-ARIMA; Ranking; Priority; Criteria

1. INTRODUCTION

Our socio-economic environment is unstable and uncertain; inflation, recessions, and increasing pollution are among the factors contributing to increasing instability. We try to resolve the problem by using a method of forecasting that permits us to evaluate the impact of the frequent changes. ARIMA models (Box - Jenkins, 1970) are flexible enough to deal with such frequent changes in time series.

The purpose of this paper is to study a set of eight criteria which when applied to the Box-Jenkins method permit an evaluation of the fitting and forecasting performance of a set of the most often applied ARIMA models to Canadian economic time series. The question of which models perform well is important for programs like the X-11-ARIMA (Dagum 1980) which automatically fits a fixed small set of models (three models in the case of the X-11-ARIMA) to the series.

Section 2 introduces eight criteria: the average forecast error for the last three years, the chi-square statistic for the randomness of the residuals, the presence of small parameters, overdifferencing, underdifferencing, correlation between the parameters, stationarity and invertibility. Section 3 discusses the criteria and summarizes the results. Section 4 ranks the models conditionally and unconditionally. Section 5 compares within-sample and out-of-sample extrapolated values for the last three years.

2. THE CRITERIA

In this section we give a brief discussion of the eight criteria used in ranking the models.

[1] Presented at (1) Business and Economic Forecasting Session of the Canadian Operational Research Symposium, Ottawa, May 1984 and (2) Business and Economic Statistics Section of the American Statistical Association Meetings, Philadelphia, August 1984.

[2] K. Chiu, J. Higginson, and G. Huot, Time Series Research and Analysis Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

Stability

The stability condition of a process Z_t is either "stationary" or "non-stationary". It indicates how well the system remembers the shocks a_{t-j}, $j = 1, 2, \ldots,$ and how fast or slowly the response of the system to any particular shock decays. For a process

$$\begin{aligned} Z_t &= a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \ldots \\ &= \psi(B)a_t, \end{aligned}$$

where $a_t \sim NID(0, \sigma_a^2)$, the filter is said to be stable if the sequence $\{\psi_i\}$ is convergent. For a general ARIMA model (p, d, q),

$$\phi(B)\,(1 - B)^d Z_t = \theta(B)a_t,$$

the stability condition is that all the λ_i of the characteristic equation

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p = (1 - \lambda_1 B)\,(1 - \lambda_2 B) \ldots (1 - \lambda_p B) = 0$$

for the process are strictly inside the unit circle, i.e. $|\lambda_j| < 1$.

Invertibility

The process Z_t may be expressed as:

$$Z_t = a_t + \pi_1 Z_{t-1} + \pi_2 Z_{t-2} + \ldots$$

The system is said to be invertible if the sequence $\{\pi_i\}$ is convergent. The criterion is considered to be of primary importance because if the invertibility condition fails, the generating function $\pi(B)$ of the π's increases without bound. This means the current event of the system depends more on events in the distant past than in the recent past, and the process is physically meaningless.

The invertibility condition for a general ARIMA model (p, d, q), is that the ν_i of the characteristic equation

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q = (1 - \nu_1 B)\,(1 - \nu_2 B) \ldots (1 - \nu_q B) = 0$$

for the process are strictly within the unit circle, i.e. $|\nu_i| < 1$.

Underdifferencing

In the AR(p) model, when one or more of the λ_i, say λ_k approaches 1; then from

$$\begin{aligned} \phi(B) &= 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p \\ &= (1 - \lambda_1 B) \ldots (1 - \lambda_{k-1} B)\,(1 - \lambda_k B) \ldots (1 - \lambda_p B) \\ &= (1 - \lambda_1 B) \ldots (1 - \lambda_{k-1} B)\,(1 - \lambda_{k+1} B) \ldots (1 - \lambda_p B)\,(1 - \lambda_k B), \end{aligned}$$

we have $\phi(B)$ approaching

$$(1 - \phi_1' B - \phi_2' B^2 - \ldots - \phi_{p-1}' B^{p-1})\,(1 - B).$$

Therefore, a differencing operator may be needed for this system, and the AR(p) model becomes an ARI(p − 1, 1) model. Furthermore, when λ_k approaches 1, we may have non-stationarity.

Overdifferencing

Consider the general ARIMA model (p, d, q) (P, D, Q)$_s$,

$$\phi(B)\Phi(B)\,(1 - B)^d(1 - B^s)^D Z_t = \theta(B)\Theta(B)a_t.$$

If any ν_i of the characteristic equation $\theta(B) = 0$ approach 1, i.e. if any $(1 - \nu_i B)$ approach $(1 - B)$, we can eliminate $(1 - B)$ from both sides.

Test of randomness for the a_t's

Correlation in the residuals is not desirable since we want an unbiased estimate of the parameters for the process.

The statistic

$$Q = n(n + 2) \sum_{k=1}^{m} (n - k)^{-1} \varrho_k^2$$

as modified by Prothero and Wallis (1976) and Ljung and Box (1978) from the Chi-square test of Box and Pierce is used.

Here n is the sample size, $k = 1, 2, \ldots, m$ are the various lags, and ϱ_k are the autocorrelations. Q is used for the testing of the randomness of the residuals.

Small Parameters

Generally speaking, when the number of parameters of a given model is increased, the mean sum of squares σ_a^2 is reduced. However, only large parameters, or those parameters significantly different from 0 can contribute to a significant reduction of σ_a^2. To check for a small parameter, we may need an F-test (Pandit and Wu 1983):

$$F = \frac{A_1 - A_0}{s} \div \frac{A_0}{N - r} \sim F(s, N - r)$$

where r is the number of parameters of the model and s is the number of parameters which are restricted to zero. N is the number of observations, A_0 is the smaller sum of squares of the restricted model, and A_1 is the larger sum of squares of the restricted model.

But in our study here, we choose two constants, 0.05 and 0.10, as our indicator of the presence of a small parameter.

Correlation of the Parameters

High positive or negative correlation between parameters reflects ambiguity in the estimated values since a range of parameter values results in models with equally good fit. Therefore, if some of the elements in the correlation matrix of estimated parameters are large in absolute value, say greater than or equal to 0.9, the model may be reduced by deleting some of the smaller parameters.

Forecasting Error

No matter how we define a good model or bad model, we still have a primary interest in the forecasting error of the model. In this paper we use the mean absolute percentage forecasting error of one-year-ahead forecast

$$MAPE = \frac{1}{N}\sum_{t=1}^{N} \left| \frac{Z_{t+\ell} - \hat{Z}_t(\ell)}{Z_{t+\ell}} \right| \times 100\%$$

where ℓ is 12 or 4, and $\hat{Z}_t(\ell)$ is the forecast with lead time ℓ.

3. EVALUATION OF THE ARIMA MODELS

The eight criteria have been put into two groups. The first group considers good fitting of parsimonious models while the second considers the quality of the forecasts. This distinction between fitting and forecasting is important; good fitting and good forecasting are not equivalent.

These criteria have been used to evaluate and rank seven of the most often applied ARIMA models, namely:

1. $(0, 1, 1)\,(0, 1, 1)_s$
2. $(0, 1, 2)\,(0, 1, 1)_s$
3. $(0, 2, 2)\,(0, 1, 1)_s$
4. $(2, 1, 2)\,(0, 1, 1)_s$
5. $(1, 1, 0)\,(0, 1, 1)_s$
6. $(2, 1, 0)\,(0, 1, 1)_s$
7. $(2, 1, 0)\,(0, 1, 2)_s$

where "*s*" is 12 if the series is monthly and 4 if it is quarterly.

These models were fitted to a sample of 167 monthly seasonal time series chosen randomly from eleven sectors of the Canadian economy: national accounts; labour; prices; manufacturing; fuel, power and mining; construction; food and agriculture; domestic trade; external trade; transportation; and finance. About 40 quarterly time series from national accounts and finance were also tested.

The series are mostly multiplicative, according to the Bell Canada model test (Higginson 1976). That is, the different components (trend-cycle, seasonal, and irregular) are multiplied together to produce the raw series. Therefore, the amplitudes of the seasonal component frequently increase with increasing levels of the trend. The multiplicative series received a logarithmic transformation before the first three and last three models were fitted. The fourth model was fitted to the untransformed series in all cases.

Looking at the non-seasonal part of an ARIMA model which is associated with the trend-cycle and extremes, we see that the models can be grouped into three classes. Class I is models 1, 2 and 3 whose ordinary part includes only one or two first differences and one or two moving average parameters. Class III includes models 5, 6 and 7 whose ordinary part includes only one first difference and some autoregressive parameters. Model 4 (Class II) forms a class by itself; its non-seasonal part is mixed. We see that the seasonal part of all models is the same except for model 7.

Although the eight criteria are analysed separately in this section, several of them are dependent. For example, we shall see that the excess of parameters in model 4 generates problems of nonstationarity, noninvertibility, under- and overdifferencing, and correlation.

In Sections 3 and 4, we test within-sample extrapolated values for the seven ARIMA models. That is, the models are fitted to the whole series thus providing the parameters to be used for calculating the forecasts for the last three years. This is the way ARIMA forecasts are evaluated in the X-11-ARIMA program.

3.1 Criteria for Fitting Parsimonious ARIMA Models

The stationarity condition requires that all the roots of the autoregressive characteristic equation be inside the unit circle. We see in Table 1 that non-stationarity occurs only for model 4, in three cases. These appear to be due to overparametrization of the model.

In order for the model to be invertible, it is necessary that the roots of the moving average characteristic equation be inside the unit circle. Only model 4 has many cases of noninvertibility, 20%, as we see in Table 2. Two explanations are possible. There is first of all the case of straightforward noninvertibility. In some other cases noninvertibility was accompanied by nonstationarity. The fact that the autoregressive part may have roots near unity might have caused autocorrelation in the residuals. The moving average parameters would then take higher values to compensate.

An important criterion in judging the appropriateness of the ARIMA models for the series is the chi-square test of Box and Pierce (1970) (modified by Prothero and Wallis in 1976, and by Ljung and Box in 1978), applied to the autocorrelation of the residuals. Table 3 shows for each of the seven models the number and the percentage of series that fail the chi-square test at different levels. We see from this table first, that within a given class of models the simpler models have higher failure rates and second, that the failure rate depends to a large degree on the class of the model. The first point is illustrated by models 2 and 6 which having one more parameter than models 1 and 5, have a higher number of series passing this test. The evidence for the second point is that moving average models appear to satisfy the

Table 1

Failure in Stationarity

	CLASS I						CLASS II		CLASS III					
CRITICAL VALUE	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6		Model 7	
	(0, 1, 1) (0, 1, 1)		(0, 1, 2) (0, 1, 1)		(0, 2, 2) (0, 1, 1)		(2, 1, 2) (0, 1, 1)		(1, 1, 0) (0, 1, 1)		(2, 1, 0) (0, 1, 1)		(2, 1, 0) (0, 1, 2)	
--	--	--	--	--	--	--	3	2%	--	--	--	--	--	--

Table 2

Failure in Invertibility

	CLASS I						CLASS II		CLASS III					
CRITICAL VALUE	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6		Model 7	
	(0, 1, 1) (0, 1, 1)		(0, 1, 2) (0, 1, 1)		(0, 2, 2) (0, 1, 1)		(2, 1, 2) (0, 1, 1)		(1, 1, 0) (0, 1, 1)		(2, 1, 0) (0, 1, 1)		(2, 1, 0) (0, 1, 2)	
--	1	1%	2	1%	3	2%	33	20%	2	1%	2	1%	1	1%

Table 3

Failure in Chi-Square

CRITICAL VALUE	CLASS I						CLASS II		CLASS III					
	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6		Model 7	
	(0, 1, 1) (0, 1, 1)		(0, 1, 2) (0, 1, 1)		(0, 2, 2) (0, 1, 1)		(2, 1, 2) (0, 1, 1)		(1, 1, 0) (0, 1, 1)		(2, 1, 0) (0, 1, 1)		(2, 1, 0) (0, 1, 2)	
1%	31	19%	18	11%	29	17%	26	16%	62	37%	21	13%	20	12%
5%	45	27%	36	22%	46	28%	41	25%	82	49%	49	29%	42	25%
10%	61	37%	48	29%	56	34%	55	33%	89	53%	60	36%	56	34%
15%	72	43%	57	34%	69	41%	66	40%	101	60%	71	43%	64	38%
20%	83	50%	62	37%	80	48%	76	46%	106	64%	80	48%	73	44%
30%	100	60%	77	46%	94	56%	88	53%	119	71%	95	57%	89	53%
40%	111	66%	97	58%	107	64%	99	59%	127	76%	104	62%	100	60%
50%	121	72%	106	63%	118	71%	113	68%	135	81%	117	70%	116	69%
60%	131	78%	121	72%	128	77%	129	77%	141	84%	127	76%	121	72%

chi-square test better than autoregressive models. This may be due to the presence of extremes in the series. At the 5% level for example, model 1 fails for 27% of the series compared with 49% for its autoregressive counterpart model 5. As well as all models of class III, the mixed model, class II, is inferior to the second model of class I.

Underdifferencing occurs when a root of the characteristic equation of the autoregression polynomial is close to unity, say a distance ξ from unity. Here ξ is set equal to 0.1. We see in Table 4 that only model 4 is underdifferenced. This may be attributed to overparametrization. Model 4 has two autoregressive parameters and two moving average parameters in its non-seasonal part. Just through the estimation, there is a moderate chance that at least one of the autoregressive parameters will be greater than or equal to 0.9.

In this discussion the critical levels chosen for overdifferencing are 0.90 and 0.95. Table 5 shows that models 3 and 4 are most often overdifferenced. Model 3 has two first differences and two non-seasonal moving average parameters. If the second first difference is not necessary, autocorrelation is created in the series that has been differenced once already. The moving average polynomial will model this introduced autocorrelation by having one of its roots close to unity. We can therefore simplify the model by eliminating one moving average parameter and one difference. As to model 4, this may be due to overparametrization.

Table 4

Failure in Underdifferencing

CRITICAL VALUE	CLASS I						CLASS II		CLASS III					
	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6		Model 7	
	(0, 1, 1) (0, 1, 1)		(0, 1, 2) (0, 1, 1)		(0, 2, 2) (0, 1, 1)		(2, 1, 2) (0, 1, 1)		(1, 1, 0) (0, 1, 1)		(2, 1, 0) (0, 1, 1)		(2, 1, 0) (0, 1, 2)	
.90	--	--	--	--	--	--	14	8%	--	--	--	--	--	--

In ARIMA modelling of a stochastic process, it is enough to consider the first two moments, that is, the mean and autocovariance. The test on the size of the parameters serves only to eliminate those that contribute very little or nothing to the explanation of the autocovariance.

Table 6 illustrates two things. First, the simplest models pass this test better than more complicated models. After a logarithmic transformation, most of the multiplicative series in the sample will follow a straight line fairly closely (except for seasonal variation), so a "first difference" model will fit them using few parameters. Adding an extra unnecessary parameter to the model will often result in its receiving a small estimate from the estimation. Second, the estimated values of the moving average parameters are small (less than .05 or .10) more often than the estimated values of the autoregressive parameters. For example at the level of 0.05, the second autoregressive parameter in model 6 is judged unnecessary 13% of the time compared with 29% of the time for the second moving average parameter in model 2. Similarly, the addition of a second seasonal moving average parameter increased the failure rate from 13% in model 6 to 43% in model 7.

Table 5

Failure in Overdifferencing

	CLASS I						CLASS II		CLASS III					
CRITICAL VALUE	Model 1 (0, 1, 1) (0, 1, 1)		Model 2 (0, 1, 2) (0, 1, 1)		Model 3 (0, 2, 2) (0, 1, 1)		Model 4 (2, 1, 2) (0, 1, 1)		Model 5 (1, 1, 0) (0, 1, 1)		Model 6 (2, 1, 0) (0, 1, 1)		Model 7 (2, 1, 0) (0, 1, 2)	
.90	8	5%	11	7%	43	26%	50	30%	7	4%	9	5%	14	8%
.95	3	2%	6	4%	19	11%	37	22%	3	2%	3	2%	6	4%

Table 6

Failure in Small Parameter

	CLASS I						CLASS II		CLASS III					
CRITICAL VALUE	Model 1 (0, 1, 1) (0, 1, 1)		Model 2 (0, 1, 2) (0, 1, 1)		Model 3 (0, 2, 2) (0, 1, 1)		Model 4 (2, 1, 2) (0, 1, 1)		Model 5 (1, 1, 0) (0, 1, 1)		Model 6 (2, 1, 0) (0, 1, 1)		Model 7 (2, 1, 0) (0, 1, 2)	
.05	15	9%	49	29%	21	13%	42	25%	12	7%	22	13%	72	43%
.10	26	16%	88	53%	43	26%	73	44%	31	19%	45	28%	114	68%

Table 7

Failure in Correlation

	CLASS I						CLASS II		CLASS III					
CRITICAL VALUE	Model 1 (0, 1, 1) (0, 1, 1)		Model 2 (0, 1, 2) (0, 1, 1)		Model 3 (0, 2, 2) (0, 1, 1)		Model 4 (2, 1, 2) (0, 1, 1)		Model 5 (1, 1, 0) (0, 1, 1)		Model 6 (2, 1, 0) (0, 1, 1)		Model 7 (2, 1, 0) (0, 1, 2)	
--	--	--	3	2%	86	51%	124	74%	--	--	--	--	--	--

High positive or negative correlations between parameter estimates are undesirable and reflect ambiguity in the estimation situation since a range of parameter combinations result in models with equally good fits. Table 7 shows that only models 2, 3 and 4 fail the correlation test, i.e. the absolute value of at least one of the correlations is ≥ 0.90. The problem is minimal for model 2, and serious for models 3 and 4 where 51% and 74% of the fits had highly correlated parameters. This may be due to overdifferencing in model 3 and the presence of too many parameters in model 4.

3.2 Criterion for Extrapolation of ARIMA Models

This criterion attempts to ensure the quality of the forecasts of the ARIMA models. We require that the average percentage forecast error of the fitted error be below a certain level.

Table 8 shows that six of the seven models are equivalent from the point of view of forecasts, i.e. the number of autoregressive and moving average parameters does not affect the forecast error of the model averaged over all the series. Of course, some models perform better for certain series.

Table 9 shows the average forecast error and standard deviation of the error under two possible outcomes: passing and failing the forecast error criterion. Not only is the failure rate of model 3 higher than that of the other models, but the table shows that when it fails,

Table 8

Failure in Forecast Error

	CLASS I						CLASS II		CLASS III					
CRITICAL VALUE	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6		Model 7	
	(0, 1, 1) (0, 1, 1)		(0, 1, 2) (0, 1, 1)		(0, 2, 2) (0, 1, 1)		(2, 1, 2) (0, 1, 1)		(1, 1, 0) (0, 1, 1)		(2, 1, 0) (0, 1, 1)		(2, 1, 0) (0, 1, 2)	
%		%		%		%		%		%		%		%
10	89	53	84	50	101	60	80	48	84	50	85	51	85	51
15	57	34	58	35	69	41	53	32	57	34	56	34	55	33
20	39	23	40	24	51	31	40	24	40	24	40	24	40	24
25	32	19	33	20	43	26	32	19	36	22	14	20	34	20
30	24	14	26	16	35	21	24	14	27	16	27	16	27	16

Table 9

Conditional Mean (M) and Standard Deviation (SD) of the Average Forecast Error

		CLASS I						CLASS II		CLASS III					
Critical Value	Out-come	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6		Model 7	
		(0, 1, 1) (0, 1, 1)		(0, 1, 2) (0, 1, 1)		(0, 2, 2) (0, 1, 1)		(2, 1, 2) (0, 1, 1)		(1, 1, 0) (0, 1, 1)		(2, 1, 0) (0, 1, 1)		(2, 1, 0) (0, 1, 2)	
		M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
15%	Pass	7%	4.0	6%	3.9	7%	4.1	6%	3.8	7%	3.9	7%	4.0	7%	3.9
	Fail	35%	22.3	36%	22.5	41%	26.4	36%	21.4	38%	24.5	37%	23.4	37%	23.0

its average forecast error is bigger. The forecast errors of model 3 are increased by its overdifferencing. However, when the forecast errors of model 3 pass the criterion, their average is as small as that of the other models.

4. RANKING OF THE MODELS

To rank the models, the eight criteria are used at different acceptance levels. Tables 10 and 11 present the overall and conditional rankings of the models. Table 10 gives the total

Table 10

Overall Ranking of the Models

	2 criteria		8 criteria*		8 criteria*		8 criteria*
	FE ≤ 15%		FE ≤ 15%		FE ≤ 15%		FE ≤ 15%
	$\chi^2 \geq$ 5%		$\chi^2 \geq$ 5%		$\chi^2 \geq$ 5%		$\chi^2 \geq$ 5%
			SP ≤ .10		SP ≤ .05		SP ≤ .05
			OD ≥ .90		OD ≥ .90		OD ≥ .95
Models	% of series that passed	Models	% of series that passed	Models	% of series that passed	Models	% of series that passed
4	52%	1	34%	6	38%	6	39%
7	51%	6	31%	1	37%	1	38%
6	49%	5	23%	2	29%	2	29%
2	48%	2	20%	5	26%	5	28%
1	44%	3	13%	7	25%	7	27%
3	41%	7	11%	3	17%	3	19%
5	32%	4	2%	4	4%	4	5%

*As well as the four criteria listed, the four other criteria mentioned in the text were imposed.

Table 11

Conditional Ranking of the Models

	2 criteria		8 criteria*		8 criteria*		8 criteria*
	FE ≤ 15%		FE ≤ 15%		FE ≤ 15%		FE ≤ 15%
	$\chi^2 \geq$ 5%		$\chi^2 \geq$ 5%		$\chi^2 \geq$ 5%		$\chi^2 \geq$ 5%
			SP ≤ .10		SP ≤ .05		SP ≤ .05
			OD ≥ .90		OD ≥ .90		OD ≥ .95
Models	% of series that passed	Models	% of series that passed	Models	% of series that passed	Models	% of series that passed
4	52%	1	34%	6	38%	6	39%
7	9%	3	6%	3	9%	3	9%
2	1%	6	4%	7	4%	1	4%
3	1%	5	2%	2	3%	4	2%

*As well as the four criteria listed, the four other criteria mentioned in the text were imposed.

success rate of the models. Table 11 gives first the total success rate of the best model; the following models are chosen according to their success with series with which all higher models have failed.

Table 10 shows that:

- when only the chi-square statistic (χ^2) and average forecast error (FE) are used as criteria, models 4 and 7, which have the most parameters, rank at the top.
- on the other hand, the use of all criteria favour the simplest models (models 1 and 6), at all levels of small parameter (SP) and overdifferencing (OD) criteria.
- models 1 and 6 usually rank close together, although model 1 has one less parameter than model 6.
- when model 6 is not first it is a close second.
- the more the criteria are relaxed, the higher the pass ratio is, although the ranking of the models remains about the same.

In table 11 we see that:

- when all criteria are used, models 1 and 6 which ranked first and second in table 10 now rank only first and third.
- second place belongs to model 3. This model, which in table 10 ranked third, fifth and sixth with total success rates of 41%, 13%, 17%, and 19%, here ranks fourth once and second three times. This is because model 3 fits well an important family of series (series with a steep trend) that all other models fit poorly.
- moving average and autoregressive models are not mutually exclusive. These two families of models are complementary and necessary in fitting and forecasting series.
- when we require only that the average forecast error be less than 15% and the chi-square statistic be greater than 5% and nothing else, the combined success rate of models 4, 7, 2 and 3 together is 63%.
- when all the criteria are used, the models chosen are simple and their combined success rate varies between 46% and 54% using the levels of 15% and 5% described just above. The success rate depends on the levels of small parameter and overdifferencing used.

Even though model 1 does not appear in the third column of table 11, it would appear there if the level of forecast error permitted were raised to 20%.

The criteria and levels used in selecting models in figures 1 and 2 are the same as are used in the second column of tables 10 and 11, except that in figure 1 the average forecast error permitted varies between 10% and 99% while in figure 2 the chi-square criteria varies between 10% and 60%.

Figure 1 shows that:

- models 1, 3 and 6 perform the best.
- the ranking of the models tends to remain the same.
- the performance of the first model increases more rapidly than that of the others, going from 23% to 59% compared with an increase from 13% to 17% for model 3. This point needs clarification. Model 1 is chosen according to its unconditional performance, while the other models are chosen according to their conditional ranking.
- the increase in performance of the models according to unconditional ranking is greater than the increase when using conditional ranking.

We see in figure 2 that

- models 1, 3 and 6 are generally the best models for any level of chi-square.
- models 1 and 6 trade places but are not mutually exclusive.

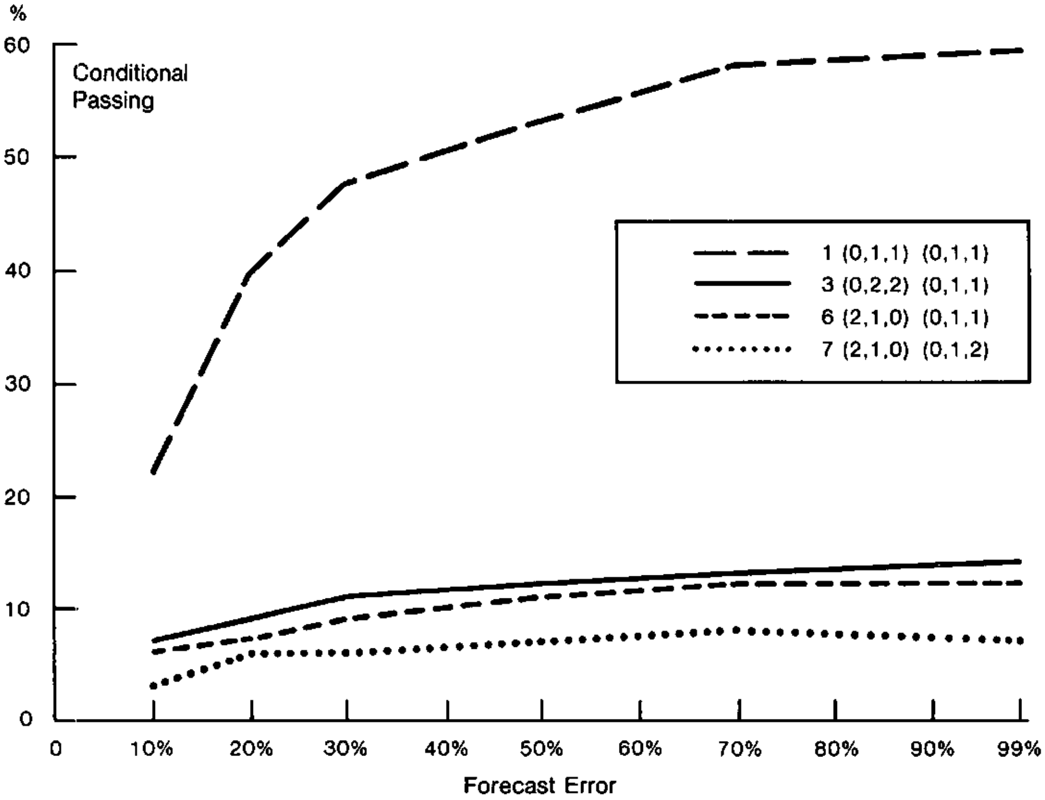


Figure 1. Model Priority Chart for Different Levels of the Forecast Criterion

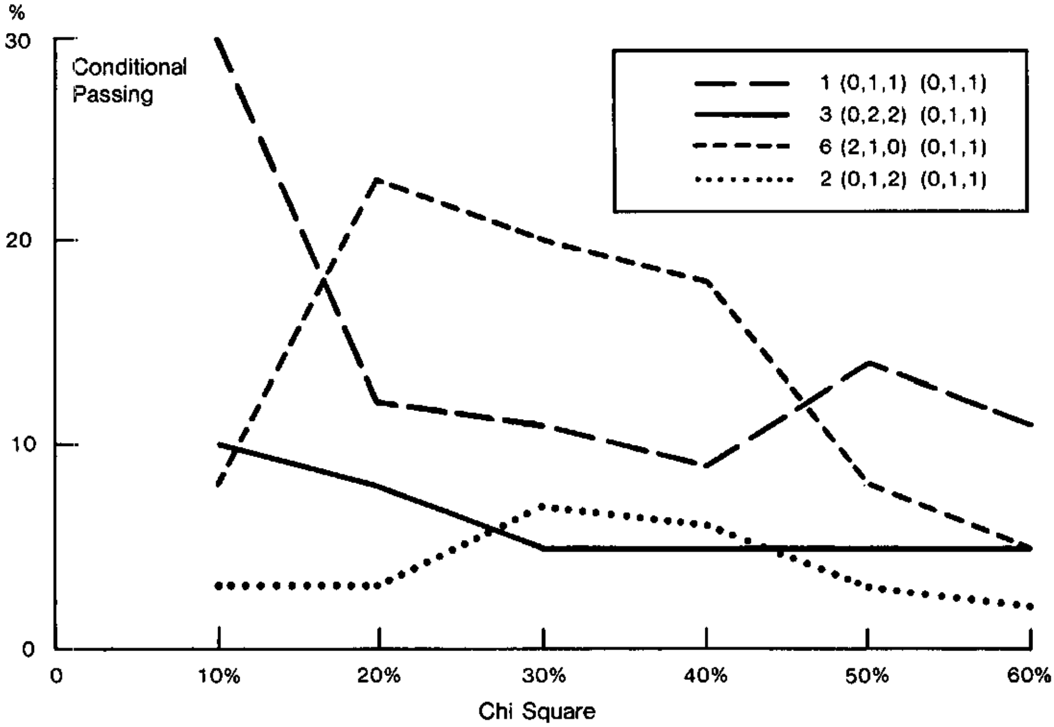


Figure 2. Model Priority Chart for Different Levels of the Chi-Square Criterion

Table 12

Conditional ranking of the ARIMA models for the sectors of the Canadian economy

Sectors	Models ranking and % of series that passed							
	first model	%	second model	%	third model	%	fourth model	%
Labour	1	79	3	14	–	0	–	0
Prices	5	50	7	17	2	8	–	0
Manufacturing	3	19	6	14	1	5	2	5
Fuel, Power and Mining	1	46	6	4	–	0	–	0
Domestic Trade	1	53	6	7	7	7	–	0
External Trade	6	21	–	0	–	0	–	0
Transportation	1	54	5	8	–	0	–	0
Finance	1	32	3	11	–	0	–	0

Table 12 presents the conditional ranking of the ARIMA models for those sectors of the Canadian economy for which we fitted twelve or more series. The criteria and levels used in ranking the models are the same as those used in the second column of tables 10 and 11. We see that

- models 1 and 6 are generally the best performers.
- the combined success rate of the models varies considerably from one sector to another, from 93% in the labour sector to only 21% in external trade.
- this success rate is at least 50% for five sectors. The rate depends on the structure of the series, changes in the structure, and the amount of irregular in the series. The rate is good considering that for two of the last three years Canada suffered a severe recession which strongly affected the structure of the series. The success rate for external trade is always low because those series are very irregular.

5. WITHIN-SAMPLE AND OUT-OF-SAMPLE FORECASTS

The within-sample forecasts are obtained by fitting the models to the entire series in order to estimate the parameters and calculate the forecasts for the last three years. The out-of-sample forecasts do not use information from after the forecast time origin. For each forecast origin, the parameters are re-estimated.

Table 13

Failure Rate in Forecast Error for
Within-Sample and Out-of-Sample Forecasts

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
	(0, 1, 1) (0, 1, 1)	(0, 1, 2) (0, 1, 1)	(0, 2, 2) (0, 1, 1)	(2, 1, 2) (0, 1, 1)	(1, 1, 0) (0, 1, 1)	(2, 1, 0) (0, 1, 1)	(2, 1, 0) (0, 1, 2)
	%	%	%	%	%	%	%
Within-sample	34	35	41	32	34	34	33
Out-of-sample	31	32	42	33	31	32	31

Table 14

Conditional and Unconditional Ranking of the Models

Unconditional ranking		Conditional Ranking	
Models	% of series that passed	Models	% of series that passed
1	40%	1	40%
6	28%	2	5%
5	27%	7	4%
2	20%	3	3%
3	14%		
7	10%		
4	2%		

Table 13 shows the rate of failure in forecast error at the 15% level for within-sample and out-of-sample forecasts. The difference between the two is small and is well within one standard deviation for each model. The X-11-ARIMA seasonal adjustment program uses within-sample forecasts because they cost less.

Table 14 has been prepared using the same criteria and levels as were used in the second columns of tables 10 and 11. The unconditional ranking is exactly the same as that in the second column of table 10. Only the success rates of the first three models differ, and in table 14, model 1 is clearly superior to the other models. However, the conditional ranking is different from that appearing in the second column of table 11.

The conditional rankings in tables 11 and 14 differ for two reasons. First, of course, table 14 uses out-of-sample forecasts. Another important reason is that the calculation of the seven other criteria was based on one year less data, and the missing year contained a severe recession. Thus the structure of the series and the choice of models is markedly different.

It appears therefore that the conditional ranking of the models for both within-sample and out-of-sample forecasts depends on the phase of the business or economic cycle in which the series ends.

6. CONCLUSION

Our objective was to rank a set of seven ARIMA models according to their fitting and forecasting of a large sample of time series.

- when only the chi-square statistic and the average forecast error are used as criteria, models 4 and 7 rank at the top.
- The use of all eight criteria favours the simplest models (1 and 6) and model 3.
- Models 1 (moving average model) and 6 (autoregressive model) rank close together in unconditional ranking, although model 1 has one less parameter than model 6.
- In conditional ranking, these two both rank highly but are not mutually exclusive. That is, moving average and autoregressive models are complementary and both are necessary in fitting and forecasting series.
- Although Model 3 ranks near the bottom, it fits well an important family of series (series with a steep trend) that all other models fit poorly.
- The nonparsimonious models (numbers 4 and 7) have a combined success rate of 61% compared to a success rate that varies between 44% and 52% for parsimonious models 1, 6 and 3.

• The combined success rate of the models varies considerably from one economic sector to another, from 93% in the labour sector to only 21% in external trade. This rate depends on the structure of the series, changes in the structure, and the amount of irregular in the series.
• It appears that the conditional ranking of the models for both within-sample and out-of-sample forecasts depends on the phase of the business or economic cycle in which the series ends.

ACKNOWLEDGEMENT

We acknowledge helpful discussions with Mr. Normand Laniel and we are very grateful to Ms Helen Lim and Mr. Alfred Papineau for their valuable computational assistance, and to Ms B. Cohen for her typing assistance.

REFERENCES

BOX, G.E.P., and JENKINS, G.M. (1970). *Times Series Analysis Forecasting and Control.* Holden Day: San Francisco.

BOX, G.E.P., and PIERCE, D.A. (1970). Distribution of residual autocorrelations in autoregressive integrated moving average time series models. *Journal of the American Statistical Association,* 65, 1509-1526.

DAGUM, E.B. (1980). *The X-11-ARIMA Seasonal Adjustment Method.* Catalogue No. 12-564E, Statistics Canada, Ottawa.

DRAPER, N.R., and SMITH, H. (1981). *Applied Regression Analysis.* John Wiley & Sons, Inc.

HIGGINSON, J. (1976). A test for the presence of seasonality and a model test. Research Paper, Time Series Research and Analysis Division. Statistics Canada, Ottawa.

LJUNG, G.M., and BOX, G.E.P. (1978). On a measure of lack of fit in time series models. *Biometrika,* 65, 297-307.

PANDIT, S.M., and WU, S.M. (1983). *Time Series and System Analysis with Applications.* John Wiley & Sons, Inc.

PLOSSER, C.I., and SCHWERT, G.W. (1977). Estimation of a non-invertible moving average process. *Journal of Econometrics,* 6, 199-224.

PROTHERO, D.L., and WALLIS, K.F. (1976). Modelling macroeconomic time series (with discussion). *Journal of the Royal Statistical Society,* AL39, 468-500.

Survey Methodology, June 1985
Vol. 11, No. 1, pp. 65–77
Statistics Canada

An Empirical Study of Some Regression Estimators for Small Domains

M.A. HIDIROGLOU and C.E. SÄRNDAL[1]

ABSTRACT

The synthetic estimator (SYN) has been traditionally used to estimate characteristics of small domains. Although it has the advantage of a small variance, it can be seriously biased in some small domains which depart in structure from the overall domains. Särndal (1981) introduced the regression estimator (REG) in the context of domain estimation. This estimator is nearly unbiased, however, it has two drawbacks; (i) its variance can be considerable in some small domains and (ii) it can take on negative values in situations that do not allow such values.

In this paper, we report on a compromise estimator which strikes a balance between the two estimators SYN and REG. This estimator, called the modified regression estimator (MRE), has the advantage of a considerably reduced variance compared to the REG estimator and has a smaller Mean Squared Error than the SYN estimator in domains where the latter is badly biased. The MRE estimator eliminates the drawback with negative values mentioned above. These results are supported by a Monte Carlo study involving 500 samples.

KEY WORDS: Small domains; regression estimation; modified regression estimator; bias; mean squared error.

1. INTRODUCTION

The synthetic estimator (SYN) has the advantage of a small variance, but the following disadvantages: (a) it can be badly biased in some domains, and ordinarily we do not know which ones; (b) consequently, a calculated coefficient of variation (cv), or a calculated confidence interval, is meaningless for such domains.

For the same model that underlies the SYN estimator one can create a nearly unbiased analogue, the generalized regression estimator (REG), which has the additional advantage that a standard design based confidence interval is easily computed for each domain estimate. A disadvantage with REG is that the estimated variance (and hence the cv and the width of the confidence interval) can be unacceptably large in very small domains. (This is, of course, a direct consequence of the shortage of observations in such domains.) Also, the REG can (although with small probability) take negative values in situations where such values are unacceptable.

It is therefore desirable to strike a balance between SYN and REG. Here, we report on an empirical study with one such compromise estimator, the modified regression estimator (MRE). It has a small (but noticeable) bias in those domains where the synthetic estimator is greatly biased; in other domains, the MRE is nearly unbiased. The MRE has the advantage of a considerably reduced variance compared to the REG estimator. In addition, the MRE has a smaller Mean Squared Error than the SYN estimator in domains where the latter is badly biased. Meaningful confidence intervals can also be easily constructed for the new MRE estimator.

[1] M.A. Hidiroglou, Business Survey Methods Division, Statistics Canada, 5-C8, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6 and C.E. Särndal, Department of Mathematics and Statistics, University of Montréal, Montréal, Québec, Canada H3C 3J7.

The paper is structured as follows. In Section 2, some of the commonly used estimators for small areas such as the direct, post-stratified and synthetic estimators are reviewed as well as some of the regression estimators given by Särndal (1981, 1984). In Section 3, the proposed modified regression estimators are introduced and discussed. In Section 4, the properties of the modified regression estimators as well as some of the other estimators are studied through a Monte Carlo simulation using business tax data. Finally, Section 5 provides some general conclusions.

2. ESTIMATORS

Let the population $U = \{1, ..., k, ..., N\}$ be divided into D non-overlapping domains $U_{1.}, ..., U_{d.}, ..., U_{D.}$. Let $N_{d.}$ be the size of $U_{d.}$. (In our empirical study, the domains are defined by a cross-classification of 4 industrial groupings with the 18 census divisions in the province of Nova Scotia. There were $D = 70$ non-empty domains, as described in Hidiroglou, Morry, Dagum, Rao and Särndal (1984).)

The population is further divided along a second dimension, into G non-overlapping groups, $U_{.1}, \ldots, U_{.g}, \ldots, U_{.G}$.

The size of $U_{.g}$ is denoted $N_{.g}$. (In our study, the groups are based on Gross Business Income classes.) The cross-classification of domains and groups gives rise to DG population cells U_{dg}; $d = 1, \ldots, D$; $g = 1, \ldots, G$. Let N_{dg} be the size of U_{dg}.

Then the population size N can be expressed as

$$N = \sum_{d=1}^{D} N_{d.} = \sum_{g=1}^{G} N_{.g} = \sum_{d=1}^{D} \sum_{g=1}^{G} N_{dg} \tag{2.1}$$

Let s denote a sample of size n drawn from U by simple random sampling (srs). Denote by $s_{d.}$, $s_{.g}$ and s_{dg} the parts of s that happen to fall, respectively, in $U_{d.}$, $U_{.g}$ and U_{dg}.

The corresponding sizes, which are random variables, are denoted by $n_{d.}$, $n_{.g}$ and n_{dg}. Note that (2.1) holds for lower case n's as well. The variable of interest, y (= Wages and Salaries) takes the value of y_k for the k:th unit (= unincorporated business tax filer). The auxiliary variable x (= Gross Business Income) takes the value x_k for the k:th unit, and x_k is known for all $k = 1, \ldots, N$.

The following estimators of the domain total $t_d = \sum_{U_{d.}} y_k$ are compared, where $\sum_{U_{d.}}$ denotes the summation over the units in $U_{d.}$.

The straight expansion estimator (EXP):

$$\hat{t}_{d\text{EXP}} = \frac{N}{n} \sum_{s_{d.}} y_k \tag{2.2}$$

The poststratified estimator (POS):

$$\hat{t}_{d\text{POS}} = N_{d.} \bar{y}_{s_{d.}} \tag{2.3}$$

where

$$\bar{y}_{s_{d.}} = \sum_{s_{d.}} \frac{y_k}{n_{d.}}$$

is the mean of the $n_{d.}$ y-values from the d:th domain. If $n_{d.} = 0$ we define the POS estimator to be zero (somewhat arbitrarily, since strictly speaking the estimator is then undefined). Neither the EXP nor the POS estimator are particularly advantageous. They serve mainly as benchmarks against which the behaviour of the following more efficient estimators will be compared.

Two versions of the SYN and REG have been investigated, the "Count" version and the "Ratio" version. The SYN estimator is based on the assumption that a given model holds for each group g. For the "Count" version a given model would lead to the assumption that the mean of each group is the same across all domains d. For the "Ratio" version, the implied model would be that the ratios of a given variable of interest over an auxiliary variable would be constant within a given group across all domains. If the assumption of homogeneity of domain characteristics does not hold within each group, the SYN estimators can be very biased. The REG estimation method as given by Särndal (1984) is motivated by the following requirements: (a) to obtain approximately design-unbiased estimates with simple variance estimates and easily calculable (and meaningful) confidence intervals; (b) to strengthen the estimates by involving sample data from all domains.

The formulas for the "Count" versions are:

Synthetic-Count estimator (SYN/C):

$$\hat{t}_{d\mathrm{SYN/C}} = \sum_{g=1}^{G} N_{dg}\bar{y}_{s_{.g}} \tag{2.4}$$

where $\bar{y}_{s_{.g}}$ is the mean of y in $s_{.g}$.

Regression-Count estimator (REG/C):

$$\hat{t}_{d\mathrm{REG/C}} = \sum_{g=1}^{G} \{N_{dg}\bar{y}_{s_{.g}} + \hat{N}_{dg}(\bar{y}_{s_{dg}} - \bar{y}_{s_{.g}})\} \tag{2.5}$$

where $\bar{y}_{s_{dg}}$ is the mean of y in s_{dg}, and $\hat{N}_{dg} = Nn_{dg}/n$. Here, $\sum_{g=1}^{G} \hat{N}_{dg}(\bar{y}_{s_{dg}} - \bar{y}_{s_{.g}})$ is a bias correction term that ordinarily carries a considerable variance contribution.

The "Ratio" versions of the SYN and REG estimators are:

Synthetic-Ratio estimator (SYN/R):

$$\hat{t}_{d\mathrm{SYN/R}} = \sum_{g=1}^{G} X_{dg}\hat{R}_g \tag{2.6}$$

with $X_{dg} = \sum_{U_{dg}} x_k$ and

$$\hat{R}_g = \frac{\sum_{s_{.g}} y_k}{\sum_{s_{.g}} x_k}$$

Regression - Ratio estimator (REG/R):

$$\hat{t}_{d\mathrm{REG/R}} = \sum_{g=1}^{G} \{X_{dg}\hat{R}_g + \hat{N}_{dg}(\bar{y}_{s_{dg}} - \hat{R}_g\bar{x}_{s_{dg}})\} \tag{2.7}$$

3. MODIFIED REGRESSION ESTIMATORS

Regression estimators introduced by Särndal (1984) were constructed by fitting a regression model to some auxiliary variables and using the resulting fitted model to create predicted values for the units in the population domain. Assuming that the sampling design, p, is an arbitrary one (not necessarily srs) with inclusion probabilities π_k (first order) and π_{kt} (second order), let the regression model be given by

$$E_\xi(y_k) = \underset{\sim}{x}_k'\underset{\sim}{\beta};\ V_\xi(y_k) = v_k$$

where the y_k are independent random variables. An estimator of $\underset{\sim}{\beta}$ is

$$\hat{\underset{\sim}{\beta}} = \left(\sum_s \frac{\underset{\sim}{x}_k' \underset{\sim}{x}_k}{\nu_k \pi_k} \right)^{-1} \sum_s \frac{\underset{\sim}{x}_k' y_k}{\nu_k \pi_k}$$

where it is assumed that the ν_k are known to multiplicative constant(s) that cancel when $\hat{\underset{\sim}{\beta}}$ is derived.

Following Särndal (1984), a nearly unbiased estimator of the unknown d-th domain total is given by

$$\hat{t}_{d\text{REG}} = \sum_{U_{d.}} \hat{y}_k + \sum_{s_{d.}} \frac{e_k}{\pi_k} \tag{3.1}$$

where $\hat{y}_k = \underset{\sim}{x}_k' \hat{\underset{\sim}{\beta}}$ is the k-th predicted value and $e_k = y_k - \hat{y}_k$ denotes the k-th residual.

We shall refer to $\sum_{U_{d.}} \hat{y}_k$ as *the synthetic term* of the estimator $\hat{t}_{d\text{REG}}$ and the second term, $\sum_{s_{d.}} e_k/\pi_k$, will be called the *correction* term.

If $s_{d.}$ is non-empty, an approximately unbiased alternative to the REG estimator (3.1) is given by

$$\hat{t}_{d\text{ALT}} = \sum_{U_{d.}} \hat{y}_k + N_{d.} \frac{\sum_{s_{d.}} \frac{e_k}{\pi_k}}{\hat{N}_{d.}} \tag{3.2}$$

where

$$\hat{N}_{d.} = \sum_{s_{d.}} \frac{1}{\pi_k}$$

is the estimated domain size.

The correction term now appears in the form of a ratio estimator,

$$\frac{\sum_{s_{d.}} \frac{e_k}{\pi_k}}{\sum_{s_{d.}} \frac{1}{\pi_k}},$$

multiplied by the known domain size $N_{d.}$ (obviously, $N_{d.}$ is known since the cell counts N_{dg} are known).

The size $n_{d.}$ being random, the ratio form will serve to reduce the variance of the correction term. The effect will be particularly noticeable in domains where the average of the residuals is clearly away from zero (that is, in domains where the model does not fit well).

If the expected sample take in the domain, $E_d = E_p(n_{d.}) = \sum_{U_{d.}} \pi_k$, were substantial (say, $E_d \geq 50$), then it is practically certain that the realized sample take, $n_{d.}$, will not be exceedingly small. For example, under srs, values $n_{d.} \leq 30$ will hardly ever occur. In such situations, the nearly unbiased estimator (3.2) can be recommended as is. It should realize important efficiency gains over (3.1), notably in domains where the model does not fit as well. But in practice one often encounters domains that are so small that the expected sample take E_d does not exceed 5. This is true for a number of domains in our study. In such cases, realized sample takes $n_{d.}$ between zero and five are very likely. Our empirical work has confirmed the intuitively obvious fact that the residual correction will, in these small domains, contribute greatly to the variance, whether the correction appears in its straight form, $\sum_{s_{d.}} e_k/\pi_k$, as in (3.1), or in its ratio form , $N_{d.}(\sum_{s_{d.}} e_k/\pi_k)/(\sum_{s_{d.}} 1/\pi_k)$, as in (3.2).

To counteract this inflated variance contribution, we modify the correction term of (3.2) in a way implying that we settle for a small bias (in domains where the model fits less well) in exchange for a reduced variance contribution when the realized sample take $n_{d.}$ is lower than expected (and it is assumed that the expected sample take is already low in itself).

The form of the new correction term will be determined by the relation between realized sample take $n_{d.}$, and expected sample take E_d. The correction term $\sum_{s_{d.}} e_k/\pi_k$ will be multiplied by $(\hat{N}_d/N_d)$ when $n_{d.} < E_d$ and by $(N_d/\hat{N}_d)$ otherwise. The resulting correction term using this adaptive "dampening factor" will have the effect of not "over-correcting" the synthetic term when some of the residuals e_k behave as outliers for small $n_{d.}$'s. The "over-correcting" may have the effect of greatly underestimating a domain d, yielding negative values when only positive values are acceptable, or conversely greatly overestimating the domain.

The resulting estimator, the modified regression estimator (MRE), incorporating these two types of realizations of $n_{d.}$, is

$$\hat{t}_{d\text{MRE}} = \sum_{U_{d.}} \hat{y}_k + F_d \sum_{s_{d.}} \frac{e_k}{\pi_k} \tag{3.3}$$

where

$$F_d = \begin{cases} \dfrac{N_{d.}}{\hat{N}_{d.}} \text{ when } n_{d.} \geq E_d \\ \\ \dfrac{\hat{N}_{d.}}{N_{d.}} \text{ when } n_{d.} < E_d \end{cases}$$

It can be shown that (3.3) is nearly unbiased conditionally on $n_{d.}$, as long as $n_{d.} \geq E_d$. For $n_{d.} < E_d$, the MRE has some conditional bias, which tends to increase the more $n_{d.}$ falls short of its expected value. At the same time, the MRE estimator is being pushed towards its synthetic term, thus benefitting from the stability (low variance) of the synthetic term. Unconditionally, the MRE estimator given by (3.3) will have a certain small bias, but a much reduced variance compared with the REG estimator.

We note a final point in favour of MRE estimator. As a result of its considerable variance in very small domains, the REG estimator will, with a small but positive probability, take values extremely removed from the true value $t_{d.}$. The value of the REG may even be negative, which is, of course, unacceptable for a variable (such as Wages and Salaries) which is by definition non-negative. Negative values of the REG estimate can occur when there exists large negative residuals e_k in the correction term of (3.1), and are especially likely when $n_{d.} < E_d$. The new MRE estimator virtually eliminates this occurence of negative estimates. In practice, if by a remote possibility the MRE takes a negative value, we recommend to redefine the MRE estimator as being equal to the always positive SYN estimator.

A natural formula for estimating the variance of (3.2) is

$$\hat{V}_p(\hat{t}_{d\text{ALT}}) = \left(\frac{N_{d.}}{\hat{N}_{d.}}\right)^2 \sum_{\substack{k \neq \ell \\ \in s_{d.}}}\sum \Delta_{k\ell} \frac{(e_k - \hat{e}_{s_{d.}})(e_\ell - \tilde{e}_{s_{d.}})}{\pi_k \pi_\ell} \tag{3.4}$$

where

$$\tilde{e}_{s_{d.}} = \frac{\sum_{s_{d.}} \dfrac{e_k}{\pi_k}}{\sum_{s_{d.}} \dfrac{1}{\pi_k}}$$

and

$$\Delta_{k\ell} = \begin{cases} 1 - \pi_k & \text{if } \ell = k \\ \\ 1 - \dfrac{\pi_k \pi_\ell}{\pi_{k\ell}} & \text{if } \ell \neq k. \end{cases}$$

We propose that the same formula may serve well to estimate the variance of the MRE estimator (3.3). It is true that (3.3) differs from (3.2) when the realized sample take falls short of the expected sample take; however, it is not foreseen that the difference will be great enough to cause serious distortion in the validity of a confidence interval for t_d centred on $\hat{t}_{d\mathrm{MRE}}$ using (3.4) as the estimated variance.

In the case of simple random sampling, and assuming for $g = 1, \ldots, G$,

$$E_\xi(y_k) = \beta_g;\ V_\xi(y_k) = \sigma_g^2;\ k \in U_{.g}, \tag{3.5}$$

we find

$$\hat{\beta}_g = \frac{\sum_{s_{.g}} y_k}{n_{.g}} = \bar{y}_{s_{.g}},$$

leading to the "Count estimator" whose modified version (MRE/C) is

$$\hat{t}_{d\mathrm{MRE/C}} = \sum_{g=1}^{G} \{N_{dg}\bar{y}_{s_{g.}} + F_d\hat{N}_{dg}(\bar{y}_{s_{dg}} - \bar{y}_{s_{.g}})\} \tag{3.6}$$

where E_d in the formula for F_d is now given by

$$E_d = E_{\mathrm{srs}}(n_{d.}) = \frac{nN_{d.}}{N}$$

with

$$\hat{N}_{dg} = n_{dg}\left(\frac{N}{n}\right)$$

and

$$\bar{y}_{s_{dg}} = \begin{cases} \dfrac{\sum_{s_{dg}} y_k}{n_{dg}} & \text{for } n_{dg} \geq 1 \\ 0 & \text{otherwise.} \end{cases}$$

The MRE/C estimator will have some bias, which is, however, ordinarily much less than that of the SYN/C estimator.

The underlying model assumptions which lead to the "ratio estimator", whose modified version is denoted as MRE/R, are for $g = 1, \ldots, G$,

$$E_\xi(y_k) = \beta_g x_k;\ V_\xi(y_k) = \sigma_g^2 x_k,\ k \in U_{.g}.$$

The MRE/R estimator is then, in the case of simple random sampling,

$$\hat{t}_{d\mathrm{MRE/R}} = \sum_{g=1}^{G} \{X_{dg}\hat{R}_g + F_d\hat{N}_{dg}(\bar{y}_{s_{dg}} - \hat{R}_g\bar{x}_{s_{dg}})\} \tag{3.7}$$

where

$$\hat{R}_g = \frac{\sum_{d=1}^{D} \hat{N}_{dg}\tilde{y}_{s_{dg}}}{\sum_{d=1}^{D} \hat{N}_{dg}\tilde{x}_{s_{dg}}},$$

and

$$X_{dg} = \sum_{U_{dg}} x_k.$$

Drew, Singh and Choudhry (1982) provided small domain estimators which, although not derived by a regression approach, have some similarity to the ones given in this paper. Their "count" version is

$$\hat{t}_{d\text{KNO/C}} = \sum_{g} N_{dg}\{W'_{dg}\tilde{y}_{s_{dg}} + (1 - W'_{dg})\tilde{y}_{s_{.g}}\} \tag{3.8}$$

while their "ratio" version is

$$\hat{t}_{d\text{KNO/R}} = \sum_{s_g} X_{dg}\left\{W'_{dg}\,\frac{\tilde{y}_{s_{dg}}}{\tilde{x}_{s_{dg}}} + (1 - W'_{dg})\frac{\tilde{y}_{s_{.g}}}{\tilde{x}_{s_{.g}}}\right\} \tag{3.9}$$

where

$$W'_{dg} = \begin{cases} \dfrac{n_{dg}}{E_{dg}} \text{ if } n_{dg} \leq E_{dg} \\ 1 \quad \text{otherwise} \end{cases}$$

with $E_{dg} = n(N_{dg}/N)$. In the present context, if W'_{dg} in (3.8) is replaced by

$$W''_{dg} = \begin{cases} \left(\dfrac{n_{d.}}{E_d}\right)\left(\dfrac{n_{dg}}{E_{dg}}\right) & \text{if } n_{d.} < E_d \\ \left(\dfrac{E_d}{n_{d.}}\right)\left(\dfrac{n_{dg}}{E_{dg}}\right) & \text{if } n_{d.} \geq E_d \end{cases}$$

we obtain $\hat{t}_{d\text{MRE/C}}$.

4. RESULTS FROM THE EMPIRICAL STUDY

In order to study the properties of the estimators discussed in the preceding sections, a simulation was undertaken. The province of Nova Scotia was chosen as our population with $N = 1678$ sampling units (unincorporated tax filers). The variable of interest, y, is Wages and Salaries. We use a single auxiliary variable, x, namely, Gross Business Income. It is assumed that $x_1, \ldots, x_N$ are known.

Domains of the population were formed by a cross-classification of four industrial groups with eighteen regions. The industrial groups were Retail (515 units), Construction (496 units), Accommodation (114 units) and Others (553 units). The overall correlation coefficients between Wages and Salaries and Gross Business Income were 0.42 for Retail, 0.64 for Construction, 0.78 for Accommodation and 0.61 for Others. The regions were the 18 Census Divisions of the province. This produced 70 non-empty domains (out of the four times 18 domains, two combinations had no units). Thus, 70 domain totals t_d are to be estimated every time a sample is drawn.

For the Monte Carlo simulation, 500 simple random samples, s, each of size $n = 419$, were selected from the population of $N = 1678$ units. The selected sample units were classified into type of industry and Census Division. The population could have been divided along a second dimension, say income groups. But for the purposes of this study, all the taxfilers were considered as belonging to one income group ($G = 1$).

The results are summarized for each small area within the industrial groups RETAIL and ACCOMMODATION using tables and graphs. For the tables (1-4), summary statistics are the relative conditional bias and mean squared error. The eight graphs, one for each of the eight estimators, are given in figure 1. In each graph, there are eighteen vertical 'distribution bands', one for each of the eighteen Census Divisions for the industrial group RETAIL. The upper and lower points of each distribution band correspond, respectively, to the 90:th and 10:th percentile of the distribution of the 500 values of $(\hat{t}_{d.} - t_{d.})/t_{d.}$. Consequently, a distribution band placed roughly symmetrically about the zero line indicates that the corresponding estimator is approximately unbiased for the domain of interest; otherwise, the estimator is biased for the domain. The shorter the band, the smaller the variance of the estimator in the domain. The abscissa measures the mean sample take for the domain.

From the tables and graphs, the following conclusions emerge: (where conclusion C states the main new results, whereas A and B resume what is known from earlier work Särndal and Råbäck (1983); Hidiroglou et al. (1984)).

A. The SYN/C and SYN/R estimators are badly biased in some domains, namely, in those domains where the underlying model fits poorly. However, they consistently have an attractively low variance, compared to the other alternatives. The Mean Squared Error of the two SYN estimators will consequently be very large in domains with large bias (poor model fit); by contrast, the Mean Squared Error is small in domains with little bias (good model fit).

B. The REG/C and REG/R estimators are essentially unbiased. Their variance, although usually much lower than that of the EXP and POS estimators, is consistently much higher than that of the SYN/C and SYN/R estimators. In the smallest domains, none of the unbiased estimators (EXP, POS, REG/C, REG/R) is attractive from the variance point of view; this is especially true for the REG estimators. This problem is remedied by the two MRE modifications of the REG estimators.

C. The two MRE estimators, MRE/C and MRE/R, are negligibly biased when the SYN estimators happen to be nearly unbiased (e.g., RETAIL, area 17); otherwise the MRE estimators have a certain bias, which, however, is ordinarily much less pronounced than that of the SYN estimators (e.g., RETAIL, area 2). The MRE estimators have considerably smaller variance and Mean Squared Error, in all domains, than the REG estimators. This tendency is particularly pronounced in the smaller domains. In comparison with the SYN estimators, we find that the MRE estimators (as expected) still have a larger variance in virtually all domains. However, the Mean Squared Error of the MRE estimators is smaller than that of the SYN estimators in domains where the latter are badly biased. In Table 6 we see, for example, that the MRE/R estimator has a smaller Mean Squared Error than that of the SYN/R in 9 out of 16 small areas. The obvious explanation is that in domains where the SYN estimator is greatly biased, the (bias)^2 constitutes an extremely large contribution to the Mean Squared Error of the SYN, whereas for the MRE estimators, the (bias)^2 is not very important. Since we do not know which domains create the large biases, the goal of producing reliable estimates in all domains is on the whole better served by the MRE method of estimation.

Table 1

Mean Sample Take and Relative Bias of Each of Eight Estimators over 500 Repeated Simple Random Samples from the Entire Population Industrial Group: RETAIL; 18 Census Divisions in Nova Scotia.

	Mean	Estimator							
Area	Sample Take	EXP	POS	SYN/C	MRE/C	REG/C	SYN/R	MRE/R	REG/R
1	1.76	−0.02	−0.13	0.12	0.02	−0.03	0.30	0.09	−0.02
2	5.45	0.00	−0.04	−0.36	−0.10	−0.02	−0.27	−0.08	−0.02
3	3.90	−0.02	0.01	−0.08	−0.02	0.00	−0.01	−0.01	0.00
4	3.02	0.01	−0.05	0.15	0.05	0.01	0.13	0.04	0.04
5	5.93	0.00	0.01	0.21	0.05	0.00	0.13	0.03	0.00
6	7.63	−0.02	−0.01	0.28	0.07	0.01	0.10	0.02	0.00
7	8.61	0.02	0.01	−0.16	−0.03	0.01	−0.18	−0.03	0.01
8	5.64	−0.02	−0.01	0.34	0.10	0.03	0.24	0.06	0.01
9	24.64	0.00	0.00	−0.02	0.00	0.00	−0.01	0.00	0.01
10	8.92	−0.02	−0.02	0.15	0.02	−0.01	0.09	0.00	−0.01
11	8.35	−0.03	−0.02	0.08	0.01	0.00	0.10	0.02	0.00
12	10.58	0.01	0.00	−0.27	−0.05	0.00	−0.18	−0.03	0.00
13	0.48	−0.04	−0.58	0.61	0.36	0.04	1.00	0.58	0.04
14	2.80	0.03	−0.03	0.33	0.11	0.00	0.24	0.10	0.02
15	4.21	0.06	−0.01	0.28	0.06	0.00	0.30	0.07	−0.01
16	2.24	0.03	−0.05	0.74	0.26	0.03	0.94	0.32	0.02
17	23.95	−0.01	−0.01	−0.02	0.00	0.00	−0.05	−0.01	0.00
18	0.54	0.07	−0.54	0.63	0.34	−0.06	0.67	0.35	−0.06

Table 2

Mean Squared Error of Each of Eight Estimators over 500 Repeated Simple Random Samples from the Entire Population Industrial Group: RETAIL; 18 Census Divisions in Nova Scotia.

	Estimator							
Area	EXP	POS	SYN/C	MRE/C	REG/C	SYN/R	MRE/R	REG/R
1	3,209	2,206	96	697	1,397	462	769	1,484
2	42,598	24,623	21,782	12,725	17,358	13,110	10,256	14,380
3	10,469	6,853	357	2,592	4,212	146	2,333	3,782
4	5,626	3,657	324	746	1,186	257	1,206	1,853
5	14,554	9,681	2,999	5,090	7,360	1,294	3,993	5,974
6	12,308	5,686	6,713	3,423	4,289	1,255	1,747	2,515
7	34,865	17,988	6,912	9,387	13,451	8,161	12,019	17,239
8	12,066	8,630	5,772	3,694	5,045	2,981	3,528	4,986
9	72,974	40,440	5,776	24,025	29,250	5,068	21,292	25,832
10	22,091	9,433	4,559	5,832	7,927	2,009	5,365	7,272
11	23,519	12,505	1,778	6,738	9,578	2,348	7,890	11,063
12	46,588	21,874	35,310	13,558	17,084	17,454	12,222	16,514
13	635	244	161	95	228	422	287	783
14	3,871	2,849	692	1,254	2,141	378	1,373	2,346
15	8,088	3,511	2,249	1,892	2,806	2,651	1,985	2,937
16	3,245	2,127	3,316	1,563	2,516	5,333	1,741	2,654
17	81,211	47,753	5,503	28,957	35,232	7,681	27,457	33,136
18	1,003	306	169	187	654	186	184	637

Table 3

Mean Sample Take and Relative Bias of Each of Eight Estimators over 500 Repeated Samples from the Entire Population
Industrial group: ACCOMMODATION; Areas: 16 Census Divisions in Nova Scotia.

	Mean	Estimator							
Area	Sample Take	EXP	POS	SYN/C	MRE/C	REG/C	SYN/R	MRE/R	REG/R
1	0.25	0.01	−0.75	−0.08	−0.06	−0.01	0.36	0.28	0.01
2	1.37	−0.06	−0.21	0.25	0.10	0.02	0.25	0.11	0.02
3	1.02	0.06	−0.26	0.19	0.09	0.04	0.12	0.06	0.03
4	0.23	−0.10	−0.77	−0.33	−0.26	−0.07	−0.15	−0.13	−0.05
5	2.04	0.03	−0.13	0.21	0.08	0.03	0.18	0.06	0.01
6	1.49	0.04	−0.13	0.17	0.10	0.03	0.03	0.02	0.01
7	1.53	0.01	−0.18	−0.29	−0.11	−0.01	−0.30	−0.12	−0.02
8	1.54	0.03	−0.19	−0.42	−0.17	−0.01	−0.26	−0.11	−0.02
9	6.83	0.01	−0.02	0.13	0.02	0.00	0.12	0.02	0.00
10	1.26	−0.01	−0.26	0.40	0.17	0.03	0.30	0.13	0.02
11	3.06	0.04	−0.02	0.51	0.21	0.08	0.40	0.16	0.06
12	1.80	0.02	−0.16	−0.08	−0.05	−0.03	−0.23	−0.10	−0.03
14	1.04	0.02	−0.33	−0.52	−0.23	−0.07	−0.32	−0.15	−0.06
15	1.54	−0.03	−0.23	−0.21	−0.13	−0.08	−0.15	−0.11	−0.08
17	3.08	−0.07	−0.05	−0.03	−0.01	0.00	−0.14	−0.07	−0.03
18	0.52	0.04	−0.54	3.26	3.20	0.60	2.97	2.92	0.50

Table 4

Mean Squared Error of Each of Eight Estimators over 500 Repeated Simple Random Samples from the Entire Population
Industrial Group: ACCOMMODATION; Areas: 16 Census Divisions in Nova Scotia.

	Estimator							
Area	EXP	POS	SYN/C	MRE/C	REG/C	SYN/R	MRE/R	REG/R
1	1,142	283	9	7	25	58	44	164
2	7,467	5,082	877	631	1,077	747	455	726
3	878	442	48	163	242	24	116	163
4	155	43	7	6	17	3	3	6
5	15,200	8,392	2,091	2,270	3,230	1,271	1,208	1,785
6	5,239	3,906	253	1,038	2,193	54	396	792
7	21,197	8,781	3,569	1,831	3,016	3,709	1,812	2,948
8	14,071	6,738	3,608	2,122	4,018	1,492	947	1,766
9	50,606	27,867	9,980	11,413	14,344	6,575	7,779	9,991
10	2,219	993	590	362	665	317	151	280
11	10,535	5,774	6,366	5,126	7,154	3,867	2,752	3,673
12	16,787	10,485	543	1,148	1,944	1,245	1,130	1,836
14	51,471	25,644	9,669	8,221	14,155	3,972	3,189	5,077
15	59,207	41,381	4,861	10,548	18,119	2,759	4,262	6,636
17	29,632	25,211	1,501	3,023	4,754	1,765	2,123	3,214
18	286	99	2,062	2,112	5,623	1,607	1,646	4,561

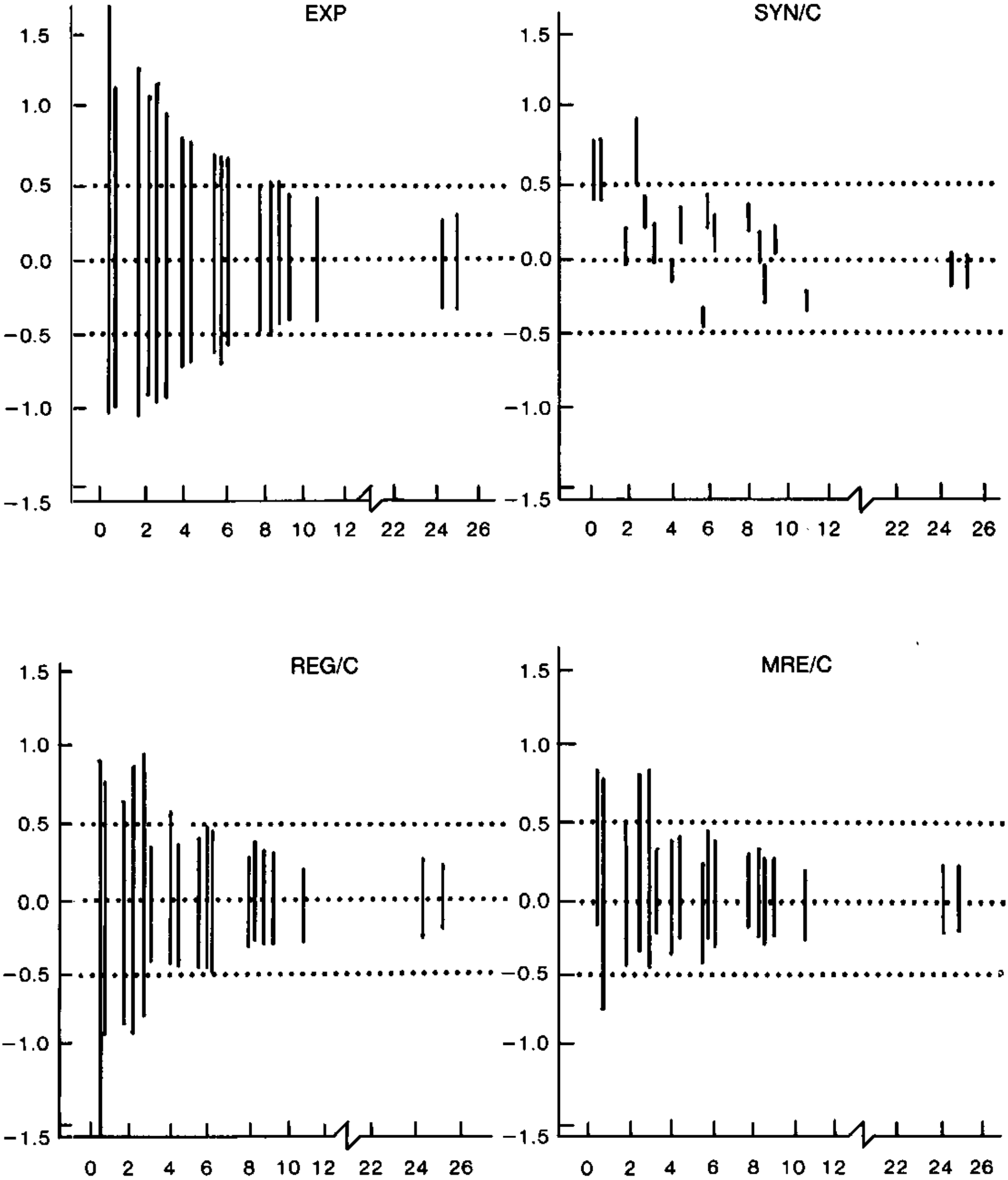
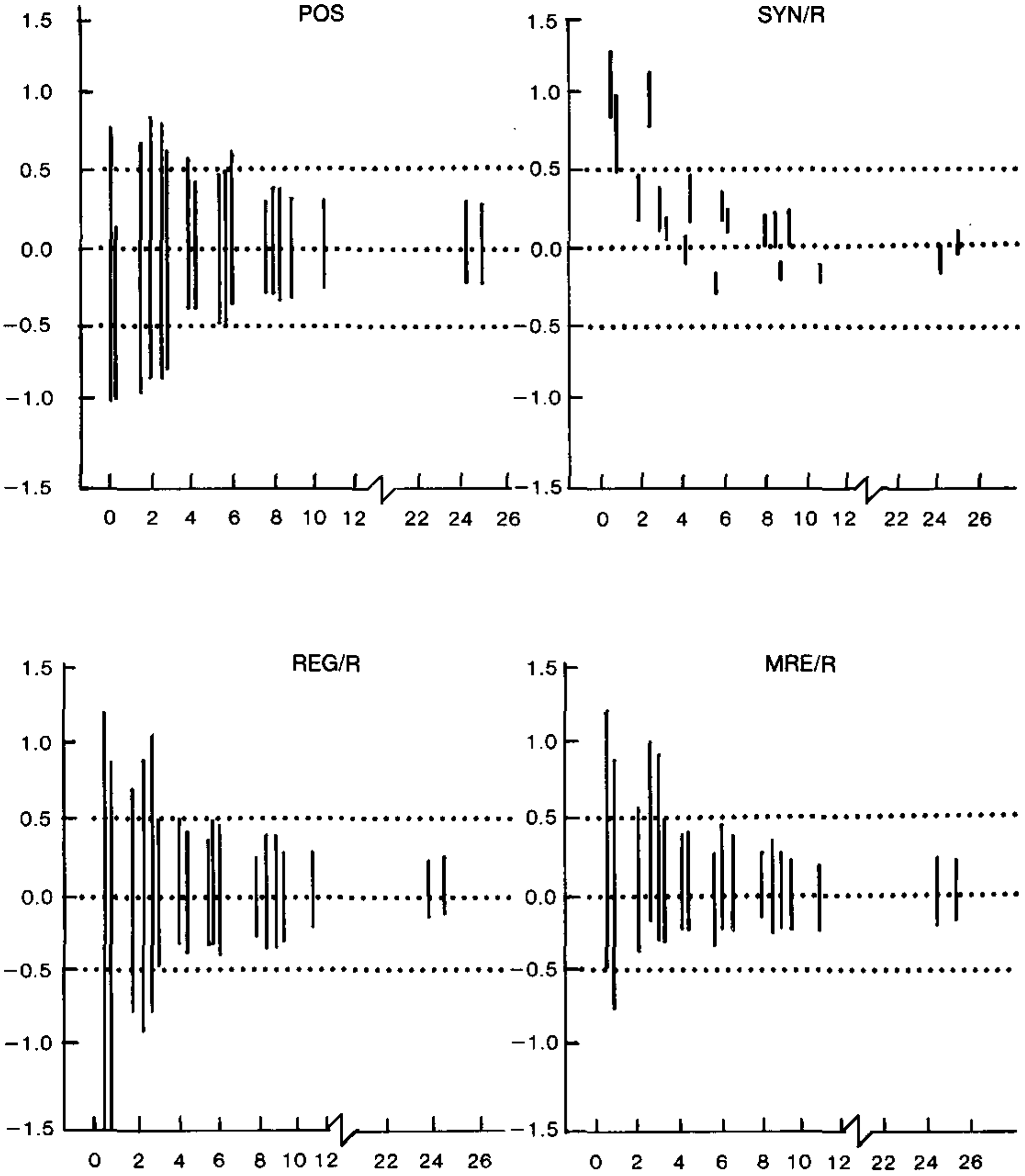


Figure 1: Distribution band of relative error for selected estimators — abscissa represents mean sample take. Industrial Group: RETAIL. Areas: 18 Census Divisions in Nova Scotia.

Figure 1 (continued)



5. CONCLUSIONS

In summary we find that the overall performance of the MRE estimators is such that we suggest them as promising alternatives for future applications of small area estimation. The recommended confidence interval procedure based on the MRE estimators is given in section 3.

We think that the MRE method presented here involves a simple mechanism for steering the estimates slightly in the direction of the stable SYN estimators, when the sample take is less than expected. This goal is also manifested (but attained by different means) in such other attempts as the empirical Bayes (Fay and Herriot, 1979) and sample-dependent (Drew, Singh, and Choudhry 1982) methods of estimation.

REFERENCES

DREW, J.D., SINGH, M.P. and CHOUDHRY, G.H. (1982). Evaluation of small area estimation techniques for the Canadian Labour Force Survey. *Survey Methodology,* 8, 17-47.

FAY, R.E. and HERRIOT, R. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association,* 74, 269-277.

HIDIROGLOU, M.A., MORRY, M., DAGUM, E.B., RAO, J.N.K. and SÄRNDAL, C.E. (1984). Evaluation of alternative small area estimators using administrative data. Paper presented at ASA meetings, Philadelphia, August, 1984.

SÄRNDAL, C.E. (1981). Frameworks for inference in survey sampling with applications to small area estimation and adjustments for nonresponse. *Bulletin of the International Statistical Institute,* 49:1, 494-513. (proceedings, 43rd session, Buenos Aires).

SÄRNDAL, C.E. and RÅBÄCK, G. (1983). Variance reduction and unbiasedness for small domain estimators. *Statistical Review,* 1983:5 (Essays in honour of T.E. Dalenius), 33-40.

SÄRNDAL, C.E. (1984). Design-Consistent versus Model-Dependent Estimation for Small Domains. *Journal of the American Statistical Association,* 79, 624-631.

Survey Methodology, June 1985
Vol. 11, No. 1, pp. 79–88
Statistics Canada

1981 Census of Agriculture Data Processing Methodology

DAVID K. HOLLINS[1]

ABSTRACT

This paper presents an overview of the methodology used in the processing of the 1981 Census of Agriculture data. The edit and imputation techniques are stressed, with emphasis on the multivariate search algorithm. A brief evaluation of the system's performance is given.

KEY WORDS: Edit and imputation; Multivariable searches

1. INTRODUCTION

This paper presents an overview of the methodology used in the processing of the 1981 Census of Agriculture data. There are 3 separate phases to the processing of the data: Data Entry, Edit, and Imputation, each of which performs a different function. First, in Data Entry, data on the questionnaires are keyed onto a computer data file. Then, in the Edit phase, computer edits are applied to the keyed data records in order to detect any inconsistent, missing, or suspicious entries. In the final phase, Imputation, actions are taken to adjust the data records so that they conform to the rules defined by the computer edits applied during Edit. The methodology involved in each of the three phases of processing is described in subsequent sections of this paper. A flow chart of the 1981 Census of Agriculture processing is given in Figure 1.

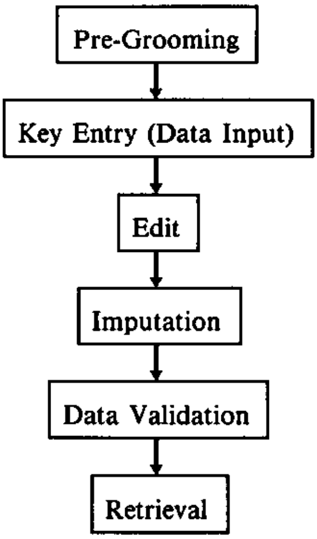


Figure 1. Overall Process Flow

[1] D.K. Hollins, Census and Household Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

The 1981 Census of Agriculture required that the same questionnaire be completed by each farm operator in Canada. The questionnaire is 8 pages long and consists of 134 questions. Questions are asked on all aspects of farm operation, including items such as types of crops grown, livestock raised, equipment maintained, and types of land use. Operators are required to answer only those sections of the questionnaire which apply to their holding.

As this paper is an overview, it is not possible to delve into the technical computer aspects of the Census of Agriculture processing. These details may be found in Shields and Yiptong (1981), on which this paper is based.

2. DATA ENTRY

In the Data Entry phase the Census of Agriculture data are transferred from the original questionnaires to a data file in computer memory. Data entry is comprised of two stages: a clerical pre-grooming process (Pre-Scan), and Key Entry.

After the questionnaires arrive at head office for processing, a clerical pre-grooming process known as Pre-Scan is performed. In this process, a clerk scans each questionnaire for response irregularities such as unreadable entries, ditto marks, and responses in incorrect locations. If valid responses can be discerned, they are recorded in the appropriate locations, if not, the questionnaire is left unchanged.

Next, in Key Entry, the data on each questionnaire are keyed into the computer. Identifying information from the front page of the questionnaire is entered in a standard fixed format. However, since farm operators are required to answer only the sections of the questionnaire that apply to their holding, a large portion of the questionnaire remains blank. To reduce keying time, a method known as "string-keying" is used to enter the remaining data. This means that the field name is keyed, immediately followed by the data value for that field. Only fields with existing data values are keyed; unanswered portions of the questionnaire are not. Because of the sparseness of the data, this method results in significant savings in keying time required.

The Key Entry process creates one Edit and Imputation Master File (EIMF) record for each of a total of approximately 320,000 questionnaires. There are 244 fields on an EIMF record, each identified by a name, generally 6 characters in length. The Key Entry operator is instructed to key "#" for any unreadable entries. If possible, a clerical correction will be performed on records containing this symbol during Edit, otherwise, the records will be corrected during imputation.

3. EDIT

The Edit phase serves two purposes. The first is to use computer edits to detect any inconsistent, missing, or suspicious entries in the data. The second is to perform a clerical correction on the defective records, or if that is not possible, then to pass the defective records on to be fixed during Imputation. A flow chart of the Edit process is given in Figure 2.

There are 3 components to the edit system: two computer edit cycles called Correction Cycles #1 and #2, and a cycle for correcting edit failures, called Correction of Rejects. Correction Cycle #1 (CC #1) consists of those edits that detect conditions that prevent the "de-stringing" (the conversion from string format to fixed format) of the keyed record (decode edits), and those edits that detect errors in the geographic and identifying information from the front page of the questionnaire (ID edits). Correction Cycle #2 (CC #2) consists of those edits that identify inconsistencies in the main body of the data (data edits). Correction of Rejects is a clerical process during which both CC #1 and CC #2 edit failures are corrected manually. Edit failures that cannot be corrected by Correction of Rejects are passed on to Imputation.

Each of the EIMF records is processed through the edit system individually.

3.1 Correction Cycle #1 (Decode and ID Edits)

Correction Cycle #1 consists of the application and resolution of two sets of edits: the decode edits and the ID edits.

The decode edits are applied first and if conditions exist that prevent the "de-stringing" of the data record, then decode edit failures will result. For example, as no two fields should have the same identifying characters, "de-stringing" will be prevented if two field names are keyed identically.

Any failed decode edits are resolved manually by the Correction of Rejects staff. This involves returning to the questionnaire to determine the cause of the edit failure, then the rekeying of the relevant data. After an attempt is made to resolve a decode edit failure, the EIMF record is re-edited by passing it through the decode edits again, forming a continuous cycle between the decode edits and the Correction of Rejects staff. This cycle is repeated until there

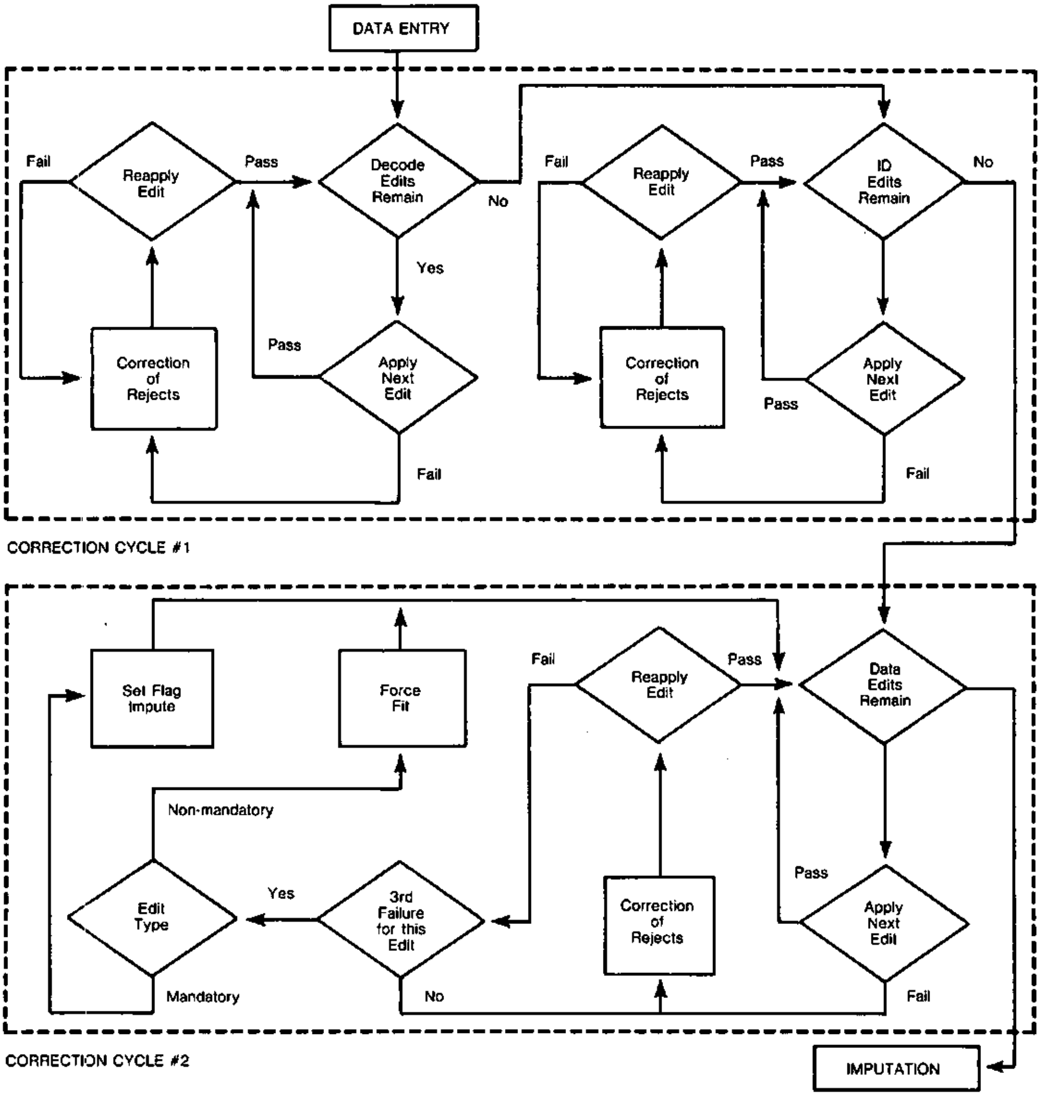


Figure 2. Edit Process Flow

are no decode edit failures remaining on the EIMF record. If a decode edit cannot be resolved directly, the most appropriate valid interpretation of the available data is employed as a final override.

After all decode edit failures have been resolved, the ID edits are applied. If any of the identifying information on the EIMF record is inconsistent or missing, then one or more ID edits will fail. These ID edit failures are resolved in an identical manner to the decode edits.

Once all of the CC #1 (decode and ID) edit failures have been resolved by the Correction of Rejects staff, the EIMF record is passed through the CC #2 edit program.

3.2 Correction Cycle #2 (Data Edits)

The data edits (CC #2) are used to detect errors in the main body of the questionnaire, as opposed to errors in coding, or in identifying information. There are two types of data edits: non-mandatory edits (75), and mandatory edits (24).

Non-mandatory edits are written to detect suspicious entries on the EIMF data records. Generally, non-mandatory edits, detecting variable values falling outside prescribed limits, are performed by comparing different fields or groups of fields on the questionnaire to determine if some data values are abnormally high or low in comparison with others. For example, a record with total farm area equalling 10 acres and containing 10,000 cattle would be flagged by a non-mandatory limit edit.

Mandatory edits are written to detect logical impossibilities on the data record, e.g., if the total number of cattle reported is not equal to the sum of the reported values for each of the different cattle types, then a mandatory edit would fail. The most complex mandatory edits are those written for the crop section of the questionnaire.

To resolve a non-mandatory edit failure, the record is sent to a Correction of Rejects clerk. The Correction of Rejects clerk first notes whether or not the edit failure is due to a keying error. If it is, the relevant data is rekeyed. If it is not, the clerk scans the questionnaire to see if the respondent has written any comments on the questionnaire that may explain the reason for the edit failure. For example, if the respondent is instructed to answer a question in tons, and tons has been crossed out and pounds written in, the response will probably fail a non-mandatory limit edit. In this case, the Correction of Rejects clerk will convert the response from pounds into tons. If the Correction of Rejects clerk can find no explanation for the edit failure, the respondent's answers are left intact on the EIMF record and are indicated acceptable. Although no changes are made to the data on the EIMF record, this is known as "force-fitting" the data.

Mandatory edit failures are handled somewhat differently to non-mandatory edit failures. To resolve a mandatory edit failure, the failed record is sent to a Correction of Rejects clerk who proceeds at first in an identical manner to that used in the resolution of non-mandatory edit failures. However, if no explanation for the edit failure can be found, instead of "force-fitting" the edit failure, the record is flagged for computer imputation.

As in CC #1, there is a continuous cycle between the Correction of Rejects staff and the CC #2 edit program. After each attempt is made to resolve a CC #2 edit failure the EIMF record is re-run through the CC #2 edit program. Unlike CC #1, however, the Correction of Rejects clerk has only 3 attempts to resolve the CC #2 edit failures on a given EIMF record. After the third attempt, the CC #2 edit program is run once again. Any remaining non-mandatory edit failures are marked "force fit" and any remaining mandatory edit failures are marked "impute". The mandatory edit failures are simply flagged at this stage. The particular fields requiring imputation are identified at the imputation stage.

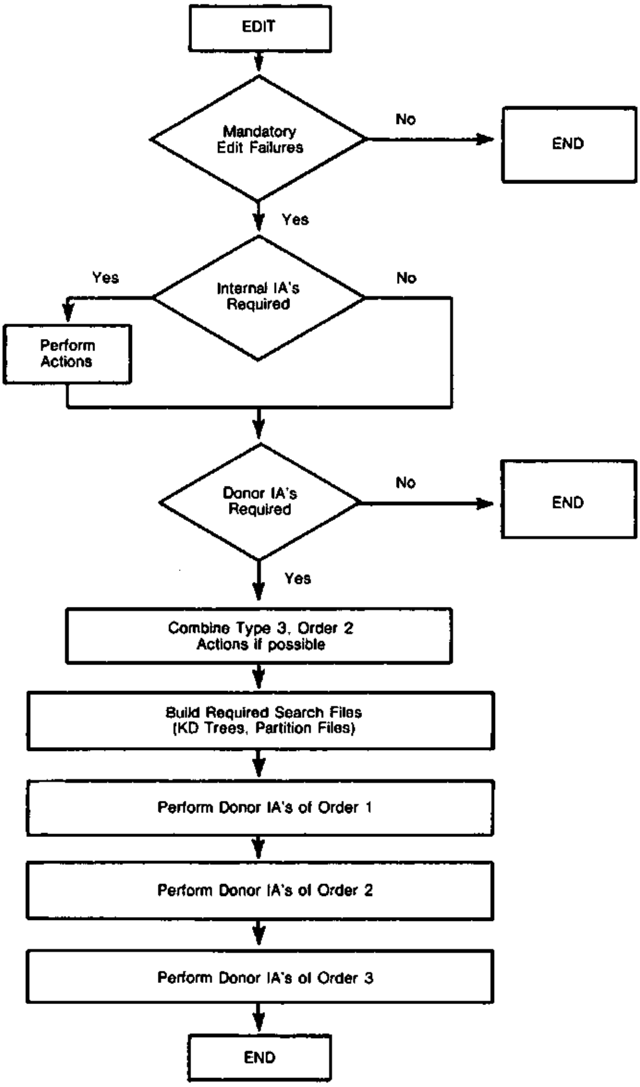

EDIT
Mandatory Edit Failures
No
END
Yes
Internal IA's Required
Yes
No
Perform Actions
Donor IA's Required
No
END
Yes
Combine Type 3, Order 2 Actions if possible
Build Required Search Files (KD Trees, Partition Files)
Perform Donor IA's of Order 1
Perform Donor IA's of Order 2
Perform Donor IA's of Order 3
END

Figure 3. Imputation Process Flow

4. IMPUTATION

The purpose of the 1981 Census of Agriculture imputation system (see Figure 3) is to resolve edit failures on the EIMF data records. As all non-mandatory edit failures are "force-fit" as described in the previous section, only the mandatory edit failures remain to be resolved by the imputation system. In order to make the EIMF data records conform to the mandatory edits, specified "imputation actions" are performed. These imputation actions (IA's), of which there are over 100, are designed so that as few fields as possible are changed on the EIMF record, e.g. totals are always adjusted to equal the sum of the parts, rather than the parts being adjusted to total the sum. Each IA has associated with it the appropriate imputation processing control information and is selected based on the field or fields requiring imputation. There are two different types of IA's performed: internal IA's, or deterministic corrections, and donor IA's.

4.1 Internal Imputation Actions

Internal IA's are performed in cases where sufficient data exists on the failed record to enable the imputation system to provide a deterministic correction for the inconsistent field(s). These internal IA's are performed in cases where the inconsistent field(s) is (are) deterministically dependent on other fields not requiring imputation. For example, an internal IA would be performed if a respondent reports quantities for the various types of cattle but neglects to report the total number of cattle. In this case, total cattle would be calculated using the sum of the quantities reported for the various types of cattle. Another situation in which an internal IA would be performed is where a respondent reports a certain quantity of a particular type of fruit tree but neglects to give the corresponding acreage. In this case, the acreage would be computed using a predetermined average density for that type of fruit tree. Internal IA's are performed in accordance with constraints to ensure that the imputed values are within reasonable bounds.

The implementation of internal IA's is more straightforward than that of donor IA's. As the internal IA is performed using data from the same record, there is no need to specify an algorithm for donor selection. The only requirement is to perform the deterministic correction specified by the appropriate internal IA. All internal IA's are performed before proceeding to donor imputation.

4.2 Donor Imputation Actions

When the inconsistent field or fields are not deterministically dependent on other consistent fields, internal IA's cannot be applied. The lack of sufficient information on the failed record to provide a deterministic correction to the inconsistent field(s) necessitates an imputation method using data contained on another record. This method, known as donor imputation, involves the transfer of data from a "clean" donor record (one which has passed all mandatory edits) to the failed record. The transferred data will restore consistency to the inconsistent field(s) on the failed record. For example, a donor IA will be performed in order to estimate the distribution for types of cattle when only the total number of cattle is reported. In this case, the distribution of cattle types present on the donor record is transferred to the failed (recipient) record.

As donor imputation requires an algorithm for locating a donor record, it is more complex to implement than internal imputation. In order to perform donor imputation, several search "parameters" must be specified.

To ensure that a "clean" donor record is geographically close to the "bad" recipient record, the country is divided into distinct geographical regions called imputation regions. The delineation of these imputation regions is based on the existing "crop district" boundaries which are defined according to characteristics such as soil type and climate. There are 59 crop districts, and thus 59 imputation regions, in Canada with an average of 5,500 farms per region. In order to be an eligible donor, a record must be in the same imputation region as the recipient record.

In order to avoid searching records that cannot donate suitable data, each donor IA also specifies the subpopulation on which the donor search is to take place. For example, if the distribution for types of cattle is being imputed, then the only records searched in order to find a donor would be members of the subpopulation where cattle have been reported. A given record may be a member of several of the 30 different subpopulations. In some cases, all clean records within the imputation region are deemed suitable donors in which case the general population in the imputation region is defined as the appropriate subpopulation.

The final constraint on the file of eligible donors is the fact that records requiring any donor imputation themselves cannot be used as donors. However, records requiring only internal imputation may be used as donors.

In summary, the file of eligible donors consists of all records not requiring donor imputation that are members of the subpopulation specified by the imputation action to be performed and that are also located in the same imputation region as the bad record.

As some records require more than one IA to be performed, there is need for a hierarchical system of imputation action execution. To specify the order in which the IA's are to be performed, every IA, both internal and donor, has one of three "orders" associated with it. IA's of order 1 are performed first, followed by IA's of orders 2 and 3 respectively.

To aid in the selection of a suitable donor record, one or more variables not requiring imputation are selected to be used as matching variables for each donor IA. These matching variables, selected by subject matter experts, are considered to be highly correlated with the field(s) requiring imputation. Both the recipient and the selected donor record should have similar matching variable values. As the use of continuous matching variables does not permit exact matches, a distance function based on the selected matching variable(s) is used to identify the closest eligible donor to the bad record.

Each donor IA has one of three possible search types associated with it. Partition searches (type 1) are performed when only 1 discrete matching variable is specified for the IA. Binary searches (type 2) are performed when only 1 continuous matching variable is specified for the IA. Multivariable searches (type 3) are performed when 2 or more continuous matching variables are specified for the IA. Each of these three search types is described individually in the following sections. Other combinations of matching variable types are not employed.

Finally, after a suitable donor has been selected and if specified in the IA control information, the donated data from the donor record are prorated before transferring them to the recipient record. For example, if the variable "number of trucks" is used as a matching variable for imputing "value of trucks", then the value of "value of trucks" assigned to the recipient record is equal to "value of trucks" of the donor, multiplied by the ratio "number of trucks" of the recipient divided by "number of trucks" of the donor.

As previously described, each donor imputation action has one of three search types associated with it. Two of these search types, binary and partition searches, are used to perform imputation actions for which only 1 matching variable is specified. The other search type, the multivariable search, is performed when 2 or more continuous matching variables are to be used.

4.2.1 Type 1 — Partition Searches

Partition Searches are performed when only 1 discrete matching variable with a small number of possible values is specified for the imputation action, e.g., as in the case where a respondent reports the total number of tractors, but neglects to give the corresponding total dollar value. Since a farmer is unlikely to have more than 3 tractors the donor population is divided into 3 partitions: 1, 2, or 3+ tractors. A donor is chosen at random from the partition to which the recipient record belongs. If there are no donor records within the partition to which the recipient record belongs, but there are donors in any of the subsequent (higher numbered) partitions, then all of the subsequent partitions are collapsed into one and a donor record is selected at random from this collapsed partition. If there are no donor records in the partition to which the recipient record belongs or in any subsequent partition, then a donor record is selected at random from the closest preceding (lower numbered) partition that contains any donor records. As these collapsing procedures are not frequently applied, no serious introduction of bias is encountered. If the donor population is empty, then the field to be imputed is assigned the maximum value allowable by the edits and the record flagged to indicate that imputation was unsuccessful. These flagged records are then reviewed by subject matter personnel who manually assign an appropriate value to the field requiring imputation.

4.2.2 Type 2 — Binary Searches

Binary searches are performed when only 1 continuous matching variable is specified for the imputation action, e.g., as in the case where a respondent reports the total value of his/her tractors, but does not give the corresponding number of machines. The entire file of eligible donor records is searched and the record that minimizes the difference between the matching variable values is selected as the donor. If two or more potential donor records are equally close, then the one that is geographically closer to the recipient (as judged from the geographic ID) is automatically selected as the donor. If the donor population is empty, then the recipient record is flagged to indicate that imputation was unsuccessful.

4.2.3 Type 3 — Multivariable Searches

Multivariable searches are performed when more than one continuous matching variable are specified for the imputation action. These are the most complex of the three search types performed by the 1981 Census of Agriculture. The method used to perform multivariable searches was adapted for use at Statistics Canada by G. Sande.

When the missing data are related to more than one continuous matching variable, it is desirable to use as a donor a record that is closest to the recipient record on all these matching variables simultaneously. This requires a multivariable search on a large donor file and has been made practical by grouping the donor population in such a way that it is not necessary to search every donor to determine the closest. This specialized grouping of records is called the K-D (Key Discriminator) tree. The same K-D tree may be used for all records requiring a certain donor IA within a particular imputation region as the file of eligible donors will remain the same in each case. However, if a different donor IA is to be performed using a different donor population, or even the same donor IA on a different imputation region, a new K-D tree must be built as the file of eligible donors will not contain the same records.

a) Building the K-D Tree

The first step in the building of the K-D tree is to perform a transformation on all of the matching variables by subtracting the mean and dividing by the standard deviation of the donor population. This allows matching variables of different scales to be specified for the same search.

After the variable transformation, the following algorithm is then used to actually build the K-D tree. It is first applied to the entire file of eligible donors, and then to all subfiles subsequently created by the algorithm.

Firstly, the range (largest value minus smallest value) is calculated for each of the matching variables specified. The median value of the variable with the largest range (or the variable with the smallest ID if there are 2 or more with the maximum range) is then calculated. The variable for which the median is calculated is called the discriminator variable. This median value is used to split the file into 2 new subfiles, the left subfile containing records with values less than or equal to the median value of the discriminator variable, and the right subfile containing records with values greater than the median value of the discriminator variable. The algorithm is then progressively re-applied to the resulting subfiles using all specified matching variables until all files become TERMINAL, at which point the building of the K-D tree is complete. A subfile becomes TERMINAL when either the range equals zero for all matching variables, i.e., all records in the subfile are identical, or if there are 16 or less records in the subfile.

The above algorithm will yield a K-D tree of the form illustrated in Figure 4.

Every record contained in the original file will be present in one and only one of the subfiles corresponding to the terminal nodes.

b) Searching thė K-D Tree

In order to locate the best possible donor, it is necessary to decide which of the terminal nodes "corresponds" to the recipient record. This is done by traversing the K-D tree, using the transformed matching variable values of the recipient record, starting with the root node and proceeding until one of the terminal nodes is reached. At each node of the tree it is determined, using the discriminator variable for that node, which of the two lower nodes the recipient record corresponds to. The K-D tree is traversed in this manner until a terminal node is reached.

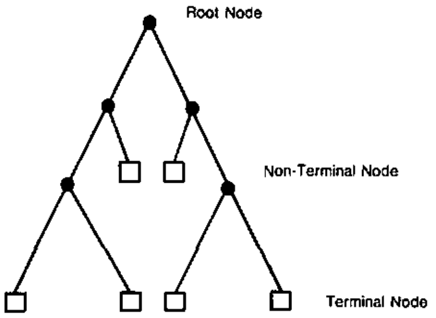


Figure 4. General Form of K-D Tree

In order to determine which donor in the chosen terminal node is closest to the recipient record, a distance function is required. Because of its ease of implementation, the distance defined by the maximum of the absolute differences between matching variables was used. The selected donor record is the one that minimizes this "distance".

Although the selected donor record is the closest to the recipient record contained in the chosen terminal node, it is possible that there are closer donor records residing in other

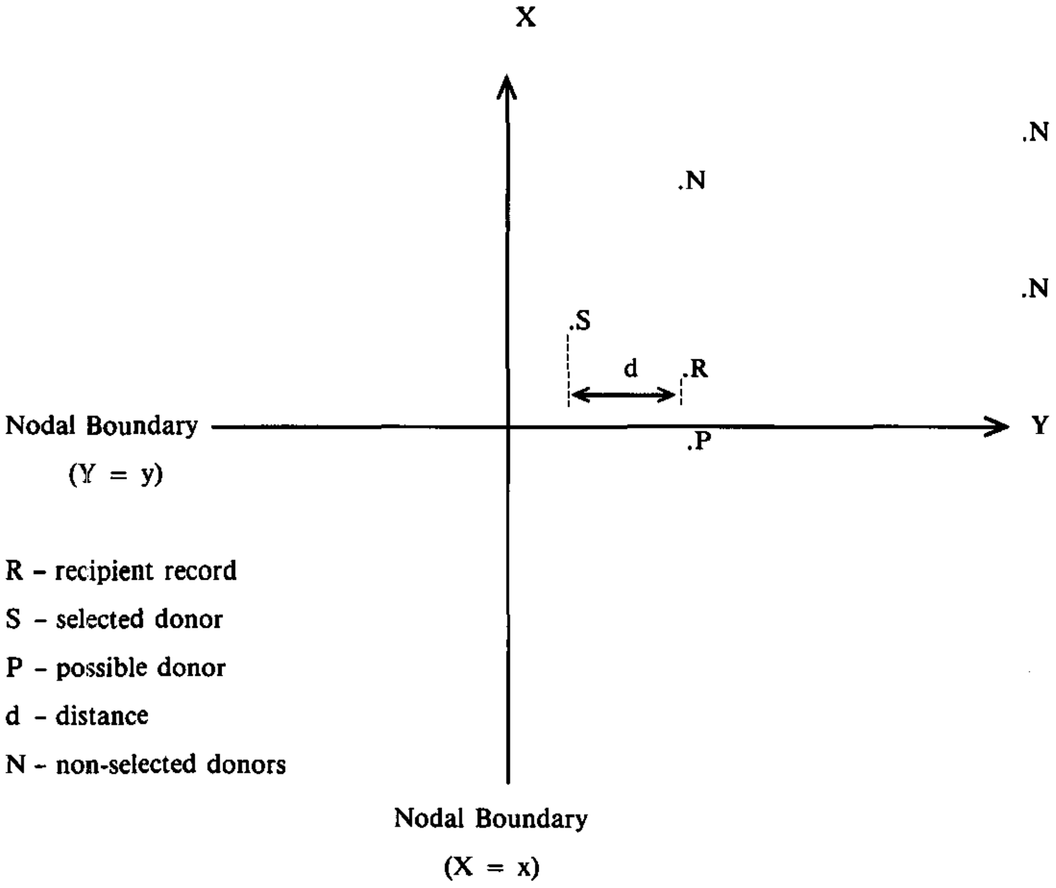


Figure 5. Closer Donors From Other Terminal Nodes (two matching variables)

terminal nodes. This may occur only if a nodal boundary exists that is closer to the recipient record than the currently selected donor record. This case is shown in Figure 5 for a donor IA involving two matching variables; X and Y. Each quadrant represents a terminal node.

It is evident that the possible donor P is closer to the recipient R than the selected donor S. This is possible because R is closer to the position of the nodal boundary Y = y than to S, and only donor records lying in the same terminal node as the recipient record may be selected.

A procedure, based on the variable values used to define the nodal boundaries and known as the bounds-overlap-ball (B.O.B.) test, is used to determine which of the other terminal nodes, if any, may contain donors closer to the recipient record than the selected donor record. Only terminal nodes that have the potential to provide closer donors are tested, and if a closer donor is found, then it replaces the previously selected donor. The B.O.B. test is applied until all nodes that may contain closer donors have been tested.

Finally, for all three search types, after the eventual donor record has been selected, the donated data values are prorated as previously described, if specified in the IA control information.

It will always be possible to select a donor unless the donor population is empty. If this occurs then the imputation region is collapsed with another and imputation is redone. It was never necessary to perform this operation in 1981.

5. CONCLUDING NOTE

A detailed evaluation, Grenier (1983), indicated that a major portion of the edit system was of little data quality benefit. This was because the Correction of Rejects procedures were unable to correct a sufficient proportion of the edit failures. For example, Correction of Rejects was unable to correct the failures resulting from a subset of 77 of the 97 edits more than 5% of the time. Also, many of the edits affected less than .1% of the population. Additionally, the Correction of Rejects procedures were highly labour intensive and created a heavy paper burden. To eliminate these inefficiencies a new computer edit system will be designed for 1986.

Statistics from the 1981 Census of Agriculture, Grenier (1983), indicated that 43% of the farms in Canada had at least one field imputed. Of this 43%:

18% required internal imputation only,
17% required donor imputation only, and
8% required both internal and donor imputation.

An analysis of the data distributions before and after imputation indicated that the imputation system did not have a serious impact at the Canada level although many of the 137,390 records imputed underwent a significant change. The system successfully handled all necessary imputations with only 58 records requiring manual imputation. The system was found to be very efficient, a processing cost of only $15,000 being incurred. Diagnostic data indicated that minor modifications to the system must be made for greenhouses, mushroom houses, community pastures, and institutions, if they are to remain in the census. Due to its successful fulfillment of the requirements, it is planned to reuse the present imputation system in 1986.

REFERENCES

SHIELDS, M., and YIPTONG, J. (1981). Census of Agriculture-1981 Imputation Specifications. Technical Report, Census and Household Survey Methods Division, Statistics Canada.

GRENIER, A.R. (1983). 1981 Census of Agriculture Evaluation Report. Technical Report, Agriculture Statistics Division, Statistics Canada.

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 (8½ × 11 inch), one side only, entirely double spaced with margins of at least 1½ 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(·)" and "log(·)", 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, ω; o, O, 0; l, 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.