Statistics Canada Statistique Canada

SURVEY METHODOLOGY

A JOURNAL
OF
STATISTICS CANADA

VOLUME 13, NUMBER 1
JUNE 1987

Canada

SURVEY METHODOLOGY

A JOURNAL OF STATISTICS CANADA

JUNE 1987

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

November 1987

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

Payment to be made in Canadian funds or equivalent

Catalogue 12-001, Vol. 13, No. 1

ISSN 0714-0045

Ottawa

SURVEY METHODOLOGY

A Journal of Statistics Canada

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

EDITORIAL POLICY

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

Submission of Manuscripts

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

Subscription Rates

The price of the Survey Methodology Journal (Catalogue No. 12-001) is $20.00 per year in Canada, $23.00 per year for other countries (payment to be made in Canadian funds or equivalent). Subscription order should be sent to: Publication Sales, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6. A reduced price of US $10.00 ($14.00 Can.) is available to members of the American Statistical Association, the International Association of Survey Statisticians, and the Statistical Society of Canada. Please subscribe through your organization.

SURVEY METHODOLOGY

A Journal of Statistics Canada
Volume 13, Number 1, June 1987

CONTENTS

Survey Methodology, June 1987
Vol. 13, No. 1, pp. 1–14
Statistics Canada

Telephone Sample Designs for the U.S. Black Household Population[1]

KATHRYN M. INGLIS, ROBERT M. GROVES, and STEVEN G. HEERINGA[2]

ABSTRACT

The two-stage rejection rule telephone sample design described by Waksberg (1978) is modified to improve the efficiency of telephone surveys of the U.S. Black population. Experimental tests of sample design alternatives demonstrate that: a) use of rough stratification based on telephone exchange names and states; b) use of large cluster definitions (200 and 400 consecutive numbers) at the first stage; and c) rejection rules based on racial status of the household combine to offer improvements in the relative precision of a sample, given fixed resources. Cost and error models are examined to simulate design alternatives.

KEY WORDS: RDD samples; Telephone surveys; Rare population samples.

1. INTRODUCTION

Surveys of rare populations lacking special frames often entail large per-unit costs relative to similar designs for the full population. When the rare population is a small subgroup of a readily identifiable population, the sample of that subgroup is often obtained by screening the larger population. Household surveys of demographic subgroups such as the U.S. Black population typically use such screening to locate eligible sample units; however, extensive screening to identify a rare population sample results in high costs per interview. In recent years telephone-sampling methods have been proposed as cost-efficient tools for sampling and interviewing rare populations. The cost of telephone interviewing is often less than face-to-face interviewing (Groves and Kahn 1979), and when screening is required to identify an eligible respondent, the cost-efficiency of telephone interviewing becomes even more marked. Still, the screening costs of telephone surveys of rare populations can be high in absolute terms.

This paper presents ways in which the screening method for telephone surveys can be refined to reduce costs while achieving desired levels of precision. In this paper we examine a variety of telephone sample designs for the U.S. Black household population. The telephone survey experiments described in this paper were conducted as part of a study of Black political attitudes and electoral behavior in the 1984 U.S. presidential election.

The use of telephone sampling and interviewing implies that Blacks living in households without telephones (about 15 percent of the U.S. Black household population) are not covered by the survey procedures. Such persons tend to be poorer and younger than those living in households with telephones (Thornberry and Massey 1983). To the extent that Blacks without telephones have attitudes and voting behaviors that are different from those with telephones, the survey estimates would differ from Black household population parameters. While not wanting to discount noncoverage error associated with telephone surveys of the Black population, this paper focuses on differential cost efficiencies and sampling error that might result from alternative approaches to telephone samples of Black households.

[1] Revision of paper presented at the 1985 American Statistical Association meetings. Research was partially supported by the U.S. Bureau of the Census and the Survey Research Center. The discussion does not necessarily represent the views of those organizations.

[2] Kathryn M. Inglis, McNair Anderson and Associates, Australia. Robert M. Groves and Steven G. Heeringa, Survey Research Center, University of Michigan, Ann Arbor, Michigan, 48106-1248, United States.

The telephone sample designs presented here are extensions of a design described by Waksberg (1978). That random digit dialing (RDD) design (commonly referred to as the Waksberg-Mitofsky design) is a two-stage cluster sample of telephone numbers. U.S. telephone numbers contain 10 digits, a three-digit area code, a three-digit central office code or "prefix", and a four-digit suffix in the range 0000-9999 (e.g., 313-764-4424). At the primary stage, a stratified sample of 10-digit telephone numbers is randomly generated, and each such "primary number" is linked to a block of 100 consecutive numbers (e.g., 313-764-4424 would be linked to the "100-series", 313-764-4400 to 313-764-4499). For household surveys, if the primary number is found to be a working household number, then its cluster of 100 consecutive telephone numbers is retained at the first stage for further sampling. If not, its "100-series" is discarded. Therefore, the probability of selection of a first stage 100-series is proportional to the number of working household numbers in that 100-series. In the second stage of sampling, equal numbers of working household numbers are selected from each of the 100-series retained at the primary stage. Therefore, the second stage sampling of households is performed with conditional probabilities of selection inversely proportional to the number of working household numbers in the 100-series. Thus, the design yields an equal probability (epsem) sample of household numbers, and clusters them so that the proportion of total numbers selected which reach households is higher than that obtained by a stratified random RDD sample. To clarify the discussion here, we refer to the 100-series banks of consecutive numbers as the primary stage unit (PSU) of the two-stage RDD design. The term "cluster" is reserved for the fixed set of working household numbers that is selected from the PSUs at the design's second stage.

In this research the sample design modifications aimed at reducing screening costs take three forms: a) stratification of telephone exchange units by proportion Black, and disproportionate allocation of the sample to high density Black strata; b) use of two-stage rejection rules based on both residential status and race of the household; and c) increase in PSU size (from 100 consecutive numbers to 200 and 400).

Stratification of the telephone population by race attempts to isolate exchange areas with high proportions of telephone subscribers who are Black. Higher sampling fractions are then applied to those strata, relative to strata with lower proportions Black. Under this disproportionate sample design, the total number of households that have to be contacted in order to obtain one interview with an eligible Black household is smaller than that for an epsem sample of the household population. Consequently, the screening costs for locating a sample of Black households are reduced. In telephone samples, the basic geographical unit for stratification is the wire center or telephone exchange, to which one or more three-digit prefixes (central office codes) may be assigned. In general, no counts of the subscriber population by racial characteristics are available for these sampling units. Thus, proxy indicators of high density Black exchanges must be used. The experiments described in this paper examined the value of such proxy indicators.

Blair and Czaja (1982) present an alteration of the Waksberg-Mitofsky RDD design which incorporates two-stage rejection rules based on both residential status and race eligibility of the household. For the Black population this method includes, at the first stage, only 100-series whose primary number was assigned to a Black household and then samples a fixed total of Black household numbers within those PSUs. In a U.S. national sample survey, Blair and Czaja found that using this design, the percentage of Black households among all household numbers chosen increased from 9 percent for the first stage to 25 percent for the second stage numbers. Given the compensating probabilities of selection in the two stages, this epsem design greatly reduces the level of screening required to obtain any given sample size of Black households. A similar alteration of the rejection rules for the two-stage Waksberg-Mitofsky design was employed in the experiments described in this paper.

In the Blair and Czaja design some of the primary stage 100-series contained too few Black household numbers to yield the number of elements per cluster required (10 in their case) for an epsem sample of Black households. In addition, relatively large screening costs are incurred at the first stage of selection for this design; over 44 primary numbers must be dialed to locate one Black household. The joint solution to these two problems is to both increase the size of the PSU and to select larger numbers of second stage elements per PSU. The analyses reported here examined the use of primary stage units of 100, 200, and 400 consecutive numbers each. The extension of the PSU definition beyond the standard 100 consecutive numbers was suggested by observations on the assignment of telephone numbers within prefixes. The following appears to be the most common pattern: 1) almost all household numbers within a prefix serve units located within the geographical boundaries of the exchange; 2) there is little geographical clustering of assignments within exchanges (i.e., neighbors do not tend to have consecutive telephone numbers, nor need they have numbers in the same prefix); and 3) there is more diversity in the percentage of household numbers among 1000-series than among 100-series within the same 1000-series of numbers. These impressions are the result of several years of household telephone sampling at the Survey Research Center. Observations 1) to 3) suggest that the expansion of the PSU definition from 100 consecutive numbers to a larger number might permit the use of larger clusters of secondary numbers with little reduction in the proportion of those numbers which are Black households.

2. THE PILOT STUDY

In two integrated experiments imbedded in a pilot survey, several design alternatives were tested. One purpose of the pilot study was to examine the ability of stratification based on civil government units, with only rough correspondence to telephone exchanges, to isolate sets of telephone numbers densely filled with black household numbers. For this, three strata of exchanges were defined:

1. "High density"- Exchanges corresponding to the central cities of large Standard Metropolitan Statistical Areas (e.g., Chicago city, for the Chicago SMSA). This identification was based on the name of the telephone exchanges in these areas.
2. "Medium density"- All other exchanges in selected southern states (Virginia, North Carolina, South Carolina, Florida, Georgia, Alabama, Mississippi, Louisiana). The vast majority of exchanges lie in only one state; those serving two states were associated with the state given in the exchange name.
3. "Low density"- The balance of exchanges in the coterminous United States.

An equal probability sample of 1400 six-digit area code/central office code prefix combinations was then systematically selected from the 34,389 such combinations listed as active on a frame which can be purchased from American Telephone & Telegraph (AT&T). Four-digit random numbers were appended to each selected six-digit stem to yield a sample of 1400 ten-digit primary numbers.

The results of the pilot study demonstrated that the three strata had vastly different proportions of Black telephone numbers. The low density stratum was found to require over six times as much screening to locate a black household as was required in the high density stratum. (This result was confirmed with more precision in the production study, discussed in the next section).

Another purpose of the pilot study was to test the use of rejection rules based on racial composition and working household status of sample numbers from PSUs of differing size. To provide increased precision in analyses related to this objective, an additional 500 primary numbers were selected from the high- and medium- density strata. The 1900 primary numbers in the combined pilot study sample were then dialed and screened for their Black household status. If the sampled primary number reached a Black household, it simultaneously identified three different PSUs. As shown in Table 1, every individual number can be viewed as belonging to a single 100-series, a single 200-series, and a single 400-series. For example, the number 313-764-4424 is a member of the 4400-4499 100-series, the 4400-4599 200-series, and the 4400-4799 400-series. To test the feasibility of expanding the PSU size, the pilot study sampled secondary numbers from each of these three hundred series. The second stage cluster sizes of Black households were set at 3 for the 100-series of the primary number, 6 for the 200-series, and 9 for the 400-series clusters. In both the primary and secondary stages of selection, if the race of the household was not known, it was assumed to be a non-Black household.

Table 1 presents the disposition of the secondary numbers by PSU type and stratum. Of most interest is the proportion of secondary numbers assigned to Black households for the different PSU definitions. For the 100-series, .134 of all secondary numbers are Black household numbers. This implies that .223 of the households sampled were Black, compared to the .25 Black households found by Blair and Czaja. For the 200-series PSUs, .124 of all secondary numbers are Black household numbers. For the 400-series, .115 of all second stage sample telephone numbers are assigned to Black households. These proportions are all within sampling error of each other (the standard error of each estimate is at least .02). That is, no significant decrease in the proportion eligible was observed when the PSU definition was expanded from 100 to 400 consecutive numbers. These rates imply that while 100-series PSUs on the average can support second stage clusters of 13 or 14 sample Black households, the 400-series might on the average support cluster sizes of 46 sample Black households. The ability to increase the Black household cluster size at the second stage of sampling enables the researcher to greatly reduce sample screening costs.

Table 1 also compares the proportion of eligible secondary numbers for PSUs sampled from the three different strata used in the pilot study. For all the PSU definitions (100, 200, 400) the same result applies — the large SMSA telephone exchanges in the high Black density stratum offer close to a doubling of the eligibility rate when compared to the rate for the overall population (.21 versus .12 or .13). The medium density stratum, consisting of non-SMSA exchanges in selected Southern states, has eligibility rates below that of the nation as a whole (between .08 and .10). The low density stratum, the remainder of the country, also has lower than average eligibility rates (between .07 and .085). Since the high density stratum covers about 36 percent of the Black household population with telephones, the chosen stratification, in combination with disproportionate allocation of the primary stage samples, is an effective tool for reducing screening costs.

3. THE PRODUCTION STUDY

The production study used the stratification plan that was developed and tested in the pilot study. A disproportionately allocated sample of 11,223 primary numbers was selected from the three Black-density strata using sampling fractions in the ratio 3:2:1 (High:Medium: Low). Although the pilot study found no significant difference in the working household rate for PSUs of 200 and 400 consecutive numbers, a conservative decision was made to use the smaller 200-series PSUs in the production study. The expected second stage cluster size for each PSU was set at 5.5 Black households (not counting the primary number). Primary and secondary stage rejection rules for the modified two-stage Waksberg-Mitofsky design were identical to those used for the pilot study. Since much larger sample sizes were used in the production study, questions about precision and relative efficiencies of the design can be addressed with more confidence.

Table 1

Pilot Study

Disposition of Secondary Numbers Selected within 100-, 200- and 400-Series by Stratum

Stratum and Disposition	Proportion of All Numbers Selected		
	100- Series	200- Series	400- Series*
High Density Black Stratum			
Black Households	.205	.201	.214
Don't Know Race	.028	.029	.032
Non-Black Households	.316	.279	.275
Nonresidential/Nonworking	.451	.491	.479
Number of Cases	(395)	(806)	(1163)
Medium Density Black Stratum			
Black Households	.104	.080	.076
Don't Know Race	.030	.018	.020
Non-Black Households	.494	.443	.420
Nonresidential/Nonworking	.372	.459	.484
Number of Cases	(231)	(560)	(878)
Low Density Black Stratum			
Black Households	.085	.084	.069
Don't Know Race	.014	.028	.027
Non-Black Households	.532	.577	.607
Nonresidential/Nonworking	.369	.311	.297
Number of Cases	(141)	(286)	(491)
Total			
Black Households	.134	.124	.115
Don't Know Race	.024	.025	.026
Non-Black Households	.442	.431	.448
Nonresidential/Nonworking	.400	.420	.411
Number of Cases	(767)	(1652)	(2532)

* Weighted estimate to compensate for the disproportionate allocation of the cluster of 9 secondary numbers across the separate 100-number ranges of the 400-series.

Table 2 presents the results from both the primary and secondary number screening for the production study. The unbiased weighted estimate for an "epsem" two-stage RDD design suggests that 13 percent of all secondary numbers were Black households (the standard error about this estimate is .6 percent). This is in close agreement with the 12 percent secondary number eligibility rate observed in the pilot study. A comparison of the results for the primary stage of selection with those of the secondary stage illustrates the large gains possible by using a two-stage design for telephone sampling of Black households. The gains under the two-stage design are most dramatic in the low density Black stratum where there is nearly a nine-fold increase in the proportion of Black household numbers from the primary to secondary stage (.011 to .090). In the high density stratum the increase is closer to a twofold one (.072 to .190). For the disproportionate allocation design, the unweighted proportions of Black households at the two stages are 3 percent (primary stage) and 15 percent (secondary stage). Comparison of these figures with the estimates for the epsem design (i.e., 2 percent and 13 percent) indicates the reduction in screening achieved by disproportionate allocation.

As in the pilot study, the percentage of Black households varies over the three strata, although the advantage to distinguishing the medium and low density strata is more evident. Across the three strata, the Black household eligibility rate for secondary numbers varies in an approximate 2:1.5:1 ratio. The three strata also differ in the total proportion of secondary numbers that are assigned to residences. The high density Black stratum has larger proportions of secondary numbers assigned to nonresidential units, probably reflecting the urbanization levels of the exchanges in that stratum.

Table 2

Production Study
Disposition of Numbers Selected by Stratum

Stratum and Disposition	Primaries	Secondaries
High Density Stratum		
Black Households	.072	.190
Don't Know Race	.035	.027
Non-Black Households	.219	.352
Nonresidential/Nonworking	.674	.431
Number of Cases	(3,128)	(6,671)
Medium Density Stratum		
Black Households	.032	.141
Don't Know Race	.020	.018
Non-Black Households	.188	.469
Nonresidential/Nonworking	.760	.372
Number of Cases	(1,879)	(2,375)
Low Density Stratum		
Black Households	.011	.090
Don't Know Race	.019	.023
Non-Black Households	.199	.505
Nonresidential/Nonworking	.771	.382
Number of Cases	(6,116)	(3,987)
Estimate for "Epsem Design"*		
Black Households	.021	.129
Don't Know Race	.021	.023
Non-Black Households	.200	.454
Nonresidential/Nonworking	.758	.394
Proportion Black Households for Disproportionate Design	.031	.150
Number of Cases	(11,123)	(13,033)

* Weighted estimates of "epsem design" rates. Weights compensate for disproportionate sampling rates used to select the Production Study sample from the three density strata.

Each PSU of 200 consecutive numbers can be viewed as two half-PSUs of 100 numbers each. Table 3 demonstrates that proportions of nonresidential numbers (.378) found in the half-PSU (100-series) in which the sample primary number fell are lower than in the other half-PSU (.409), but this difference is not statistically significant at the .05 level (standard error about .02). Similarly, the proportion of Black households is somewhat larger in the 100-series of the primary number (.133) than in the adjacent 100-series (.125). Again, this difference is not likely to be found in most replications of the experiment. Table 3 provides another perspective on the results in Table 2, showing only a negligible reduction in the proportion eligible in 100-series adjacent to that of the primary numbers.

The average eligibility rate - proportion of Black households - across PSUs should not be the only criterion for evaluating the sample design. In order to implement an epsem design within strata, each PSU in the design must have a sufficient number of Black households to support the designated number of second stage sample Black households. Thus, the distribution over PSUs of the proportion eligible is also of interest. Figures 1, 2 and 3 contain histograms describing the distribution over all the PSUs of the proportion of Black households by stratum. The stability of the three distributions varies because the number of sample PSUs is about four times greater in the high density stratum than the other two (224 PSUs in the high density stratum to about 60 in the medium and low density strata). The shapes of the distributions, however, appear to be very different for the three strata. The distributions for the low and medium density strata are highly skewed, with 60 percent of PSUs in the medium density stratum and 65 percent of PSUs in the low density stratum having 5 to 20 percent Black households. These eligibility rates correspond to a maximum of 10 to 40 sample Black households for the 200-series PSUs from the low and medium density stratum. In the production study the low density stratum contained several PSUs that would not permit those cluster sizes (6 of the 63 PSUs in that stratum are estimated to have fewer than 10 Black households). The distribution in the high density stratum is much more uniform (4 of the 224 PSUs estimated to have fewer than 10 Black households).

These distributions of percentage Black households by PSU deserve more discussion. Given our current understanding of the assignment of residential numbers to available banks of numbers, there is no reason to believe that within an exchange (or a prefix) there are general tendencies to assign different residential areas to different 100-series. That is, within an exchange serving both Black and non-Black households the hypothesis of assignment of numbers without regard to the race of the subscriber is a strong one. Stated alternatively,

Table 3

Production Study
Disposition of Secondary Numbers by Whether in
Same 100-Series as Primary Numbers

Status	Disposition	
	Same 100-Series as Primary Number	Adjacent 100-Series
Black Households	.133	.125
Don't Know Race	.024	.022
Non-Black Households	.465	.444
Nonresidential/Nonworking	.378	.409
Number of Cases	(6,522)	(6,511)

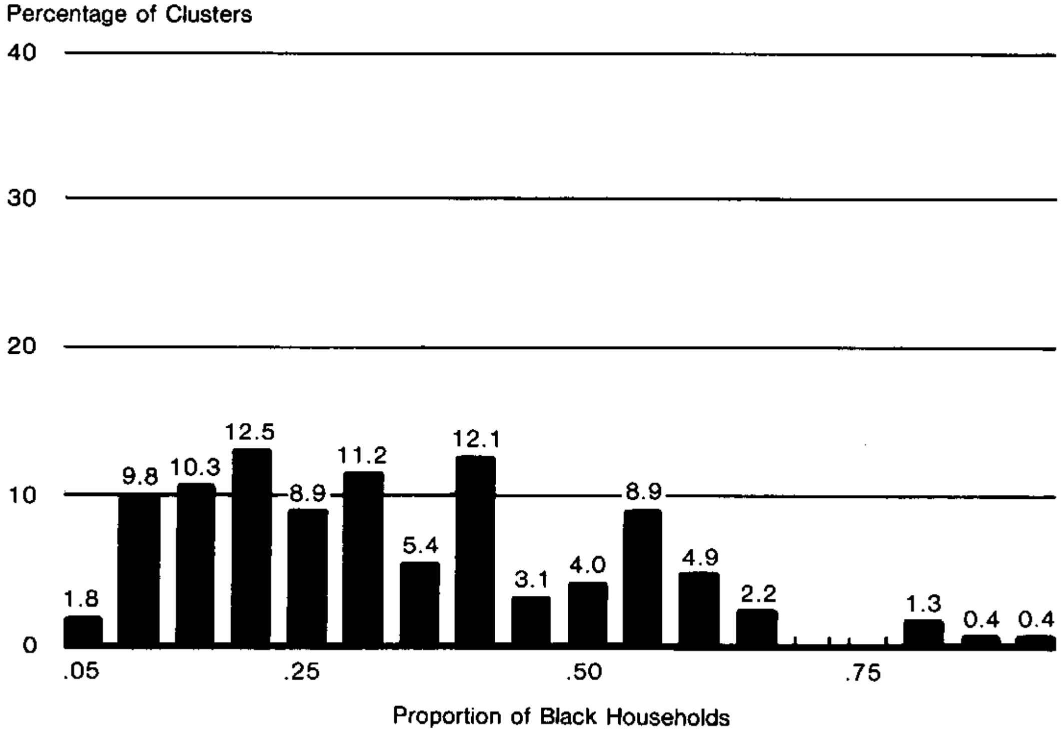

Figure 1. Percentage of High Density Clusters By Proportion of Black Households

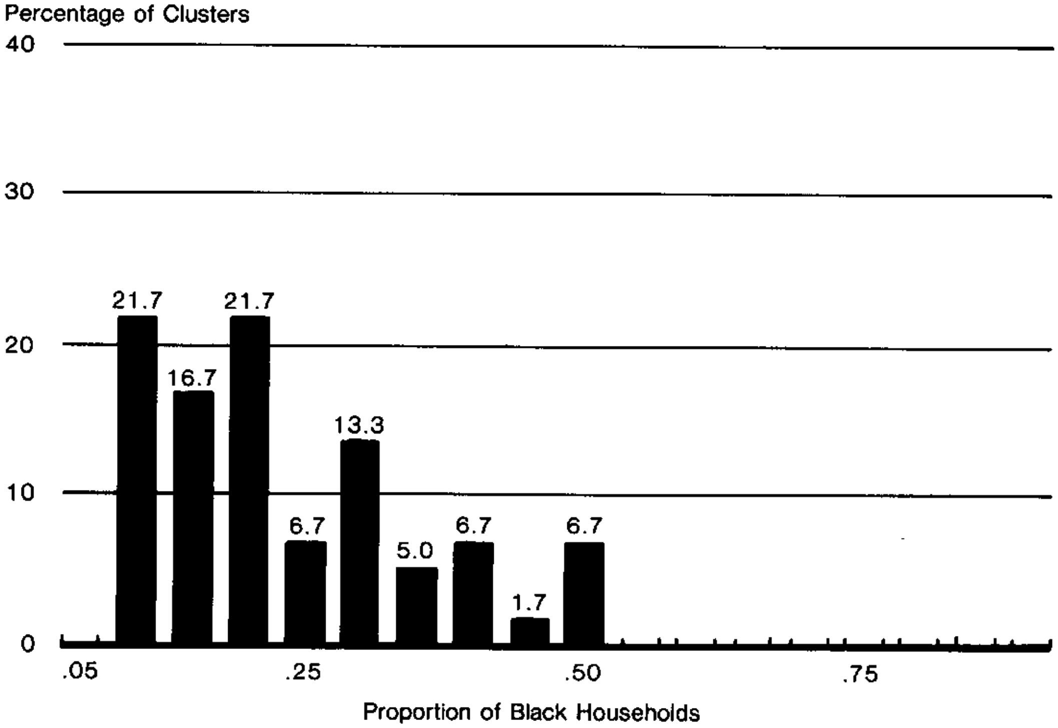

Figure 2. Percentage of Medium Density Clusters By Proportion of Black Households

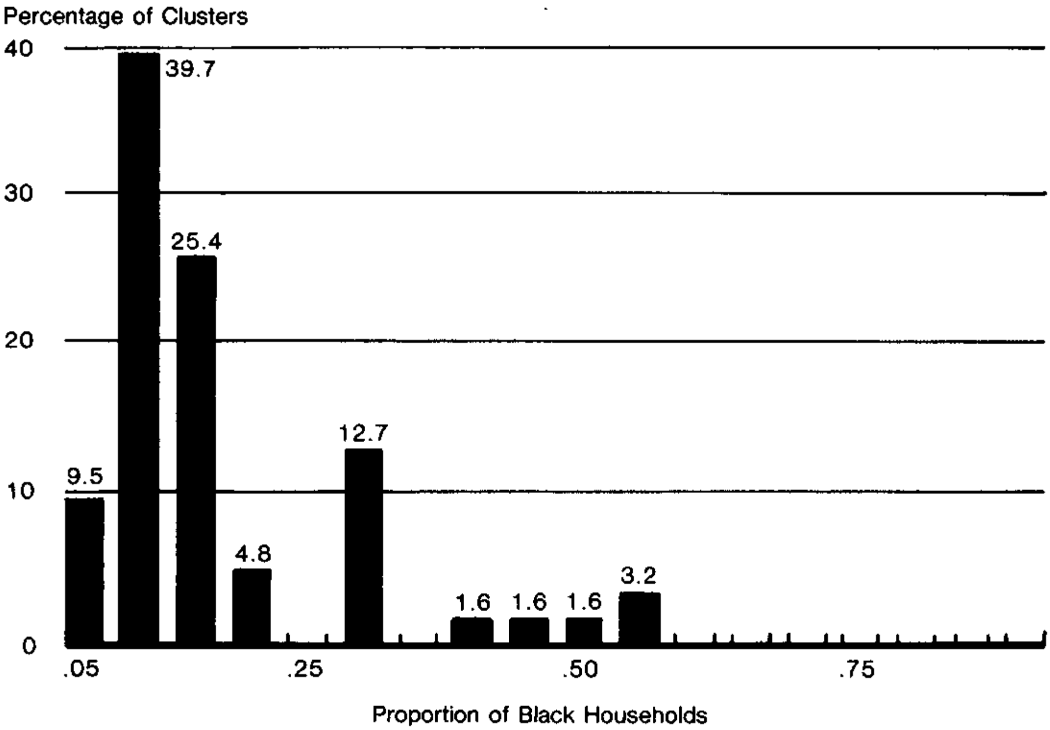

Figure 3. Percentage of Low Density Clusters By Proportion of Black Households

unless the exchanges are subdivided into wire centers that correspond to the residential locations of Black households, there is no *a priori* reason for large amounts of clustering of Black households within 200-series. Following this logic, the more uniform distribution in the high density stratum reflects, we believe, the variability in proportions of Blacks among the telephone populations in the different exchanges in the stratum.

4. SAMPLING VARIANCE PROPERTIES

To achieve greater cost-efficiency in the RDD sampling of Black households it is advantageous to use both large clusters of sample households per PSU (i.e., for a fixed sample size, a smaller number of PSUs) and disproportionate allocation of PSUs to strata of exchanges which vary in their proportion of Black telephone households. While both greater clustering and disproportionate allocation of the sample improve cost-efficiency, the overall precision of the sample is affected by the increased clustering effects and added design effects due to the non-optimal weighting that is required to compensate for the unequal selection probabilities for households from the three density strata. Increased design effects of sample estimates due to non-optimal weighting are described in Kish (1976). The clustering influence on the design effect for the modified RDD procedures is developed in the following paragraphs.

Ceteris paribus, the larger the number of sample elements chosen per PSU the higher the design effect (the ratio of the sampling variance of the given design to that of a simple random sample with the same number of elements). The model often used is $Deff = 1 + \rho(b - 1)$, where $Deff$ is the design effect, ρ is the intracluster correlation for the statistic, and b is the number of sample elements per PSU. Others have shown for many variables on the total U.S. household population that the intracluster correlations for the 100-series tend to be smaller than those generally found in area probability sample clusters (see Groves, 1978). This may not be the case for the Black population for 100-series, and there are no empirical estimates available concerning intracluster correlations for 200-series clusters. The expectation prior to estimating sampling errors was that there would be no change in the intracluster correlations between the 100- and 200-series. This hypothesis reflects the understanding of the assignment of telephone numbers within exchanges that was described above.

Based on sampling errors estimated from the production study data set, the average design effect for a selected set of seven survey statistics is 1.28 for the 100-series and 1.30 for the 200-series. The 100-series average design effect was estimated from those cases which fell into the 100-series of the primary number, while the cases from the entire 200-series were used in computing the average 200-series design effect. Thus, the average cluster size of completed interviews is 2.0 for the 100-series (coefficient of variation, .043) and 3.4 for the 200-series (coefficient of variation, .029). These design effects reflect all the stratification, clustering and weighting in the design and also the fact that the variability in the cluster sizes in the 100-series is greater. (The rejection rule forced an equal number of sample Black households at the 200-series but not necessarily at the 100-series level.) Given that the average design effects for the 100-series and the 200-series are close to one another (1.28 to 1.30), the dominant influence on the sampling variance appears to be non-optimal weighting required by the disproportionately allocated sample design, with little loss in precision due to PSU size alone (moving from the 100- to the 200-series clusters).

Table 4 (page 11) presents the synthetic intracluster correlations by stratum for the seven survey statistics used to compute the estimate of average design effect. The estimates of synthetic intracluster correlations were obtained from the design effect, following Kish's model of $Rho = (Deff - 1)/(b - 1)$, and are unweighted so as to remove the confounding effect of weighting on the synthetic estimates. The estimates in the table tend to be unstable due to the small number of clusters in each stratum, the small average cluster size of completed interviews, and its associated coefficient of variation. These sample design features complicate our inference about clustering effects in the 100- versus the 200-series. Overall, the 100-series estimates of intracluster correlation are somewhat higher than those in the 200-series. We believe that this reflects more an instability in the estimated synthetic correlation than a real difference in clustering effects. We believe that these estimates provide little evidence that there is a change in the intracluster correlation between the 100- and 200-series.

5. OPTIMAL DESIGN FEATURES

The previous sections of the paper address the effect of alternative sample features on cost-efficiency and sampling variance. Survey costs and errors are often combined at the design step to address whether "optimal" features of the survey can be identified. This approach attempts to identify the design which offers minimum variance for a fixed set of resources allocated to the survey. Given the data in this research we can estimate the optimal choices of two design attributes: a) number of sample elements per PSU, and b) allocation of the sample across the three "Black-density" strata.

To determine the optimal cluster size we use a total cost model, $C = C_o + C_a a + C_b ab$, where C_0 represents fixed costs, C_a is the sampling and screening cost for each sample cluster, of which a are selected, and C_b is the sampling, screening and interviewing cost

Table 4

Production Study
Synthetic Intracluster Correlations
for 100- and 200-Series Clusters for Seven Statistics by Stratum

	Synthetic Intracluster Correlation*					
Statistic	High Density Black Stratum		Medium Density Black Stratum		Low Density Black Stratum	
	100-Series	200-Series	100-Series	200-Series	100-Series	200-Series
Proportion Very Satisfied with Life as a Whole	.021	-.002	-.172	-.042	-.238	-.116
Proportion Who Think They Are Better Off Financially Than One Year Ago	.113	.075	.094	.069	.206	.049
Proportion Who Will Vote for Mondale	.189	.021	.086	-.087	-.436	-.046
Proportion Who Attend Church	.013	.017	-.009	-.078	.035	-.110
Proportion in Same City or Town All of Life	-.078	.001	.058	.114	.221	.248
Proportion Voted in 1980 Presidential Election	-.045	-.035	-.101	-.013	.364	.356
Proportion Who Think Reagan Will Be Elected President	-.045	-.045	-.545	-.078	.124	-.105
Average	.024	.005	-.084	-.016	.039	.039

* These estimates are unweighted.

associated with each interview obtained, of which there are b in each cluster. Because the proportions of Black households vary across the three strata in the design, the C_a and C_b parameters vary across strata (see Table 5). The optimal cluster size is computed as $\sqrt{C_a(1 - \rho)/(C_b\rho)}$ (Kish, 1965). Using cost data from the production survey, Table 5 presents estimated optimal cluster sizes for overall means and proportions with three alternative levels of intracluster correlation; .005, .01, and .02. (These values are similar to those obtained for attitudinal and behavioral variables in the actual surveys.) The C_a and C_b cost estimates for each stratum also appear. The Table shows that the optimal cluster sizes are largest in the low density stratum, reflecting the high screening costs in that group. Note also that these optimal cluster sizes tend to be larger than those actually used in the survey, $\bar{b} = 6.5$.

Note further that the optimal cluster sizes are similar for 100- and 200-series PSUs and the loss of cost-efficiency of the 200-series relative to that of the 100-series is minor and similar optimal cluster sizes result. (The sampling variance estimates also imply that intracluster correlations in the 100- and 200-series clusters are similar.)

The optimal cluster sizes in Table 5 generally exceed the levels that could be supported with a 100-series PSU definition. That is, a large proportion of 100-series PSUs would not have a sufficient number of Black household numbers to fulfill the designated second stage cluster size. For that reason alone, the 200-series is favored. Even with 200-series, the specified second stage cluster sizes could not be obtained for some PSUs in the low density stratum. (This suggests the true optimal cluster size solution should be constrained to reflect the capacities of the PSUs and the approach used here is useful to guide practical decisions on cost-efficiency, but does not reflect some extreme conditions.)

Table 5

Cost Parameters and Optimal Number of Sample Elements Per Cluster, by Stratum for 100- and 200-Series Clusters and Different ρ Values

Stratum and Cluster Definition	Optimal Cluster Size			Cost Parameters	
	ρ = .005	ρ = .01	ρ = .02	C_{ha}	C_{hb}
High Density Stratum					
100	15.9	11.2	7.9	\$50.81	\$40.11
200	15.9	11.2	7.9		\$39.78
Medium Density Stratum					
100	22.4	15.8	11.1	\$114.09	\$45.18
200	21.3	15.0	10.6		\$50.00
Low Density Stratum					
100	29.8	21.0	14.8	\$309.98	\$69.52
200	29.9	21.1	14.8		\$69.18

The second design decision evaluated is the choice of sample allocation to strata. The survey used sampling fractions in the ratio of 3:2:1 from the high density to the low density stratum. We explored the optimal allocation across strata, assuming that the optimal cluster sizes were chosen in each stratum (as shown in Table 5). Given a fixed cluster size in each stratum, b_h, we set the sampling fraction in the h-th stratum, f_h, proportional to $\sqrt{(Deff_h S^2{}_h)/(C_{ha}/b_h)}$, where $Deff_h$ is the design effect for the statistic in the h-th stratum, $S_h{}^2$ is the element variance in the h-th stratum, C_{ha} is the sampling and screening costs for PSUs in the h-th stratum, and b_h is the number of sample elements per cluster in the h-th stratum.

Table 6 presents optimal ratios of sampling fractions for various combinations of element variances in the three strata and the various ρ values. The Table shows that the optimal allocations across strata are relatively insensitive to changes in ρ values (for the range of ρ values that are likely given this design). If the strata with higher densities of Black households have element variances at least equal to that of the low density stratum, an oversampling of those strata is desirable. (This reflects the much lower costs in those strata.) The 3:2:1 ratio of sampling fractions is best when the ratio of strata standard deviations is about 1.7:1.5:1. An examination of the data obtained from the survey suggests that many variables have ratios of standard deviations across the three strata close to 1:1:1. For such variables the optimal ratio of sampling fractions is 1.7:1.4:1, given the optimal cluster sizes shown in Table 5. (With the cluster size of 6.5 actually used in each stratum, the optimal fractions have the ratio 2.5:1.6:1.) Both these ratios of sampling fractions suggest that the oversampling actually used in the production study created a loss of precision per unit cost, relative to that corresponding to the optimal sampling fractions.

Table 6

Optimal Allocation of the Sample Across Strata for Overall Means, Given Optimal Cluster Sizes in Each Stratum, for Various Relative Standard Deviations Across Strata and Values of Intracluster Correlations

Ratios of Within Stratum Standard Deviations (High:Med:Low)	Ratios of Optimal Sampling Fractions (High:Med:Low)
$\rho = .005$	
3 : 2 : 1	5.2 : 2.7 : 1
1.7 : 1.5 : 1	3 : 2 : 1
1 : 1 : 1	1.7 : 1.4 : 1
.33 : .5 : 1	.6 : .9 : 1
$\rho = .01$	
3 : 2 : 1	5.2 : 2.7 : 1
1.7 : 1.5 : 1	3 : 2 : 1
1 : 1 : 1	1.7 : 1.4 : 1
.33 : .5 : 1	.6 : .9 : 1
$\rho = .02$	
3 : 2 : 1	5.1 : 2.7 : 1
1.8 : 1.5 : 1	3 : 2 : 1
1 : 1 : 1	1.7 : 1.3 : 1
.33 : .5 : 1	.6 : .9 : 1

6. SUMMARY

Rare population sampling forces the survey statistician to consider combinations of PSU and cluster definitions, stratification, and alterations of measures of size which are not typically found in cross-section samples. This research found that these traditional sample design techniques can be adapted to increase the efficiency of two-stage telephone samples for the Black household population with telephones.

First, this research found that even the rough correspondence between telephone exchanges and large cities and states permitted stratification that successfully discriminated exchange groups with vastly different eligibility rates. The high density stratum had over twice the proportion of Black households as did the low density stratum. This permits control over screening costs in sample implementation. With other rare populations which are residentially segregated, similar results are expected.

Second, the use of rejection rules based on subpopulation eligibility effectively reduced screening costs within PSUs. This increases the eligible proportion of secondary numbers from twofold to ninefold, depending on which density stratum was considered.

Third, use of a larger PSU (200- versus 100-series of consecutive numbers) produced no serious loss of eligibility. Hundred series densely filled with eligible numbers tend to be adjacent to others densely filled. This is a discovery concerning the practice of assigning numbers by telephone companies. This fact permits larger numbers of sample numbers per PSU, another key feature in reducing the costs of the Black population sample.

Despite great pressures for cost reduction in rare population samples, it is important to balance errors and costs explicitly in choosing the final design. In this research such cost and sampling error modeling suggested that disproportionate allocation of the sample to Black-density strata is desirable. In addition, it is most efficient to select a relatively large set of secondary numbers per PSU. This set is sufficiently large that the 200- or 400-series PSU definition must be used.

Although we have applied this design only to the Black population, its performance should be similar for other residentially segregated populations. This includes income groups, certain occupational groups, and ethnic groups.

In addition, the discoveries of this research may also have implications for cross-section samples. Increasing the PSU size from 100 to 200 consecutive numbers may be advantageous in a two-stage RDD design for sampling the general telephone household population. The larger 200-series would provide twice as many numbers to select from and, as with the rare population, the proportion of eligible numbers would tend to be similar to that found in the 100-series. Therefore, given low intracluster correlation values, the cluster size of eligible numbers for a design could be set much closer to the optimal size. Because all PSUs selected would be able to support the chosen number of sample numbers, the achieved cluster size of eligible numbers should also be less variable over PSUs and therefore the impact of compensating weighting on the variance of estimates should not be great.

REFERENCES

BLAIR, J., and R. CZAJA (1982). Locating a special population using random digit dialing. *Public Opinion Quarterly*, 46, 585-590.

GROVES, R.M. (1978). An empirical comparison of two telephone sample designs. *Journal of Marketing Research*, 15, 622-631.

GROVES, R.M., and KAHN, R.L. (1979). *Surveys By Telephone*. New York: Academic Press.

KISH, L. (1976). "Optima and proxima in linear sample designs", *Journal of the Royal Statistical Society*, Ser. A, 139, 80-95.

KISH, L. (1965). *Survey Sampling*. New York: John Wiley.

THORNBERRY, O.T., and MASSEY, J.T. (1983). Coverage and response in random digit dialed national surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 654-659.

WAKSBERG, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 40-46.

Survey Methodology, June 1987
Vol. 13, No. 1, pp. 15-28
Statistics Canada

Comparing Telephone and Face-to-Face Interviewing in the United Kingdom

W.M. SYKES AND M. COLLINS[1]

ABSTRACT

This paper presents results from methodological experiments comparing telephone and face-to-face interviewing in surveys of the general population. The relatively low level of telephone ownership in the United Kingdom, especially among the less privileged, argues the need for a dual-mode approach combining telephone interviews with face-to-face interviews for those without telephones. This approach depends on the absence of differential mode-effects on the answers obtained or on the ability to account for these effects when they occur.

KEY WORDS: Telephone interviewing; Dual-mode interviewing; Social surveys; Response rates; Data quality.

1. INTRODUCTION

The choice of a mode of data collection for a survey depends upon the availability of facts about the alternatives. In the U.K., such facts about telephone interviewing have just recently begun to emerge. The necessary comparisons between telephone interviewing and other data collection modes have been carried out only in the last two years. This delay is surprising given the lively debate about the merits and drawbacks of telephone interviewing and the attention which the issue has received in other countries.

Two studies conducted by the Survey Methods Centre at Social and Community Planning Research comparing telephone and face-to-face interviewing provide the focus for this paper. Carried out in 1983 and 1984, these studies examine some of the central issues: the public's willingness to take part in telephone surveys and the kind, quality and volume of data that can be collected. The studies are described in Section 2 and their results presented in Sections 3 and 4. Reference is also made to another British study - an experiment carried out in 1985 by the Market Research Development Fund - and to the larger volume of methodological research conducted in other countries, particularly the United States.

2. THE SCPR STUDIES

Our research program reflected telephone ownership which is low by North American standards: about 75% of households possessed telephones in 1983. Non-coverage is substantial and crucial, for social researchers, because of its bias towards less affluent sectors of British society. In this context, the main objective was to evaluate dual-mode interviewing, where telephone owners would be interviewed by telephone, and non-owners face-to-face.

The first study provided two comparisons towards this evaluation: between an experimental dual-mode sample and a larger national sample interviewed face-to-face; and between two samples of telephone owners, one sample interviewed by telephone, the other interviewed face-to-face. In this paper, we focus on the latter comparison, which addresses the question

[1] W.M. Sykes and M. Collins, Survey Methods Centre, Social and Community Planning Research, 35 Northampton Square, London EC1V 0AX, England

that lies at the heart of any evaluation of the dual-mode approach: are telephone and face-to-face data compatible or are there modal differences between them? If there are modal differences, the data cannot be "added" together and treated as a single data set without the kind of adjustments not usually possible in a one-time survey. The second study concentrated only on this direct comparison between the two interview methods among telephone owners.

2.1 Study 1

The first study was conducted alongside the 1983 British Social Attitudes Survey, which is here referred to as the "main" survey. This survey involved face-to-face interviews of about an hour, covering a wide range of political, economic, social and moral issues.

The sample for the main survey was about 1,750, and was representative of adults aged 18 or over living in private households. For practical reasons, the sample was confined to those at addresses in the Electoral Register. People living in institutions (though not private households at such institutions) were excluded, as were the 4% of adults known to live at addresses not on the Electoral Register (Todd and Butcher 1982).

A multi-stage design was used with four stages of selection: 103 constituencies in England and Wales and 11 local authority districts in Scotland were selected with probability proportional to electorate; within each a single polling district was selected, again with probability proportional to electorate; from each polling district, 23 addresses were selected with probability proportional to the number of electors registered at the address. At the final stage, one person at each address was selected by the interviewer, using an adaptation of the Marchant-Blyth procedure (Blyth and Marchant 1973).

For the experiment, a parallel sample of about 800 addresses (seven per area) was selected from the same 114 sampling points. These addresses, together with all the names in the Electoral Register, were submitted to British Telecom's telephone number-retrieval facility. The facility yielded telephone numbers for 65% of the submitted addresses. Most of the difference between this retrieval rate and the level of telephone ownership - around 75% at the time - can be explained by ex-directory numbers: about 12% of telephone numbers in Great Britain are ex-directory, with regional and other variations as noted by Collins and Sykes (1987). Other problems in tracing telephone numbers seem to have had little effect.

The following procedure was used by British Telecom for retrieving telephone numbers: once the correct telephone exchange area had been identified by the address, the subscriber's name was looked up in the directory. Specific address details (i.e., the street name) helped distinguish between subscribers with identical names. Since it is not clear from the Electoral Register which of the names at an address is that of the subscriber, British Telecom was asked to check every name before abandoning a search.

The telephone numbers obtained were systematically assigned to four sub-samples. Two of these were interviewed by telephone using a questionnaire expected to take about 20 minutes to complete. The questions were drawn from all sections of the main Social Attitudes questionnaire. The other two sub-samples were interviewed by telephone using a longer questionnaire - estimated at 40 minutes - that was also drawn from the main survey questionnaire. Sub-samples allocated to both the 20-minute and the 40-minute questionnaires were sent a letter before the telephone calls. The other sub-samples received no advance warning of the survey. In all cases the selection of a respondent for interview was on the same basis as for the main survey.

Experimental sample addresses for which no telephone numbers could be obtained from British Telecom were given face-to-face 20-minute interviews. Combined with those obtained by telephone, these interviews formed a dual-mode survey that was compared with the main face-to-face interview survey (Sykes and Hoinville 1985).

A more direct examination of interview mode effects was sought by submitting a systematic sub-sample of 600 of the main sample addresses (five in each area) to British Telecom's number-retrieval service. In this case, numbers were returned for 55% of the addresses (the variability of the success rate of the British Telecom number-retrieval service remains unexplained). Comparisons were then made between those who were interviewed by telephone and those who could have been interviewed by telephone but were interviewed face-to-face. By restricting comparisons to the telephone-accessible population, we controlled for effects attributable to differences between the compared populations rather than to differences in the mode of data collection.

2.2 Study 2

The second experiment concentrated on this direct comparison. About 2,300 addresses were selected from the Electoral Register, as in Study 1, and were sent to British Telecom for telephone numbers (with in this case, a 61% retrieval rate). Addresses for which telephone numbers were retrieved were split into three sub-samples. One group was interviewed by telephone using "pencil and paper" methods; another was interviewed using Computer Assisted Telephone Interviewing (CATI); the third was interviewed face-to-face. Our experiment with CATI was a practical failure (for a number of reasons), but the other two sub-samples again give us a direct comparison between people interviewed by telephone and people who could have been interviewed by telephone but were interviewed face-to-face. The questionnaire, designed to take 25 minutes, consisted of a sub-set of questions from the 1983 British Social Attitudes Survey.

2.3 Limitations on the Comparisons between Interviewing Modes

Three factors could limit comparisons between the answers obtained face-to-face and those obtained over the telephone. First, differential non-response (as discussed in Section 3) could have led to differences in the composition of the respondent sets. This possibility was tested using a number of demographic and socio-economic variables believed to be associated with certain attitude variables. Significant differences between the respondent sets suggest that, quite apart from any differences between the modes in overall response levels, certain kinds of people are more likely to participate in a telephone rather than a face-to-face survey, and vice versa. The variables examined were: age within sex, marital status, household composition, economic status, socio-economic group and geographical location. No statistically significant evidence of differential non-response was found in the first study. In the second study, two variables showed statistically significant differences between the telephone and face-to-face samples: household composition (the telephone respondents included a higher proportion of childless couples under 60, while the face-to-face sample had a higher percentage of couples with young children and teenagers); and socio-economic group (intermediate and junior non-manual workers and those in "other" occupations had greater representation in the telephone sample than face-to-face, and "homemakers" were a higher proportion of the face-to-face sample). These differences may well represent only sampling fluctuations, but they should lead to some caution in the interpretation of differences between the answers of the two samples.

The second possibility is of different levels of skill or supervision between the telephone and face-to-face interviewers. Six telephone interviewers were employed on the first experimental survey. Two were fully trained and experienced face-to-face interviewers, but the remainder had had no previous interviewing experience and so received basic interviewer training as well as the special telephone interviewing training that all six interviewers underwent. The second study involved 10 interviewers, three of whom had worked on the previous study. As in the previous study, a supervisor was present to listen in, advise on interviewing technique when necessary and check for obvious errors in completed questionnaires.

The face-to-face interviewers for both studies were drawn from Social and Community Planning Research's panel of about 300 regularly employed face-to-face interviewers. Their training in basic interviewing techniques was similar to that given to the telephone interviewers. However, for the most part, the face-to-face interviewers were more experienced than their telephone counterparts. Differences between the two groups of interviewers should, therefore, be kept in mind, especially differences suggesting lower quality in the telephone interviews.

The third factor is the questionnaires. The main Social Attitudes questionnaire, comprising about 100 questions, was divided into five broad topic areas: employment, education, health and housing, issues of social class, and racial and sexual equality. The experimental questionnaires were composed of those questions considered most important in the main survey. These questions were chosen to represent the full range of question types in the main questionnaire.

As a result, the experimental questionnaires covered a range of topics (including some "sensitive" issues) and included questions involving different kinds of response tasks and levels of complexity. The order of the questions on the Social Attitudes Survey was maintained for both the 20-minute and 40-minute experimental questionnaires used in the first study and for the 25-minute questionnaire used in the second study. Thus the 40-minute questionnaire was not made up of the short questionnaire followed by a further 20 minutes of questions: rather, questions from the 20-minute version were spread throughout. Alterations to question wording were made only when unavoidable; for example, re-wording to adjust for the necessary absence of showcards. The Social Attitudes Survey questionnaire consists largely of closed questions, so few of the results from our experiments relate to open questions.

All of these limitations should be kept in mind when examining our results, but they are largely inevitable in such comparative studies. As described above, we have tried to identify and minimize them. They are of great concern only when our results suggest mode effects that might confound the effects of other variables: most of our results do not point to this. Thus the limitations should be considered only as potential sources of effects counteracting mode effects we might otherwise have found - surely a less serious threat to the validity of our conclusions.

3. RESPONSE RATES

In the U.K., doubts about the feasibility of telephone interviewing, particularly for social surveys, stem from concerns not only with the level of communication possible, and its effect on both cognitive and affective dimensions of the interview, but also with the general social acceptability of this use of the telephone. In Britain, it is a common belief among researchers that "cold calls" from strangers are likely to be treated with circumspection: a call from a telephone interviewer may be regarded as inappropriate and intrusive.

A common counter argument points out the possible advantages telephone interviewing has over face-to-face interviewing, particularly in inner city areas. Escalating personal and property crime has led to increasing suspicion of strangers, which means falling response rates and the installation of devices such as entry-phones that make it harder for personal interviewers to contact respondents. By telephone, contact will also certainly be made at an address if someone is there, and, if not, subsequent attempts are not expensive.

Table 1 shows the response rates for both studies conducted by the Survey Methods Centre.

Table 1
SCPR Experiments: Response Rates

	Study 1		Study 2	
Bases	Telephone (429)	Face-to-Face (313)	Telephone (730)	Face-to-Face (631)
	%	%	%	%
Completed interviews	53	60	46	68
Partial interviews	1	-	-	-
Refusal (no selection)	5	2	21	6
Refusal (proxy)	9	5	7	4
Refusal (selected person)	11	18	10	11
No contact[a]	3	1	8	4
Selected person never in	3	3	3	2
Ill, away, language problems	2	5	2	4
Other[b]	13	6	4	2

Study 1: $\chi^2 = 3.72$ d.o.f. = 1 $0.05 < p < 0.1$
Study 2: $\chi^2 = 66.22$ d.o.f. = 1 $p < 0.001$
Studies 1 and 2 combined: $\chi^2 = 59.46$ d.o.f. = 1 $p < 0.005$

comparisons with only two categories: completed interviews and non-completed interviews.

[a] Includes "Ring no answer" and "Permanently engaged".

[b] Includes "Broken appointments", "Too old", "Incapacitated", "No connection", "Right number, wrong address".

Response to these studies, for both the telephone and face-to-face components, was relatively low. (We would normally expect personal interview response rates of over 70% before reissue of refusals.) This owes something to the nature of the surveys - general purpose surveys are notoriously difficult to "sell" to respondents. The same argument can also be applied to the only other major British methodological comparison survey, carried out by Marplan on behalf of the Market Research Development Fund. This study used the same sampling method as our own experiments and also included a wide range of general questions, under the title *Lifestyle in the 1980's*. In this case, the response rates obtained were 45% by telephone with a sample base of 1697 and 67% face-to-face with a sample base of 1233 (Market Research Development Fund 1985). In both our studies, the response rate was lower for telephone interviews: barely half of the issued addresses yielded interviews. As Table 1 shows, the difference was on the borderline of non-significance for Study 1 but was statistically significant for Study 2 and for Studies 1 and 2 in combination.

The difference might be attributed to our relative lack of experience with telephone interviewing, but it is consistent with findings from other countries. For example, in the United States lower response rates - mostly arising from the higher incidence of refusals to cooperate - have been reported by a number of authors (e.g., Hochstim 1967; Henson, Roth and Cannell 1977). The position is summarized by Groves and Kahn, who write:

> "The response rate of national surveys remains at least five percentage points lower than that expected in personal interview. This has been a rather stable comparison despite changes over time in training of interviewers, monitoring techniques, feedback procedures from monitors, and techniques of introducing the survey to the respondent." (Groves and Kahn 1979; p. 219)

These findings suggest that sociological and psychological explanations of resistance to the telephone approach may be more appropriate than explanations of interviewer and general methodological inexperience. However, the first SCPR study appears to have been rather more successful than either the second or the MRDF study. It has been suggested that this difference was due to the interest and excitement surrounding the first experiment. This may have communicated itself to the interviewers (for example, researchers were continually "dropping in" to observe the proceedings), thus affecting their success rates. Certainly, experience with face-to-face surveys suggests that interviewer morale and energy are important for good response rates.

In the SCPR studies two survey conditions were varied to assess their impact on telephone response rates. For the first survey, half the telephone respondents were asked to do 20-minute interviews and the other half did 40-minute interviews (respondents were told the length of the interview towards the end of the introduction), and in both surveys advance letters giving notice of the interview were sent to a random half of the telephone sample.

Table 2 shows that response for the 40-minute interview was lower than for the 20-minute interview, although the difference between the overall distributions was not significant. The main single reason for this lower response was the higher direct refusal rate, possibly indicating that respondents were less willing to undertake the longer interviews. However, very few respondents who had agreed to participate terminated an interview prematurely - even with the longer interview.

Different strategies may be needed for longer questionnaires. While it may be reasonable to request respondents to take part in a 20-minute interview at the time when first contact is made, a system of appointments may be more successful where more interviewing time is required. Wiseman and McDonald (1979) suggest that refusal rates are likely to be lower when interviewers are instructed to make call-back appointments should the respondents indicate that they are busy.

In other studies, sending advance letters to potential telephone respondents has been found to improve response rates. For example, Dillman, Gallegos and Frey (1976) obtained refusal rates which were, on average, 6% lower for respondents receiving advance letters (compared with 14%). As Table 3 shows, in the SCPR experiments response rates were slightly higher among respondents who had been sent an advance letter (no record was kept of whether letters had been received) although the differences were not statistically significant.

To explore why respondents refuse to be interviewed by telephone, 55 refusers to the first study were followed-up to see whether they would have co-operated at the first contact if they had been approached personally. Forty said that the method of interview would have made no difference to their decision, and only a very small number of these people subsequently agreed to be interviewed. Most of the rest said they would have taken part if they had been approached face-to-face and eventually completed a face-to-face interview (13 out of 15).

Because face-to-face refusers were not followed up, we do not know if a proportion of this group would have preferred to be approached by telephone.

3.1 Response Differences and Data Quality

The public's perception of the proper use of the household telephone may effect not only response rates, but also the kinds of questions respondents will be prepared to answer. Of even greater concern, however, is the type of communication possible between interviewer and respondent and its potential effect on the measurements made.

Face-to-face communication takes place both verbally and non-verbally, while the telephone has only limited channel capacity with exchanges between interviewer and respondent restricted to what is said and so-called paralinguistic cues: tone of voice, pauses and so on (Miller and Cannell 1982).

Table 2
SCPR Experiments: Effects of Interview Length (Study 1)

Bases	40-Minute (206)	20-Minute (223)
	%	%
Completed interviews	48	59
Refusal	27	23
Other	25	18

$\chi^2 = 4.7$ d.o.f. $= 2$ $0.10 > p > 0.05$

Table 3
SCPR Experiments: Effects of Advance Letters on Response Rates

	Study 1		Study 2	
Bases	Letter (215)	No Letter (214)	Letter (388)	No Letter (392)
	%	%	%	%
Completed interviews	55	51	48	43
Refusal	23	27	37	38
Other	22	21	15	19

Study 1: $\chi^2 = 1.09$ d.o.f. $= 2$ $p > 0.5$
Study 2: $\chi^2 = 2.8$ d.o.f. $= 2$ $p > 0.2$
Studies 1 and 2 combined: $\chi^2 = 3.49$ d.o.f. $= 2$ $p > 0.1$

The possible implications for survey measurements of the telephone's limited channel capacity are numerous. For example, the absence of visual aids may increase the difficulty of some response tasks. "Voice only" communication may not convey the full meaning behind respondents' words (making it difficult, for example, to probe open-ended questions) and may not reveal if they actually understand the questions. There may also be limitations on the interviewer's ability to perform his or her role. Can verbal signals, for example, replace the non-verbal cues that convey interest and attention to the respondent, or those that help control the interview? Can the interviewer hold the concentration of the respondent, particularly in long interviews? Conversely, is the absence of visual stimuli a desirable reduction in the many sources of variability in survey data? Finally, does the greater social distance in the telephone interview make the respondent more or less comfortable in revealing sensitive information such as income, or information with a strong social desirability component?

SCPR's experiments addressed some of these issues.

3.1.1 General Comparisons

Given the different refusal rates of the interviewing modes, it is surprising that there are few other general differences. This result has been replicated in many studies in the U.S. (Groves and Kahn 1979; Lucas and Adams 1977; Jordan *et al.* 1980; Colombotos 1969; Wiseman 1972), and in other countries such as Denmark (Kormendi *et al.* 1986). Simple straight-forward questions asked identically by telephone and face-to-face yield similar distributions of response.

In the SCPR studies the marginal distributions of response yielded by the different modes of interview were compared and differences were tested for statistical significance using chi-squared tests. These tests were performed on unweighted data. However, tables in the text, unless otherwise indicated, show distributions of data weighted to take account of any differences between the number of people listed on the Electoral Register and those found at an address. Such differences occurred in approximately 25% of cases, in each of which the data were weighted by the number of persons aged 18 or over living at the address divided by the number of electors listed on the Register for that address. Weighted tables are given to allow readers to decide if they might draw different conclusions from telephone survey data and face-to-face survey data when both sets have been prepared according to routine procedures.

Standard chi-squared tests were performed even though the data arose from a multi-stage sample. It has been shown (see, for example, Holt, Scott and Ewings 1980) that underestimating true variability by ignoring sample design will generally lead to test statistics which are too large, and hence to the false rejection of null hypotheses (i.e., to anti-conservative tests). For the Social Attitudes Survey, however, estimation of true standard errors for attitudinal variables yields Design Factors (the ratio of the complex standard error to the simple random sampling standard error) which are rarely above 1.2 (Jowell and Witherspoon 1985). Further, the literature argues that in 2-way tests of independence the consequences of clustering are likely to be less severe (Holt, Scott and Ewings 1980). As a result, we feel justified in using standard chi-squared tests to avoid the large amount of computation necessary for corrected statistics. If anything, this approach will overstate the significance of differences between interview modes.

In the first study we looked at 95 questions and parts of questions and in the second study 69. The results are shown in Table 4. It is clear that in both studies the results accorded with those of other researchers: the interviewing modes yield significantly different distributions of answers for only a very small percentage of questions. A similar finding emerged from the MRDF study.

3.1.2 Comparisons for Particular Question Forms

Despite the general result, research in the U.S. has shown that there are specific kinds of questions for which differences in response distributions do occur. For example, Groves and Kahn (1979) demonstrated a tendency for respondents to give truncated answers to open-ended items over the telephone. This might be due to the faster pace of telephone interviewing, as noted, for example, by Dillman (1970) and Williams (1977). Both interviewers and respondents tend to speak more quickly on the telephone and to avoid silent pauses. The swifter pace of telephone interviews was shown in our second experiment. As Table 5 shows, with an interview designed to take 25 minutes, 10% of the telephone interviews were conducted in under 20 minutes, compared with 5% of face-to-face interviews. At the other extreme, 41% of face-to-face interviews took more than half an hour compared with under a third of the telephone interviews.

Ball (1980) suggests that the greater speed may occur because the norms of telephone conversations require both the interviewer and respondent to work to maintain the conversational flow. This may leave respondents with less time to think about their answers. Certainly, silences seem to make people uncomfortable – in a study by Jordan (1980) routine pauses in the interview were described as interminable by interviewers. Undoubtedly there are many other contributing factors: even the absence of visual distractions may be important.

Although SCPR's experimental studies did not carry any open-ended items, the MRDF study included a number of spontaneous awareness measures. Comparisons of telephone and face-to-face results appear consistent with the findings discussed above. One example

Table 4
Differences in Marginal Distributions of Response:
Telephone vs. Face-to-Face

Bases	Study 1 (95)	Study 2 (69)
	%	%
No significant difference	91	87
Significant at 5%	7	9
Significant at 1%	2	4

Table 5
Interview Length by Mode of Interview (Study 2)

Unweighted Bases	Telephone (354)	Face-to-Face (360)
	%	%
Minutes		
Under 20	10	5
20–29	63	53
30–40	22	33
40+	6	8

$\chi^2 = 17.6$ d.o.f. $= 3$ $p < 0.01$

Table 6
Comparisons of Responses on an Open Question (MRDF Survey)

What do you like about .... soup?

Bases	Telephone (700)	Face-to-Face (601)
	%	%
Number of answers		
None	33	22
One	58	61
Two	7	14
Three or more	1	2
Average	0.77	0.96

$\chi^2 = 32.2$ d.o.f. $= 3$ $p < 0.01$

is given in Table 6, which shows that a third of telephone respondents gave no answers, compared with under a quarter face-to-face. Also, the average number of responses given over the telephone was significantly lower.

We might assume that more or longer answers mean more valid reporting, and this would imply a need for techniques to improve open questions on telephone surveys. At the extreme, it might be concluded that open questions have only limited use on telephone surveys, for example when only the first information spontaneously offered by respondents is wanted. This assumption needs, however, to be tested: here we can only report the effect.

Differences between response distributions have also been reported for attitude scale questions asked identically face-to-face and over the telephone. Telephone respondents tend towards "acquiescence" and "extremeness" response bias (Jordan, Marcus and Reeder 1980; Groves and Kahn 1979). With the agree/disagree scales used by MRDF, the telephone sample showed a slight tendency to agree more. However, no difference in the spread of responses was found - there was no evidence of a greater tendency towards extremeness.

3.1.3 Sensitive Questions

Concerning the types of question that can be used in telephone surveys, researchers have paid much attention to sensitive questions - those that deal with private or personal information and those for which certain responses are more clearly socially acceptable. Initial views about the likely effects of asking sensitive questions over the telephone were divided. Those who felt that respondents would be less willing to answer truthfully said that the lack of the interviewers' reassuring presence would make respondents less likely to be frank and open. The opposite view - that respondents would give more valid answers - maintained that greater social distance, by preserving anonymity, would encourage truthful responses.

Most evidence supports the latter view (Colombotos 1965; Wiseman 1972; Henson, Roth and Cannell 1974; Locander 1974; Rogers 1976). The major exception is reported by Groves and Kahn (1979), who found telephone respondents to be reticent about their financial status and other sensitive issues.

Our studies support the hypothesis that telephone surveys work well for sensitive questions. For instance, in our first study 14 questions were isolated as potentially sensitive and tested for mode-effects. Three illustrative examples of such questions are given below:

i) How would you describe yourself?:
(Read out) ...
... as very prejudiced against people of other races
... a little prejudiced
... or, not at all prejudiced?

ii) Do you think, on the whole, that Britain gives too little or too much help to Asians and West Indians who have settled in this country, or are present arrangements about right?

iii) Finally in this section, I would like you to tell me whether, in your opinion, it is acceptable for a homosexual person to be a teacher in a school?

No significant differences in the marginal distributions of response were found. For several questions, however, there was a somewhat greater tendency to give socially desirable answers in face-to-face contact. In other words, the questions seemed to be less sensitive over the telephone. For example, 28% of respondents interviewed by telephone admitted to having been questioned by police over the past two years in connection with a crime, compared with 20% of face-to-face respondents.

Sensitive questions in the MRDF study also showed a slight tendency for telephone respondents to give more "honest" answers, although on individual questions differences in the distributions were generally not significant. For example, when asked to describe themselves on a number of dimensions, telephone respondents were more likely to say they were "attractive" (mean score of 2.81 out of 4 compared with 2.72 face-to-face) and were more ready to give an answer at all (88% gave an answer compared with 75% face-to-face).

Questions about income have generally been regarded as potentially problematic in telephone surveys, both in respondents' willingness to answer and in the answers given. Under-reporting of income levels is the main expectation, although in practice this may be hard to distinguish from under-estimation resulting from higher non-response in the upper income brackets. A study by Locander and Burton (1976) suggests that the validity of income data may depend on the question format. In a comparison of four question formats, under-reporting of income resulted from a method that first asked "Is your income more than $2,000?" gradually increasing the figure until the first "no" response. However, over-reporting of income was encouraged by a similar method that began with the highest income category. The method used for the telephone surveys in the SCPR experiments was similar to the first type described above. It most closely approximates the response task set by the face-to-face income question in which a card indicating broad income bands, starting with the lowest, was used to guide the respondents' choice. Over the telephone, the ranges were read to respondents starting at the lowest levels. The results are shown in Table 7.

In neither study was there any mode difference in respondents' willingness to answer the income question. Differences in the distribution of answers, in this case a possible under-reporting of income, were only apparent in the first study.

3.1.4 Complex Questions

In both SCPR studies a number of questions were identified in advance as likely to pose particular response problems for telephone respondents. These included questions with one or more potentially difficult concepts, long questions and questions with large numbers of response options. Such "complex" questions appear to be no more problematic for telephone respondents than for those interviewed in person. For example, of 19 "complex" questions

Table 7

Gross Household Income: SCPR Studies

	Study 1		Study 2	
Bases[a]	Telephone (183)	Face-to-Face (170)	Telephone (297)	Face-to-Face (352)
Income				
less than £5,000	38	27	28	28
£5,000–£9,999	42	37	37	38
£10,000 or over	21	35	35	35

Study 1: χ^2 = 10.08 d.o.f. = 2 $p < 0.01$
Study 2: χ^2 = 0.11 d.o.f. = 2 $p > 0.9$

Bases	(217)	(199)	(344)	(405)
Don't know/ Not answered	16%	15%	14%	13%

[a] "Don't know" and "Not answered" excluded.

identified on the first study (12 of which had been asked with the aid of show-cards face-to-face), only one showed any evidence of mode-effects.

4. SUMMARY AND CONCLUSIONS

Since telephone ownership in the United Kingdom remains relatively low, particularly for certain sectors of the population, telephone interviewing is unlikely to replace face-to-face interviewing for surveys that must include the less advantaged. But its potential in combination with traditional face-to-face procedures has gained recognition. For example, the U.K. Labour Force Survey uses telephone interviewing for second and subsequent interviews with eligible respondents who have indicated a willingness to be contacted by telephone.

Crucial to the success of dual-mode surveys is the absence of differential mode effects. The results reported here provide a largely optimistic outlook. With a few exceptions there were no statistically significant differences between the distributions of answers obtained face-to-face and those given over the telephone.

However, the relatively low response rates to telephone surveys poses problems that need to be overcome. High refusal rates can reduce the cost-effectiveness of using the telephone. More importantly, they increase the chances of introducing bias into the sample. Further research to explore ways of improving telephone response rates is necessary to realize the potential of the method in the United Kingdom.

ACKNOWLEDGEMENTS

This research was carried out within the SCPR Survey Methods Centre, funded as a Designated Research Center by the Economic and Social Research Council (grant HR 3333). The authors are grateful to the editors and referees for their extensive and useful comments on earlier drafts of this paper.

REFERENCES

ARONSON, S. (1971). The Sociology of the Telephone. *International Journal of Comparative Sociology*, 12, 153-167.

BALL, D.W. (1968). Towards a Sociology of Telephones and Telephoners. In *Sociology and Everday Life* (ed. Marcello Truzzi), Englewood Cliffs, New Jersey: Prentice Hall.

BERGSTEN, J.W. (1979). Some Methodological Results from Four Statewide Telephone Surveys Using Random Digit Dialing. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 239-243.

BISHOP, Y.M.N., FIENBERG, S.E., and HOLLAND, P.W. (1975). *Discrete Multi-variate Analysis: Theory and Practice*. Cambridge, Mass: MIT Press.

BLANKENSHIP, A.B. (1977). *Professional Telephone Surveys*. London: McGraw-Hill.

BLYTH, W.G., and MARCHANT, L.J. (1973). A self-weighting random sampling technique. *Journal of the Market Research Society*, 15, 157-162.

CANNELL, C.F., OKSENBERG, L., and CONVERSE, J.M. (1979). Experiments in interviewing techniques. Research Report, Institute for Social Research, University of Michigan.

CHRISTOFFERSEN, M.N. (1984). The quality of data collected at telephone interviews. Danish National Institute of Social Research, Copenhagen.

COLLINS, M. (1983). Telephone interviewing in consumer surveys. *Market Research Society Newsletter*, October.

COLLINS, M., and SYKES, W. (1987). The Problems of Non-Coverage and Unlisted Numbers in Telephone Surveys in Britain. *Journal of the Royal Statistical Society*. Ser. A, 150, (forthcoming).

COLOMBOTOS, J. (1965). The effects of personal vs. telephone interviews on socially acceptable responses. *Public Opinion Quarterly*, 29, 457-458.

COLOMBOTOS, J. (1969). Personal versus telephone interviews: effect on responses. *Public Health Reports*, 84, 773-782.

COOMBS, L., and FREEMAN, R. (1986). Use of telephone interviews in a longitudinal fertility study. *Public Opinion Quarterly*, 28, 112-117.

CZAJA, R., BLAIR, J., and SEBESTIK, J.P. (1982). Respondent selection in a telephone survey: a comparison of three techniques. *Journal of Marketing Research*, 19, 381-385.

DE MAIO, T.J. (1984). Refusals in telephone surveys: when do they occur? Paper presented at the 39th Annual Conference of the American Association for Public Opinion Research.

DILLMAN, D. (1970). *Mail and Telephone Surveys: The Total Design Method*. New York: John Wiley.

DILLMAN, D., GALLEGOS, J., and FREY, J. (1976). Reducing refusal rates for telephone interviews. *Public Opinion Quarterly*, 40, 66-78.

FALTHZIK, A. (1972). When to make telephone interviews. *Journal of Marketing Research*, 9, 451-452.

FITTI, J.E. (1979). Some results from the telephone health interview survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 244-249.

FLEISHMAN, E., and BERK, M. (1979). Survey of interviewer attitudes towards methodological issues in the national medical care expenditure survey. Paper presented at the Third Biennial Conference on Health Survey Research and Methods, Reston, Virginia.

GROVES, R., and KAHN, R. (1979). *Surveys by Telephone: A National Comparison with Personal Interviews*. New York: Academic Press.

GROVES, R.M., MAGILAVY, L.J., and MATHIOWETZ, N.A. (1981). The process of interviewer variability. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 438-443.

GROVES, R.M., and MATHIOWETZ, N.A. (1984). Computer assisted telephone interviewing: effects on interviewers and respondents. Survey Research Center, University of Michigan.

HENSON, R., ROTH, A., and CANNELL, C.F. (1974). Personal vs. telephone interviews and the effects of telephone re-interviews on reporting of psychiatric symptomatology. Research Report, Survey Research Center, University of Michigan.

HOCHSTIM, J.R. (1967). A critical comparison of three strategies of collecting data from households. *Journal of the American Statistical Association*, 62, 976-986.

HOLT, D., SCOTT, A.J., and EWINGS, P.D. (1980). Chi-squared tests with survey data. *Journal of the Royal Statistical Society*, Ser. A, 143, 303-320.

IBSEN, C.A., and BALLWEG, J. (1974). Telephone interviews in social research: some methodological considerations. *Quality and Quantity*, 7, 181-192

JOWELL, R., and WITHERSPOON, S. (1985). *British Social Attitudes: The 1985 Report*. Aldershot: Gower.

JORDAN, L.A., MARCUS, A.C., and REEDER, L.G. (1980). Response styles in telephone and household interviewing: a field experiment. *Public Opinion Quarterly*, 44, 210-222.

KAHN, R.L., and GROVES, R.M. (1977). *Comparing telephone and personal interview systems*. Survey Research Center, University of Michigan.

KORMENDI, E., EGSMOSE, L., and NOORDHOEK, J. (1986). Datakvalitet ved Telefon-interview. Socialforskingsinstituttet, Studie 52, Copenhagen.

LOCANDER, W.B., and BURTON, J.P. (1976). The effect of question form on gathering income data by telephone. *Journal of Marketing Research*, 13, 189-192.

LOCANDER, W.B., SUDMAN, S., and BRADBURN, N. (1974). An investigation of interview method, threat and response distortion. *Proceedings of the Social Statistics Section, American Statistical Association*, 21-27.

LUCAS, W.A., and ADAMS, W.C. (1977). *An Assessment of Telephone Survey Methods*. Santa Monica, California: Rand Corporation.

McCULLAGH, P., and NELDER, J.A. (1983). *Generalised Linear Models*. London: Chapman and Hall.

MILLER, P.V., and CANNELL, C.F. (1982). A study of experimental techniques for telephone interviewing. *Public Opinion Quarterly*, 46, 250-269.

Market Research Development Fund, (1985). *Comparing telephone and face-to-face surveys*. Marplan Ltd.

OKSENBERG, L., COLEMAN, L., and CANNELL, C. (1984). Voices and refusal rates in telephone surveys. Unpublished manuscript.

O'NEIL, M., GROVES, R., and CANNELL, C. (1979). Telephone interview introductions and refusal rates: experiments in increasing respondent cooperation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 252-255.

ROGERS, T.F. (1976). Interviews by telephone and in person. *Public Opinion Quarterly*, 40, 51-65.

SCHMIEDESKAMP, J.W. (1962). Reinterviews by telephone. *Journal of Marketing*, 26, 28-34.

SYKES, W., and HOINVILLE, G. (1985). Telephone interviewing on a survey of social attitudes: a comparison with face-to-face procedures. Social and Community Planning Research Center.

TODD, J., and BUTCHER, R. (1982). *Electoral Registration in 1981*. London: OPCS.

WILLIAMS, E. (1977). Experimental comparisons of face-to-face and mediated communication. *Psychological Bulletin*, 84, 963-976.

WISEMAN, F. (1972). Methodological bias in public opinion surveys. *Public Opinion Quarterly*, 34, 105-108.

WISEMAN, F., and McDONALD, P. (1979). Noncontact and refusal rates in consumer telephone surveys. *Journal of Marketing Research*, 16, 478-484.

Survey Methodology, June 1987
Vol. 13, No. 1, pp. 29-43
Statistics Canada

Issues in the Use of Administrative Records for Statistical Purposes

G.J. BRACKSTONE[1]

ABSTRACT

Demands for statistics on all aspects of our lives, our society and our economy continue to grow. At the same time statistical agencies share with many respondents a growing concern over the mounting burden of response to surveys. One result of the search for alternative methods of satisfying statistical demands has been an increased emphasis on the use of administrative records for statistical purposes. This paper reviews recent experience at Statistics Canada in this area and discusses obstacles to the greater use of administrative records. Approaches to rendering administrative systems more useful for statistical purposes are reviewed, together with some important concerns related to information protection and record linkage.

KEY WORDS: Indirect estimation; Survey frames; Survey evaluation; Access; Confidentiality.

1. INTRODUCTION

Demands for statistics on many aspects of our lives, our society, our economy and our environment continue to grow. This may be due in part to our increased ability to handle and manipulate large sets of data as we move into the so-called information age, and it may also be a reflection of the increasing complexity of our social and economic systems and our desire to understand them better. Whatever their cause we face these demands in a climate of tight budgetary constraint for government statistical agencies. At the same time, statistical agencies are sensitive to the increased burden that would be imposed on respondents by an increase in survey-taking activity to meet these demands.

These factors have led to the exploration of other means of satisfying these statistical demands. Prominent among these alternative means is the increased use of existing administrative systems as sources of statistical data. This is not a new idea. For many years, statistical data have been a by-product of administrative processes in domains such as vital statistics, imports and exports, health care, and education. We will describe later how this usage of administrative data has spread more recently to statistics on businesses and on families and individuals.

The first sections of the paper describe the variety of types and uses of administrative records, illustrating some of their uses in Statistics Canada's program. The heavy dependency of Canada's statistical system on administrative records will be apparent. Section 6 discusses issues of accessing administrative sources and making them more appropriate for statistical use. Finally, a brief review of privacy concerns related to administrative record use is provided.

2. TYPES OF ADMINISTRATIVE RECORD

Administrative records come in many shapes and sizes. An important distinction is between those administered nationally (usually by the Federal Government) and those

[1] G.J. Brackstone, Assistant Chief Statistician, Statistics Canada, 26-J R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario K1A 0T6.

administered sub-nationally (e.g., by provinces or municipalities). For the latter to be useful nationally, agreement between jurisdictions is required on items such as definitions, standards, record formats, and procedures. Such agreement is not always easy to achieve, particularly in domains that are constitutionally within provincial jurisdictions.

Administrative records vary in terms of their purpose, and their purpose is a prime determinant of their coverage and quality, and therefore of their statistical usefulness. Six broad categories of purpose can be distinguished.

(1) *Records maintained to regulate the flow of goods and people across borders.*

These include records of imports, exports, immigration and emigration. The coverage and content of the resulting administrative records depend on the particular laws and regulations to be enforced, and on the success of their enforcement. Typically such laws are well enforced. Immigration records, by definition, exclude illegal immigrants but otherwise are complete. However, since emigration from Canada is not controlled, no direct administrative emigration records exist. Administrative records on Canadian imports tend to be more accurate than those on exports since the former require more detailed documentation in order to assess their liability for duty.

(2) *Records resulting from legal requirements to register particular events.*

Examples include births, deaths, marriages, divorces, business incorporations or amalgamations, licensing, etc. Typically coverage and quality of records collected for this purpose are very high in Canada, since evidence of this type of registration is necessary to obtain rights or benefits.

(3) *Records needed to administer benefits or obligations.*

Examples include taxation, unemployment insurance, pensions, health insurance, and family allowances. The coverage and content of these records are highly program dependent. The population to which they apply may be very well covered, but for political or administrative reasons the definition of this population may not be the most useful definition analytically.

(4) *Records needed to administer public institutions.*

These include, for example, records related to schools, universities, health institutions, courts and prisons. Such records tend to focus on the institutional caseload rather than on the individuals passing through the institution. On the other hand, they usually provide very complete aggregate statistics on the population using these institutions. In Canada, many administrative records in this category fall within provincial jurisdiction.

(5) *Records arising from the government regulation of industry.*

Examples include records in the areas of transportation, banking, broadcasting and telecommunications. They also include records arising from the management of the supply or the price of some commodities, especially in the agriculture area.

(6) *Records arising from the provision of utilities.*

These include electricity, phone and water services. Their coverage of subscribers and the quality of information associated with services and billing are normally good. Many of these services are administered at the provincial or municipal levels.

Administrative records also vary in terms of the processes by which they are assembled. Most administrative processes with wide coverage are now automated, but differences in hardware and data formats (both between jurisdictions, and between the administrative agency and the statistical agency) have to be faced. Increased automation also leads to an increasing amount of modification to the originally reported records by the administrative agency before they are received by the statistical agency. While enhanced control of the quality of incoming forms may be beneficial to the final quality of the administrative file, additional work is required by the statistical agency to understand and evaluate the effects of any preliminary processing carried out by the administrative agency. In some administrative systems, the individual records remain at their local source and only aggregates are assembled centrally. This practice restricts the statistical agency's ability to evaluate the quality of the data and limits flexibility in statistical analysis of the data.

Finally, records differ in terms of their accessibility. Legal and regulatory provisions often govern access to, and use of, administrative records for secondary, including statistical, purposes. This topic is addressed further in Section 6.

3. USES OF ADMINISTRATIVE RECORDS

The statistical uses of administrative records may be categorized into four main areas. Most statistical applications of administrative records fall into one of these four categories or represent combinations or variations of these uses.

(1) *Direct Tabulation*

This includes the counting of units in files, cross-classification by attribute, and the aggregation of quantitative variables associated with each unit. Statistics on vital events and on external trade are important examples. Other examples include the publication of monthly counts of unemployment insurance claimants, and of beneficiaries by province, age, sex and length and type of benefit, and annual summaries of income distributions for each county based on the personal income tax file.

(2) *Indirect Estimation*

This category includes cases where data from administrative records comprise one of the inputs into an estimation process. For example, individual tax returns for the same taxfiler are linked from one year to the next in order to produce partial estimates of migration which can be weighted up with reference to census-based benchmarks. These estimates of migration then feed into Statistics Canada's population estimation program (which also makes use of administrative data on births, deaths and immigration). A second example is the use of taxation data for small businesses in lieu of seeking survey data from them. These tax-based data, adjusted if necessary, are combined with survey-based data for large businesses to provide industry aggregates.

Also within this category are uses that involve the linkage of different administrative or statistical files to produce estimates. For example, the linkage of the death register with files of individuals exposed to particular hazards in order to estimate differential mortality rates, or the linkage of records from tax files, unemployment insurance files, and manpower training files in order to analyse labour market attachment and adjustment.

(3) *Survey Frames*

In this category we include the use of administrative records to create, supplement or update frames to be used for censuses or surveys. A primary example is the use of payroll deduction information submitted by employers to Revenue Canada. The questionnaire which has to be completed by new payroll deduction account holders is a valuable means of identifying new businesses or changes in the structure of existing ones. Although in Canada we do not have a register of housing units, a second example would be the use of building permits or new telephone or electricity connections as signals of possible new housing units.

(4) *Survey Evaluation*

This category covers the use of administrative records for checking, validating or evaluating survey-derived data. This may be done either at the individual unit level, or at an aggregate level. Several census evaluation studies in the past have used immigration and taxation records to evaluate census questions on immigration and income, respectively, while family allowance records have been used in checking the census coverage of children.

An important determinant of how a particular administrative source will be used is the perceived quality of the administrative records compared to corresponding survey information. In some instances administrative records are used to evaluate survey responses, while in others survey-based data provide a means of benchmarking administrative-based estimates. The quality of administrative records has to be assessed in each individual case. In general, their quality for statistical purposes depends upon at least three factors:

(i) the definitions used within the administrative system;

(ii) the intended coverage of the administrative system;

(iii) the quality with which data are reported and processed in the administrative system.

Weaknesses in any of these three factors can affect the statistical usefulness of the administrative records. The timeliness with which they are available is also an important consideration. Some of the potential limitations that need to be considered when deciding on the statistical use of administrative records have been described elsewhere (e.g., see Brackstone 1984). The strengths and weaknesses of administrative records compared to those of censuses and surveys are summarized in Table 1.

To illustrate the utilization of administrative records in Canada we will describe two areas of application within Statistics Canada. The first deals with the production of business statistics; the second addresses the production of statistics on individuals and families.

4. ADMINISTRATIVE DATA AND BUSINESS SURVEYS

Statistics Canada is currently in the throes of a complete redesign of the infrastructure and strategy on which its business surveys program is based. In particular this involves the redesign of the business register (the frame for business surveys), the re-thinking of the role

Table 1

Comparison of Censuses, Surveys and Administrative Records as Sources of Statistical Data

Factors	Censuses	Surveys	Administrative records
1. Coverage	Aim at complete coverage of the population	Some surveys exclude certain sectors of the population (e.g., Indian reserves, remote areas)	Target populations are defined by administrative requirements
2. Content	Wide range of data items allows extensive cross-classification	Usually covers a narrow range of topics but in more depth than a census	Restricted to variables required for administrative purposes
3. Concepts/definitions	Can be based on the requirements of social and economic analysis	Can be based on the requirements of social and economic analysis	Defined by administrative requirements
4. Small area estimates	Available as a result of aim at complete coverage	Unavailable in most cases	Available, provided individual records are geographically coded to small areas
5. Quality control	Can be designed to minimize errors	Smaller size allows for even tighter control than in censuses	Under the control of the administrative agency and may not receive attention except for key variables
6. Cost	Expensive	Relatively low cost per survey, although the cumulative cost of a regular survey over a 5-year inter-censal period may be large	Relatively inexpensive if initial collection costs attributed to the administrative program
7. Frequency	Every 5 or 10 years (depending on topic)	May be annual, quarterly or monthly depending on topic	May be annual or monthly depending on administrative program
8. Timeliness	Data available six months to 2½ years after Census Day	Repeated regular surveys produce results in a few weeks. Ad hoc surveys may require several months	Dependent upon the administrative process. An annual file may not be available in a clean form until well into the following year
9. Stability	Changes are under the control of statisticians who respond to user needs	In repeated surveys, changes are infrequent to allow comparisons over time	Changes may occur due to legislative or regulatory change, or due to changes in administrative practice
10. Respondent burden	Heavy but infrequent. Reduced through the use of sampling	Light on average, though heavy for those selected	No additional burden

and use of tax data within the program, and the development of a consistent strategy for the design of both annual and sub-annual business surveys. This redesign was motivated by needs to:

(a) overcome some noticeable data quality weaknesses in the current program;

(b) better integrate data from different surveys;

(c) minimize respondent burden by making maximum use of tax data;

(d) reduce resources required for maintaining survey frames.

A more detailed description of this project can be found in Colledge (1987).

Income tax and payroll deduction data play a prominent role in the conduct of business surveys. Annual tax returns submitted by corporations (T2) and by individuals (T1) are available to Statistics Canada under the Statistics Act. The payroll deductions of income taxes by employers are also available. Statistics Canada makes use of these data from business for two distinct purposes:

(i) maintenance of its frame of businesses;

(ii) substituting income tax data for survey data.

4.1 Frame Maintenance

The maintenance of a frame of businesses is a complex task. This complexity stems primarily from the complex structure and inter-relationships of many businesses, particularly large ones, and from the difficulty of keeping track of the very large number of births and deaths occurring among small business. The term "business" itself needs careful definition. In fact a distinction must be made between legal structures (incorporated companies, etc.), operating structures (the way companies organize and operate themselves), and statistical structures (the units for which data are required for analytic purposes). A hierarchy of units can be defined within each of these structures. In the case of the statistical structure, Statistics Canada has defined a hierachy comprising, from top down, enterprises, statistical companies, establishments and locations. The task of frame maintenance thus involves not only updating for births and deaths but also keeping track of changes in the relationships between the various units within complex businesses, including the relationships between the statistical and operating hierarchies.

The proposed frame strategy calls for the continuous maintenance of the current corporate structure of all companies above a certain threshold size (which varies with industry), including the relationship of this structure to tax reporting units. Companies updated in this way will account for at least 70% of economic activity in each industry.

An activity known as "profiling" is used to determine the internal structure of complex businesses. This involves interviewing officers of the business to understand their operating structure and identify the appropriate statistical units. An important source of information on changes to business (births and restructuring) is Revenue Canada's payroll deduction (PD) system. The activation of a new PD account by an employer is treated as a signal that something has happened. Such signals are followed up with the business to identify whether a frame update is required. Other signals will be obtained from annual tax returns, from responses to regular surveys, and from routine profiling.

In the case of smaller companies, where the structure is usually simpler but the turnover is faster, no attempt is made to define the various types of unit and their inter-relationships.

Instead, administrative data are used directly. Two alternative lists of businesses are made available as a basis for surveying - one is the most recent set of annual tax returns; the second is the current set of PD accounts. In both cases, all units above the threshold are removed. These two lists overlap and the most appropriate one is used in each particular survey. The PD-based list, which is more current since PD accounts may be opened or closed at any time during the year, is preferred for sub-annual surveys. It has the disadvantage of excluding non-employers.

4.2 Substituting for Survey Data

In the interests of minimizing both response burden and costs, tax data are used to replace survey data where feasible. The concepts and definitions underlying tax data do not uniformly coincide with the survey definitions required to assure consistency in the System of National Accounts or for other analytic purposes. Therefore care has to be taken in selecting from tax returns the data items that come closest to the required survey definitions. Furthermore, tax data do not contain the full range of variables required by many annual business surveys. In particular, they lack production statistics.

A further problem in utilizing tax data lies in establishing the relationship between the unit for which a tax return is submitted and the unit(s) to be surveyed. This is a problem particularly for the large complex businesses referred to earlier.

The strategy that has been developed for annual surveys is to make use of tax data primarily for small businesses where there is usually a one-to-one relationship between the taxfiler and the business. This approach significantly reduces the response burden on small businesses, without unduly affecting the quality of final data, since the bulk of economic activity is reported through the survey returns of larger companies.

It is clear from this brief overview of the new business survey strategy and infrastructure that there is a fundamental dependence on tax data for the continuing functioning of the program. This requires a very close working relationship between Statistics Canada and Revenue Canada so that the impact of administrative and procedural changes in the tax system can be assessed and prepared for in advance.

5. SOCIO-ECONOMIC DATA FROM ADMINISTRATIVE SOURCES

A systematic effort to develop data on individuals, families and households from administrative records was initiated in the late 1970s. The original motivation for this work was the rising costs of census-taking and the search for cheaper alternatives. It quickly became apparent that the statistical potential of administrative records on individuals in Canada lay in supplementing the quinquennial census through the provision of data for small areas intercensally, rather than in replacing the census. It is not possible to achieve the coverage, geographic specificity, and range of individual, family and household characteristics required from a census with the existing administrative record systems. Nevertheless, the emulation of census coverage using a combination of administrative record systems is being pursued, together with the study of the possibility of replacing some census questions with data derived from administrative sources.

This section will concentrate on the use of administrative records to supplement census data inter-censally. The focus of the developmental work has been on administrative record systems that are national in scope (e.g., income tax, unemployment insurance (UI), family allowance, old age security) rather than systems that are administered at provincial or lower levels (e.g., health insurance, driver's licences, municipal assessments). In the latter case the problem of standardization across jurisdictions is added to the other problems inherent in the statistical use of administrative records.

The annual individual tax file (T1) has proven to be the principal source of statistical data on individuals. The first use of this file was its direct tabulation to produce statistics on income and labour force participants by age and sex for provincial and sub-provincial areas. Identification of geographic location of taxfilers is based on the postal code indicated on the record. A file that provides a conversion from postal code to the various levels of census geography (province, county, municipality, electoral district, etc.) has been developed. Special tabulations can also be produced for user-defined areas described in terms of postal codes.

Data derived in this way are, of course, based on the concepts, definitions and regulations implicit in the Income Tax Act. These may not conform to definitions desired for analytic purposes (e.g., some forms of social assistance which are not taxable may be excluded). Income can be broken down by source - in particular, employment income can be separated. Variables available for cross-classification are limited (e.g., age, sex and marital status). Occupation, though asked on the tax form, is not reported nor coded with sufficient quality to be statistically useful. The coverage of these data is limited by the need to file a tax return. Low income individuals and dependents are therefore under-represented. Over time, changes to tax law can have a significant impact on coverage; e.g., the introduction of the Child Tax Credit, that required low income earners to file a tax return in order to claim the credit, led to a marked increase in coverage in 1978 compared to the previous year.

Despite these reservations, data produced by direct tabulation from income tax files provide a useful inter-censal source of small area income data. A recent publication from Statistics Canada made use of this source to produce data for Forward Sortation Areas, i.e., the first three characters of the postal code (Statistics Canada 1987). Since a prime concern in the publication of data for small areas is to ensure that no individual data can be deduced from aggregate totals for small areas, data are not provided for areas with less than 100 taxfilers.

A second use of the individual tax file is for estimating annual migration. This is achieved by matching individuals on tax files for two successive years and comparing the Census Division (or county) code assignment for each year. If there has been a change in code, it is assumed that the taxfiler has migrated. Demographic and tax exemption information are used to estimate the total number of persons who have migrated with the taxfiler. In a final stage, since the tax file does not cover the whole population, an adjustment is made to estimate the total number of migrants from year to year. Since 1981, tax-based migration estimates have been used in Statistics Canada's population estimates program. A full description of the methodology for estimating migration from tax records can be found in Norris and Standish (1983).

While data on individual incomes can be derived from tax data as described earlier, more analytic and policy interest focuses on family income. To derive family income from the individual tax file requires the capacity to identify and match records of individuals belonging to the same family. Development of family income data in this way has been proceeding with encouraging results. A description of methodology and results can be found in Auger (1987).

A second important administrative source of data on individuals is the unemployment insurance (UI) system. Files of both claimants and beneficiaries are available to Statistics Canada. The UI claimant and beneficiary files contain individuals who, for a variety of reasons, may be entitled to UI benefits. Not all of these individuals are considered to be unemployed according to the standard international definition of unemployment as incorporated in the Labour Force Survey (LFS), the source of published unemployment rates.

If a closely corresponding category in the UI system can be found, these files can be used to tabulate counts of "unemployed" for small areas. However, since even the best choice of category in the UI system does not correspond exactly with the definition of "unemployed" used in the LFS, attention has to be focused on how to integrate or reconcile these two sources of data. For example, monthly counts for small areas from the UI system might be used as indicators of changes in unemployment at the local level which could be calibrated to reliable LFS estimates at a higher geographic level (e.g., the province). Various methods of estimation along these lines have been investigated (e.g., regression estimation, SPREE - structure preserving ratio estimation), though without as yet any final conclusion as to the most appropriate method. A description of this work can be found in Trottier and Choudhry (1985) while Feeney (1987) describes a similar approach in the Australian context. A time series modelling approach which exploits the correlated structure of the error over time appears very promising (Choudhry and Hidiroglou 1987).

These examples have illustrated that, in the case of statistics on individuals, the primary uses of administrative records are for direct tabulation and as input into estimation processes. This contrasts with the examples from the business side where frame maintenance and substitution for survey responses were the main uses.

While these two examples represent two important developing areas of administrative record use in Statistics Canada, they cover only a small fraction of the administrative files used by the Agency. There is, for example, a widespread and long-standing use of administrative records in the social institutions area (education, health, justice) both for creating survey frames and for obtaining statistical data. Current developmental work on telephone surveying and on address registers is using administrative records to develop frames of dwellings or households. A recent internal survey identified more than 50 administrative systems being used for statistical purposes. These covered the full range of types and uses described in Sections 2 and 3, and included examples from areas as varied as disease registries, motor vehicle licences, aircraft landings, milk marketing boards, fuel sales tax, municipal construction records, and customs and excise.

6. ACCESSING AND INFLUENCING ADMINISTRATIVE SYSTEMS

It is clear from this review of the use of administrative records for statistical purposes, that administrative records are a vital input to many of Statistics Canada's programs. This leads to a consideration of measures the Agency can take to protect the supply of data from administrative sources, and perhaps to make them more useful for statistical purposes. In this section we will deal with the two primary issues of obtaining access to administrative records, and influencing their content, design or associated procedures.

6.1 Access

The legal authority for access to administrative records is provided by Section 12 of the Statistics Act (1971):

> "A person having the custody or charge of any documents or records that are maintained in any department or in any municipal office, corporation, business or organization, from which information sought in respect of the objects of this Act can be obtained or that would aid in the completion or correction thereof, shall grant access thereto for those purposes to a person authorized by the Chief Statistician to obtain such information or such aid in the completion or correction of such information."

While this provision appears to give fairly broad access rights, it is not without limitations. In some cases, legislation governing the administrative process places restrictions on access or secondary use of the administrative data. This leads to a confrontation of legislation that will at best delay the negotiation of access. In some cases, access for statistical purposes is specifically permitted.

Enabling legislation is a necessary but not sufficient condition for the productive utilization of administrative records. A co-operative approach to the development and utilization of administrative records for statistical purposes is likely to be far more effective in obtaining access to administrative records than an approach involving legal arguments and sanctions. Indeed, once access is obtained, the subsequent step of influencing design or procedures is only achievable if there is a spirit of co-operation between the administrative and statistical agencies.

Access to administrative records by Statistics Canada is strictly a one-way street. Individual micro-data are provided from the administrative agency to the statistical agency, but only confidentiality-protected aggregate data can flow back. The only exception to this rule is the case where the administrative agency depends on the statistical agency to organize, format, edit, process, or restructure its records, and a version of the original micro-data is passed back to the supplying agency.

6.2 Influencing Change

We have already alluded to the potential impact of changes in administrative regulations or practices on resulting statistics. Discontinuities in time series based on administrative records can be caused by simple changes in the coverage of a program, the introduction of an incentive to join or leave a program, or procedural changes that affect quality or completeness of records. Thus the statistical agency has to guard against, and react to, externally imposed changes.

There are other kinds of changes that the statistical agency might like to see implemented. A frequent frustration of the statistician trying to use administrative records is the feeling that the administrative records could be so much more useful if only relatively minor changes were made. For example, the addition of an extra question, the use of a different concept, the coverage of an additional subgroup, or the introduction of a quality check might significantly enhance the statistical value of the records. On the other hand, why should the administrative agency contemplate changes not required for the primary administrative process, changes which would probably in some measure add to the cost and complexity of the administrative process?

The challenge for a statistical agency is to persuade the administrators that the benefits from such a change outweigh any additional administrative costs. This is made harder to the extent that the benefits do not accrue to the department responsible for the administrative system, but to separate policy-making departments and other statistical users.

It is usually easier to build statistical requirements into a system from its inception than to make changes to a system that is already operational. Therefore, a mechanism that would allow statistical requirements to be considered during the design, or the major redesign, of an administrative system is preferable to one that only tries to adjust existing systems. A topical case in Canada is in the area of tax reform, currently under consideration by the government. This could significantly change the collection of business data in Canada. Involvement of statisticians in the design of such a system could greatly enhance the statistical benefits derived from the system. Of course, the institution of a new administrative system is a relatively rare occurrence, so that adjustment to existing systems is also necessary if statistical benefits are to be obtained in the short run. On the other hand, the comparative rarity of design or redesign of major administrative systems strengthens the argument for not missing opportunities to influence such exercises when they do arise.

6.3 Mechanisms

A variety of measures or mechanisms, some bilateral involving the statistical agency and a specific administrative department, others of a broad government-wide nature, can assist the statistical agency in accessing and influencing administrative systems. These include:

(i) bilateral committees at a senior level to review and discuss issues of mutual interest, including problems related to the supply of administrative data;

(ii) feedback of statistical data to the administrative agency to demonstrate both usefulness of the data and, perhaps, weaknesses arising from administrative practices;

(iii) provision of technical advice or services in support of the administrative agency's own statistical activities;

(iv) a government information collection policy that requires, for example, any data collection activity plan (statistical or administrative) to be reviewed by a central agency;

(v) statistical planning in the form of a requirement that each new program proposal include a plan for acquiring the statistical information needed to monitor and evaluate the program;

(vi) promotion of the use of standard statistical definitions (e.g., family, business establishment, unemployed) in administrative systems;

(vii) audits that identify the use of administrative records as a cost-efficient alternative to other means of acquiring information;

(viii) political instruction to make greater use of particular administrative systems or seek alternatives to survey-taking;

(ix) removal of legislative impediments to access or use of administrative records for statistical purposes.

Statistics Canada's experience in dealing with other federal government departments has been most successful in cases where close bilateral arrangements have been developed. The introduction of senior bilateral committees in the early 1980s was supportive of such arrangements, and in some cases instrumental in creating them. Government-wide measures such as information management and statistical planning have been less successful in facilitating administrative record use. Government audits and cabinet directives have provided impetus to activities aimed at increasing administrative data use, but the increased use itself is again dependent upon close working relationships with particular departments. While it is convenient to characterize the statistical agency as the progressive agency trying to break down unreasonable barriers to administrative data use, it must also be recognized that there may be inertia to the associated changes within the statistical agency itself. Staff whose careers have been based on survey design and survey-taking may need convincing that budgetary restrictions and data needs now necessitate combining these with other approaches.

Since the above comments have focused on federally administered systems, we will add a few words about provincial records. While some of the above measures apply equally to provincially administered records, the fundamental problem in dealing with subnational jurisdictions is that of adherence to common standards. Differing provincial needs and priorities, facilitated by increasing technological capacity, will lead to divergent administrative systems in the absence of any centralizing force. Statistics Canada has used a variety of mechanisms in the past in attempts to encourage conformity, but with only mixed success. As with federal government custodians of administrative records, mutual benefit has to be the major incentive to conformity. Federal-provincial committees exist in several subject areas. The Vital Statistics Council, consisting of provincial registrars of vital events and representatives of Statistics Canada, is a successful and long-standing example. Such committees have developed and monitored conventions for reporting certain data items in the past. For example, the framework for municipal finance reporting was developed as a result of federal-provincial meetings on municipal financial statistics.

7. CONFIDENTIALITY, PRIVACY AND PUBLIC RELATIONS ISSUES

Even with the legal authority to exploit administrative records and co-operative administrative agencies to supply them, careful consideration has to be given to the public perception of the use of administrative records beyond their original purpose. Since the effectiveness, if not the survival, of a statistical agency depends critically upon the continuing co-operation and trust of respondents, it must take extreme care before embarking on any activity with the potential to undermine that co-operation or trust.

Public awareness and concern over privacy and related issues of information access and control have risen in many countries in recent years. In Canada, passage of the Privacy Act in 1982 bore witness to this mounting concern. The Privacy Act requires, *inter alia* and with some exceptions, that an index of all personal information banks under the control of federal government institutions be published periodically, that individuals have the right of access to information about themselves contained in such information banks, and that personal information be used only for purposes consistent with the purpose for which it was obtained. One of the exceptions to this last provision is that personal information may be disclosed

> "... to any person or body for research or statistical purposes if ... the purpose for which the information is disclosed cannot reasonably be accomplished unless the information is provided in a form that would identify the individual to whom it relates, and ... a written undertaking (is obtained) that no subsequent disclosure of the information will be made in a form that could reasonably be expected to identify the individual to whom it relates." (Privacy Act 1982 Section 8(2)(j)).

This provision covers the use of administrative records for statistical purposes as far as the Privacy Act is concerned. However, this Section is subject to any other Act of Parliament so that a clause forbidding such use in an Act governing an administrative process would have precedence.

While the Privacy Act and other Acts recognize statistical work as a legitimate secondary use of administrative records under certain conditions, this alone will not allay public concern over the existence of data banks that could be used to an individual's detriment. It is doubtful whether the average citizen appreciates the distinction between statistical use, where the identity of the individual record is of no lasting interest, and administrative use, where the essence of the individual record is the particular unit to which it relates. It would be easier to explain and utilize this distinction if we could state unequivocally that identifiers are never needed for statistical purposes. Unfortunately this is not the case. Several legitimate statistical techniques do require identifiers in intermediate data manipulations. These techniques all involve some form of matching data from different files or different occasions, and identification is required to ensure that the correct records are matched. Once the matching has been accomplished the records can be anonymized provided no subsequent linkage is planned. Examples include the requirement for names in a population census to ensure coverage and permit coverage measurement, longitudinal studies using administrative records, epidemiological investigations, and evaluation studies to check survey responses against administrative sources. Explaining why identifiers are needed when identity is of no interest is an interesting challenge facing the statistical agency.

A further source of concern may relate to the undertaking of confidentiality itself. Despite Statistics Canada's record of confidentiality protection there are doubtless respondents who are skeptical about the protection their information enjoys. This concern may be heightened by the use of enumerators who are known to respondents, particularly in small communities. Some respondents seem to assume there is a high degree of information exchange actually taking place between federal departments, and in some cases do not distinguish between different departments of government.

An additional concern may relate, not to the trustworthiness of the present custodians of information banks, but to a fear that personal information cannot be protected against future violation, either illegally, or by a legitimate elected authority with different views on privacy. Protection against this possibility would require the removal of all identifying information from statistical data bases.

This public concern over privacy and the manipulation of personal information requires the statistical agency to consider measures it can take to prevent or minimize negative public reaction to its legitimate use of administrative records for statistical purposes. Since this is essentially an issue of public perception, it is important that the statistical agency be open about its practices, and that any of the following measures that are implemented are clearly visible to the interested public.

(a) Public communications to respondents and users should continually stress the importance attached to confidentiality of all individual (micro) data acquired by the statistical agency.

(b) The one-way nature of micro-data flow should be stressed. Micro-data flow into the statistical agency, but only confidentiality-protected aggregates or summaries flow out. This applies equally to survey or census data and data from administrative records.

(c) The benefits of administrative record use in terms of reduced respondent burden and savings to the taxpayer should be emphasized. Such claims should be supportable by real measures of cost and respondent burden savings.

(d) An explicit and public policy on record linkage stipulating the conditions under which the statistical agency will undertake such activities can be helpful both in demonstrating careful consideration and control of linkage activities, and in forestalling linkage requests that would violate the conditions.

(e) The Privacy Act requires that individuals be informed of the purpose for which any personal information is being collected. Administrative agencies should be encouraged to ensure that statistical purposes are included in such statements. Even though statistical purposes may be a permissible secondary use of administrative records, their explicit mention on the collection form will serve to avoid subsequent surprise.

(f) The physical security that surrounds the use of sensitive administrative records should be clearly visible, and perhaps even tighter than that in use generally within the Agency. For example, in Statistics Canada, the divisions having primary custody of tax data are housed in limited access areas within buildings that are themselves subject to security checks on entry.

(g) Exemption of statisticial files from examination by security or intelligence services is an important element in maintaining public trust in the absolute confidentiality of data provided to the statistical agency. An exemption for Statistics Canada data (the sole institutional exception within government) was provided when the new Canadian Security and Intelligence Service was formed in 1983.

While the above points represent some specific measures that can be taken to avoid or respond to public reaction to the use of administrative records, ultimately the statistical agency must have strong political support for this kind of activity. The political credit to be gained from demonstrated reductions in costs and respondent burden, coupled with strong political assurances of the protection of individual data, provide a strong platform for politicians to dispel public concern over the use of administrative records for statistical purposes. At the same time they must immediately and unambiguously confront and correct any suggestion that statistical records be used for administrative purposes.

8. CONCLUSION

Administrative records are and will continue to be an increasingly important source of statistical data. The relative strengths and weaknesses of data derived from administrative systems, in terms of cost, coverage, quality, relevance and timeliness, in comparison to census- or survey-based data, dictate the manner in which these sources of data are most effectively used. Current uses of administrative records include direct tabulation, indirect estimation, substitution for survey responses, frame construction and maintenance, and data evaluation. These uses now permeate most statistical programs and can be expected to extend even further in the future.

In Canada, administrative records have become part of the fabric of our statistical system. Their use has been one of the means by which Statistics Canada has been able to maintain its programs in the face of declining budgets. In the process, respondent burden has been reduced and new, or more frequent, data series have become available. Since we do not have administrative registers as such, considerable attention has been paid to issues of coverage and the joint use of both administrative and survey-based data to ensure valid estimation of universe totals. The use of record linkage techniques, though requiring careful controls, has proven to be very valuable, particularly for business data, longitudinal labour market studies, and epidemiological work.

With the growing use of administrative records, statistical agencies are becoming increasingly dependent upon other agencies for the uninterrupted flow of input data to their statistical programs. Whatever the legislative and policy environment in which the statistical agency operates, the establishment of close co-operative arrangements with supplying agencies is crucial. The ability of the statistical agency to influence the design or redesign of administrative systems rests on a mutual understanding of the requirements of the two agencies. Establishment of a government-wide policy or principle that the statistical agency should have a voice in decisions regarding the design of administrative systems, or more generally, in proposals for meeting the statistical needs of new programs, can help the statistical agency in this regard, but is no substitute for the fostering of close co-operation with administrative agencies.

A variety of mechanisms can be considered to assist the statistical agency in gaining the access and influence it requires within the government system. The applicability and effectiveness of each mechanism will depend upon the underlying legislative and political climate, and on the mandate and status of the statistical agency within the government apparatus. Statistics Canada's experience has been that close bilateral working relationships with administrative departments, based on a principle of mutual benefit, is the most effective approach. Political support for the use of administrative records is important and has been forthcoming through recent government decisions related to budget reductions.

ACKNOWLEDGEMENTS

This paper has drawn upon ideas and work of many people at Statistics Canada. In particular, parts of this paper incorporate material prepared in collaboration with Michael Colledge, Ivan Fellegi and John Leyes.

REFERENCES

AUGER, E. (1987). Family data from the Canadian personal income tax file. Paper presented at 1987 American Statistical Association meetings.

BRACKSTONE, G.J. (1984). The impact of technological change on census-taking, *Estadistica*, 36, 43-60.

CANADA, STATISTICS ACT (1971). *Statutes of Canada 1970-71-72*, c.15.

CANADA, PRIVACY ACT (1982). *Statutes of Canada 1980-81-82*, c.11.

CHOUDHRY, G.H., and HIDIROGLOU, M.A. (1987). Small area estimation: Some experiences at Statistics Canada. *Proceedings of the 46th Session of the International Statistical Institute*, (forthcoming).

COLLEDGE, M.J. (1987). The Business Survey Redesign Project: Implementation of a new strategy at Statistics Canada. *Proceedings of the Third Annual Research Conference, U.S. Bureau of the Census*, (forthcoming).

FEENEY, G.A. (1987). The estimation of the number of unemployed at the small area level. In *Small Area Statistics, An International Symposium*, (Eds. R. Platek, J.N.K. Rao, C.E. Sarndal and M.P. Singh), New York: John Wiley.

NORRIS, D.A., and STANDISH, L.D. (1983). A technical report on the development of migration data from taxation records. Working Paper, Statistics Canada.

ROWEBOTTOM, L.E. (1978). The utilization of administrative records for statistical purposes. *Survey Methodology*, 4, 1-15.

STATISTICS CANADA (1987). *Urban FSA and rural postal code summary data*. Catalogue No. 17-602, Statistics Canada.

TROTTIER, I., and CHOUDHRY, G.H. (1985). Model based unemployment estimates for small areas. In *Small Area Statistics, An International Symposium '85 (Contributed Papers)*, (Eds., R. Platek and M.P. Singh), Laboratory for Research in Statistics and Probability, Carleton University/ University of Ottawa.

Survey Methodology, June 1987
Vol. 13, No. 1, pp. 45-62
Statistics Canada

Statistical Properties of Crop Production Estimators

CAROL A. FRANCISCO, WAYNE A. FULLER, and RON FECSO[1]

ABSTRACT

The National Agricultural Statistics Service, U.S. Department of Agriculture, conducts yield surveys for a variety of field crops in the United States. While field sampling procedures for various crops differ, the same basic survey design is used for all crops. The survey design and current estimators are reviewed. Alternative estimators of yield and production and of the variance of the estimators are presented. Current estimators and alternative estimators are compared, both theoretically and in a Monte Carlo simulation.

KEY WORDS: Crop surveys; Yield estimation; Two phase sample; Variance estimation.

1. INTRODUCTION

The National Agricultural Statistics Service (formerly known as the Statistical Reporting Service), U.S. Department of Agriculture, conducts objective yield surveys of corn, cotton, soybeans, rice, grain sorghum, sunflowers and wheat in states which are major producers of these field crops. Similar yield surveys are conducted in a number of other countries.

While field sampling procedures for each crop differ in terms of plot sizes, plot location methods, and vegetative and fruit measurement techniques, all surveys rely on the same basic design. A four-step sampling procedure is used. A description of this survey design is contained in Section 2. Section 3 describes the estimators of average crop yield and the variance estimators, evaluates them and explores alternative estimators. Conclusions and recommendations are presented in Section 4.

2. OBJECTIVE YIELD SURVEY DESIGN

The first two steps of sample selection produce the sample of area segments used in the June Enumerative Survey conducted by the National Agricultural Statistics Service (NASS). The area frame for each state is stratified by land use. For example, the State of California is divided into 12 land use strata. Each land use stratum is subdivided into areas called frame units. The size of a frame unit varies; the actual size of any given frame unit depends upon available boundary designations, available ancillary information, political boundaries, and so forth. Once frame units are established, the number of area segments in each frame unit is determined by dividing the total area of each frame unit by the target segment size. The target size is a function of the land use stratum into which the frame unit falls. For example, in California the target segment size is one half square mile in the orchard stratum and one square mile in all other cropland strata. Frame units typically contain between one and 30 area segments.

[1] Carol A. Francisco, Syntex Laboratories Inc., 3401 Hillview Avenue, Palo Alto, California 94304; Wayne Fuller, Department of Statistics, Iowa State University, Ames, Iowa 50011; and Ron Fecso, Survey Research Branch, National Agricultural Statistics Service, U.S. Department of Agriculture, Washington D.C. 20250.

Each land use stratum is substratified on the basis of geography. To develop the geographic substrata, frame units within each land use stratum are ordered by county in such a manner that adjacent counties that are agriculturally similar are placed together (Fecso 1978). Substrata are formed from sequential groups of area segments. Thus, substrata contain area segments that are agriculturally similar and geographically close together. Within a given land use stratum, substrata have an equal number of segments and equal area (within rounding). Detailed information on the area frame design is available in Fecso and Johnson (1981) and Houseman (1975).

For purposes of variance estimation, it is the substrata within land use strata that are the sampling strata. Henceforth, the land use substrata will be referred to simply as strata.

The first step in sampling from the area frame is the selection of frame units within each stratum. The number of frame units allocated to a stratum depends on the agricultural nature of the stratum. Typically, eight to 15 frame units are drawn in cropland strata; whereas in agri-urban, city, and nonagricultural strata four to five frame units are drawn. Frame units within strata are selected at random with probability proportional to the number of area segments in the frame unit. At the second step, one area segment is chosen at random from each selected frame unit. Thus, each area segment within a stratum has an equal probability of selection.

Although the frame unit is the primary sampling unit for this design, because the frame units are selected with probability proportional to the number of segments and one segment is selected per sampled frame unit, the segment can be treated as the primary sampling unit. In our study, steps one and two in the sampling procedure are considered as one procedure, and the sample of segments will be treated as a stratified single stage simple random sample. Since the average sampling rate is about one percent, the finite population correction term will be ignored in our analysis.

The third and fourth steps in the sampling procedure involve the selection of fields and of plots within selected fields. As part of the June Enumerative Survey, all selected area segments are screened for fields which have been planted or are scheduled to be planted with the crop of interest. These fields are listed by segment number and order of enumeration within segment. A systematic sample of fields is selected with selection probabilities proportional to the product of the field area and the inverse of the probability of selection of the area segment in which the field is contained. Hence, the number of sampled fields per segment varies, and large fields within a segment can be selected more than once.

At the fourth and final step, two plots of roughly equal area are placed in each selected field using a random row and pace method of location. Where rows are not readily distinguishable, and in the case of wheat, a random number of paces along the field edge and a random number of paces into the field are used to locate plots. A further exception occurs in the wheat objective yield survey. For this survey the first plot is randomly located and the second plot is placed in a fixed position relative to the first plot. In the event that a large field is selected more than once during the third step of the sampling procedure, additional sets of two plots are independently sampled. Because plots are always sampled in pairs, we call the pair of plots the secondary unit. A maximum of eight plots (that is, four secondary units) per field is imposed.

3. ESTIMATION PROCEDURES

Formally, the sample is a two phase sample with subsampling in the second phase. Table 1 contains a schematic description of the sample. The phase one sample is a stratified simple

Table 1

Sampling Procedure for the Objective Yield Survey

Phase/Sampling Unit	Selection Procedure	Sampled Number[1]	Data Collected
Phase One			
Primary Sampling Unit: Segment	equal probability within strata	n_h	crop acres
Phase Two			
Primary Sampling Unit: Segment	unequal probability	K_h	crop acres, estimated production[2]
Secondary Sampling Unit: Pair of Plots	equal probability	m_{hk}	estimated production from plots

[1] Number is per stratum for primary sampling units and is per segment for secondary sampling units.

[2] Segment production is zero if the crop acreage is zero and is estimated from plot determinations if the crop acreage is positive.

random sample of segments. The phase two sample is composed of all segments with zero crop acres and a probability-proportional-to-crop-acres sample of segments with the crop. The sample of segments is the result of a probability-proportional-to-area systematic sample of first phase fields planted with the crop. A sample of secondary units, where each secondary unit is a pair of plots, is selected from the segments in the phase two sample that have the crop. Because the secondary unit is always a pair of plots, we will henceforth refer to secondary units and no longer speak of plots. We will also ignore the fact that the operational units used to locate the plots are fields and speak only of the sampled segments.

Notice that two types of segments are observed at phase two – those that have zero acres of the crop and those that have non-zero acres. The total number of second phase segments is K. The acres and the total production are known (both equal to zero) for an observed segment with zero acres. For second phase segments with positive acres, a subsample of secondary units is used to estimate production.

Let M_{hk} be the number of secondary units in segment k of the h-th stratum. Without loss of generality, M_{hk} could be assumed to be equal to A_{hk}, where A_{hk} is the crop area in segment hk. Equality requires only the choice of an appropriate scale for area.

Section 3.1 examines the yield estimator that is currently used. Conditions under which this estimator is unbiased for state average yield are investigated. A simple estimator of the variance of estimated yield is discussed in Section 3.2. Estimators of the unconditional variances of the yield and production estimators are developed in Section 3.3. A Monte Carlo study of estimators is given in Section 3.4.

3.1 Currently Used Yield and Production Estimators

Estimates of the state average yield are currently computed as though the sample were an equal probability simple random sample of secondary units. The estimator is the simple average yield of secondary units with positive acreages. That is, the estimated average yield per acre is

$$\bar{y} = D^{-1} \sum_{h=1}^{L} \sum_{k=1}^{n_h} \sum_{\ell=1}^{m_{hk}} Y_{hk\ell}\,\delta_{hk\ell}, \tag{3.1}$$

where

$$\delta_{hk\ell} = 1 \quad \text{if } A_{hk} > 0,$$

$$\delta_{hk\ell} = 0 \quad \text{if } A_{hk} = 0,$$

$$D = \sum_{h=1}^{L} \sum_{k=1}^{n_h} \sum_{\ell=1}^{m_{hk}} \delta_{hk\ell}, \tag{3.2}$$

m_{hk} is the number of sampled secondary units selected in segment hk, L is the number of strata, and $Y_{hk\ell}$ is the estimated yield per acre for secondary unit ℓ of segment hk. If the crop acreage in a segment, A_{hk}, is zero, then $m_{hk} = 1$ and $Y_{hk\ell} = 0$, by definition. The total number of observed secondary units for segments with positive acres is D.

Expression (3.1) can be written in the convenient operational form

$$\bar{y} = D^{-1} \sum_{t=1}^{D} Y_t, \tag{3.3}$$

where the subscript t replaces the triple subscript $hk\ell$ and the summation is over secondary units in segments with positive crop acres.

The estimator of average crop yield per acre (3.1) is a type of combined ratio estimator. This can be shown by using conditional selection probabilites to rewrite $\bar{y}$. In the NASS scheme, segments are selected systematically with probabilities proportional to expanded size, and segments with sufficiently large expanded acreage are included with certainty. The number of secondary units allocated to certainty segments is proportional to the size of the segment, up to rounding error. The rounding is performed by the systematic selection scheme. Let $\pi_{hk\ell}$ be the conditional probability that secondary unit ℓ in segment k of stratum h is selected, given the sample of segments selected at the first phase of the sampling procedure. We have

$$\pi_{hk\ell} = D \left(\sum_{h=1}^{L} N_h n_h^{-1} \sum_{k=1}^{n_h} M_{hk} \right)^{-1} N_h n_h^{-1} \tag{3.4}$$

for secondary units in segments with $A_{hk} > 0$, where N_h is the population number of segments in stratum h, M_{hk} is the number of secondary units in segment k of stratum h, and n_h is the number of segments in stratum h selected at the first phase. The conditional probability of observing a segment with zero acres at the second phase is one.

Then the mean estimator given in (3.1) can be written as

$$\bar{y} = \frac{\sum_{h=1}^{L} N_h n_h^{-1} \sum_{k=1}^{K_h} \sum_{\ell=1}^{m_{hk}} \pi_{hk\ell}^{-1} Y_{hk\ell}}{\sum_{h=1}^{L} N_h n_h^{-1} \sum_{k=1}^{K_h} \sum_{\ell=1}^{m_{hk}} \pi_{hk\ell}^{-1} \delta_{hk\ell}}, \tag{3.5}$$

where $N_h n_h^{-1}$ is the inverse of the first stage selection probability, K_h is the number of second phase segments drawn from stratum h, and $K = \Sigma K_h$. Given an appropriate scale, the numerator of (3.5) is an estimator of the total production and the denominator is an estimator of the total area. It can be shown that the numerator is an unbiased estimator by taking expectations, conditioning on the first phase sample units and then averaging over first phase samples. The denominator is a stratified estimator of the total number of secondary units. By the nature of the sampling, the number of sampling units is proportional to acreage and one can choose the scale so that the number of secondary units is equal to acreage. Hence, $\bar{y}$ can be viewed as the ratio of an unbiased estimator of the total production of the crop to an unbiased estimator of the total area under the crop.

To estimate total state production, NASS multiplies $\bar{y}$ *by* $\hat{A}$, where $\hat{A}$ is the estimator of total crop acreage defined by

$$\hat{A} = \sum_{h=1}^{L} N_h n_h^{-1} \sum_{k=1}^{n_h} A_{hk}. \tag{3.6}$$

Thus, the estimated total production is

$$\hat{Y} = \hat{A}\,\bar{y}. \tag{3.7}$$

3.2 Simple Variance Estimators

Under the assumption of simple random sampling of secondary units from the entire set of secondary units available at the second phase, the estimated variance of $\bar{y}$ conditional on the second phase segments is

$$\hat{V}_2(\bar{y}) = D^{-1}(D-1)^{-1} \sum_{t=1}^{D} (Y_t - \bar{y})^2, \tag{3.8}$$

where the subscript 2 on $\hat{V}$ is used to denote conditional variance and the subscript t on Y replaces the triple subscript $hk\ell$. The sum over t is the sum over the D secondary units in segments with postive acres.

Because of the simplicity of expression (3.8), it has been suggested that it be used as an estimator of the unconditional variance. It has also been suggested that the variance of the estimated total state production be estimated with

$$\hat{V}_*(\hat{Y}) = \hat{A}^2 \hat{V}_2(\bar{y}) + \bar{y}^2 \hat{V}(\hat{A}) + \hat{V}(\hat{A}) \hat{V}_2(\bar{y}), \tag{3.9}$$

where $\hat{A}$ is defined in (3.6) and $\hat{V}(\hat{A})$ is the usual variance estimator for a stratified estimated total,

$$\hat{V}(\hat{A}) = \sum_{h=1}^{L} N_h^2 n_h^{-1} (n_h - 1)^{-1} \sum_{k=1}^{n_h} (A_{hk} - \bar{A}_h)^2, \tag{3.10}$$

and

$$\bar{A}_h = n_h^{-1} \sum_{k=1}^{n_h} A_{hk}.$$

The estimator (3.9) is an estimator of the variance of a product based on an implicit assumption that $\bar{y}$ and $\hat{A}$ are uncorrelated.

Evaluation of the extent to which the estimator (3.9) tends to underestimate the variance of $\hat{Y}$ is difficult. We can express the unconditional variance of $\bar{y}$ as

$$\begin{aligned} V(\bar{y}) &= V_1 \{E_2(\bar{y})\} + E_1 \{V_2(\bar{y})\} \\ &= V_1 \{\hat{A}^{-1} \sum_{h=1}^{L} N_h n_h^{-1} \sum_{k=1}^{n_h} Y_{hk.}\} + E_1 \{V_2(\bar{y})\}, \end{aligned} \tag{3.11}$$

where $Y_{hk.} = M_{hk} \bar{Y}_{hk.}$ is the total for the k-th segment in stratum h, and E_1 and V_1 denote the expectation and variance, respectively, with respect to first phase sampling.

The estimator $\hat{V}_2(\bar{y})$ is unbiased for the second component of expression (3.11) under simple random sampling of secondary units. Because sampling at phase two of the NASS scheme is done systematically, $\hat{V}_2(\bar{y})$ is a biased estimator of $V_2(\bar{y})$. The nature and extent of this bias depends upon the correlation structure of the list used in sample selection at the second phase. Also affecting the bias in $\hat{V}_2(\bar{y})$ as an estimator of the true variance is the fact that formula (3.8) was derived under an assumption of replacement sampling at phase two. To the extent that phase two sampling is actually done without replacement (because samples are drawn systematically from the list of expanded segment acreages, a segment is sampled more than once only if it is large), $\hat{V}_2(\bar{y})$ will overestimate $V_2(\bar{y})$.

The estimator $\hat{V}_*(\hat{Y})$ contains no estimator of $A^2 V_1 \{E_2(\bar{y})\}$, and this produces a negative bias. However, estimation of that component is not easy, even under the simplifying assumption of probability-proportional-to-size sampling at phase two. Because of these considerations, the performance of $\hat{V}_*(\hat{Y})$ will be studied by Monte Carlo methods in Section 3.4.

3.3 Alternative Estimators of Variance

An alternative approach to the estimation of $V(\bar{y})$ is to view the sample as a two phase sample, as shown in Table 1, and to assume that the unconditional probability of selecting a segment to receive a secondary unit is proportional to the conditional probability given the first phase segments.

Let π_{hk} be the conditional probability that segment k in stratum h is included in the second phase, given the first phase sample of segments. We have

$$\pi_{hk} = min\,(1, M_{hk}\pi_{hk\ell}), \tag{3.12}$$

where $\pi_{hk\ell}$ is a constant within segment hk. If $\pi_{hk} = 1$ and the segment is selected to receive more than one secondary unit, it is assumed that the secondary units are independently drawn.

Let π_{hk}^{*} be the unconditional probability that an observation is made on segment k in stratum h at phase two. If $A_{hk} = 0$, then π_{hk}^{*} is the unconditional probability that segment hk is selected to receive at least one secondary unit. If $A_{hk} = 0$, then π_{hk}^{*} is equal to the probability that segment hk is selected at the first phase of sampling. Let

$$\begin{aligned} \pi_{hk}^{*} &= \frac{n_h}{N_h} && \text{if } A_{hk} = 0\,, \\ \pi_{hk}^{*} &= \pi_{hk}\frac{n_h}{N_h} && \text{if } 0 < \pi_{hk} < 1\,, \end{aligned} \tag{3.13}$$

where π_{hk}, defined in (3.12), is the conditional probability that the hk-th segment is selected in phase two, given the first phase sample.

In our analysis we assume the π_{hk}^{*} to be fixed. This will be so and the probability π_{hk}^{*} will be the true unconditional probability if π_{hk} is a specified multiple of M_{hk} where the multiple is fixed before sample selection. Expression (3.13) will be an approximation if π_{hk} is a function of the segments selected at the first step of the selection procedure.

Expression (3.13) is proportional to M_{hk} for $M_{hk}\pi_{hk} \le 1$.. If $M_{hk}\pi_{hk\ell} > 1$, then the number of selected secondary units is greater than or equal to one. The correct number of secondary units to allocate to such segments to maintain a self-weighting sample of secondary units is $M_{hk}\pi_{hk\ell}$. In practice, the number of secondary units observed as a result of probability-proportional-to-size systematic sampling never differs from $M_{hk}\pi_{hk\ell}$ by more than one.

To simplify the remaining computations, we assume that the systematic sampling design contains no rounding error. In other words it is assumed that the number of secondary units observed per segment is equal to the number required for a self-weighting sample. Thus, it is assumed that the number of secondary units observed in a segment drawn as part of the second phase of sampling is

$$\begin{aligned} m_{hk} &= 1 && \text{if } 0 < \pi_{hk} < 1\,, \\ m_{hk} &= M_{hk}\pi_{hk\ell} && \text{if } \pi_{hk} = 1\,. \end{aligned} \tag{3.14}$$

Under this assumption, an unequal probability combined ratio estimator of the mean yield is equivalent to estimator (3.1). The combined ratio estimator is

$$\bar{y}_r = \hat{M}_r^{-1} \sum_{h=1}^{L} \sum_{k=1}^{K_h} \pi_{hk}^{*-1} M_{hk} \bar{y}_{hk.}\,, \tag{3.15}$$

where

$$\begin{aligned} \bar{y}_{hk.} &= m_{hk}^{-1} \sum_{\ell=1}^{m_{hk}} Y_{hk\ell} && \text{if } A_{hk} > 0\,, \\ \bar{y}_{hk.} &= 0 && \text{if } A_{hk} = 0\,, \end{aligned}$$

$$\hat{M}_r = \sum_{h=1}^{L} \sum_{k=1}^{K_h} \pi_{hk}^{*-1} M_{hk} .$$

In expression (3.15) and the remaining expressions of this section, the reader can read $\hat{A}_r$ (total area) for $\hat{M}_r$ (total secondary units), if so desired.

In the following discussion, replacement sampling of segments with probabilities proportional to the area of a crop within the segment is assumed as an approximation to the probability-proportional-to-size systematic sampling scheme of the second phase. An estimator of the variance of $\bar{y}$ under the assumption of replacement sampling is

$$\hat{V}(\bar{y}_r) = \hat{M}_r^{-2} \sum_{h=1}^{L} K_h (K_h - 1)^{-1} \sum_{k=1}^{K_h} (\pi_{hk}^{*-1} u_{hk} - \bar{u}_{h.})^2 , \tag{3.16}$$

where

$$u_{hk} = M_{hk} (\bar{y}_{hk.} - \bar{y}_r) ,$$

$$\bar{u}_{h.} = K_h^{-1} \sum_{k=1}^{K_h} \pi_{hk}^{*-1} u_{hk} .$$

An estimator of the total production is

$$\hat{Y}_r = N \bar{M}_n \bar{y}_r , \tag{3.17}$$

where

$$\bar{M}_n = \sum_{h=1}^{L} W_h n_h^{-1} \sum_{k=1}^{n_h} M_{hk}$$

N is the total number of segments in the population and $W_h = N^{-1} N_h$. The Taylor approximation of the unconditional variance of the approximate distribution of $\hat{Y}_r$ is

$$V\{\hat{Y}_r\} = N^2 [\bar{M}_N^2 V\{\bar{y}_r\} + 2\bar{M}_N \bar{\bar{y}}_N C\{\bar{y}_r, \bar{M}_n\} + \bar{\bar{y}}_N^2 V\{\bar{M}_n\}] , \tag{3.18}$$

where $\bar{y}_r$ is given in (3.15), $\bar{M}_n$ is defined in (3.17),

$$\bar{M}_N = N^{-1} \sum_{h=1}^{L} \sum_{k=1}^{N_h} M_{hk} ,$$

$$\bar{\bar{y}}_N = \left(\sum_{h=1}^{L} \sum_{k=1}^{N_h} M_{hk} \right)^{-1} \sum_{h=1}^{L} \sum_{k=1}^{N_h} Y_{hk.} ,$$

$Y_{hk.} = M_{hk} \bar{Y}_{hk.}$ is the total for the k-th segment in stratum h, and $C\{\bar{y}_r, \bar{M}_n\}$ is the covariance between $\bar{y}_r$ and $\bar{M}_n$.

Under the unequal-probability-fixed-take procedure, the estimator $\bar{y}_r (\doteq \bar{y})$ is approximately conditionally unbiased for the mean yield for the $n = \Sigma n_h$ segments in the first phase sample. The mean yield of the n segments is

$$\bar{\bar{y}}_n = \bar{M}_n^{-1} \sum_{h=1}^{L} W_h n_h^{-1} \sum_{k=1}^{n_h} Y_{hk.} .$$

Therefore, the covariance between $\bar{y}_r$ and $\bar{M}_n$ is the covariance between $\bar{M}_n^{-1}\ \bar{Y}_n$ and $\bar{M}_n$, where

$$\bar{Y}_n = \sum_{h=1}^{L} W_h n_h^{-1} \sum_{k=1}^{n_h} Y_{hk.} .$$

Using the common approximation for a ratio, the covariance between $\bar{y}_r$ and $\bar{M}_n$ can be approximated by

$$C[\bar{M}_n^{-1}\ \bar{Y}_n,\ \bar{M}_n] \doteq C\{(\bar{Y}_n - \bar{\bar{y}}_N \bar{M}_n)\bar{M}_N^{-1},\ \bar{M}_n\}$$

$$= \bar{M}_N^{-1}[C\{\bar{Y}_n,\ \bar{M}_n\} - \bar{\bar{y}}_N\ V\{\bar{M}_n\}] . \qquad (3.19)$$

If the probability of observing the pair $(Y_{hk.},\ M_{hk})$ is proportional to π_{hk}^*, an estimator of the covariance between $\bar{Y}_n$ and $\bar{M}_n$ is

$$\hat{C}\{\bar{Y}_n,\ \bar{M}_n\} = \sum_{h=1}^{L} W_h^2\, n_h^{-1}\, \hat{S}_{MYh} \qquad (3.20)$$

where

$$\hat{S}_{MYh} = K_h (K_n^{-1})^{-1} \left(\sum_{j=1}^{K_h} \pi_{hj}^{*-1} \right)^{-1} \sum_{k=1}^{K_h} \pi_{hk}^{*-1} (M_{hk} - \bar{M}_h^*)(M_{hk}\bar{y}_{hk.} - \bar{y}_{h..}^*) ,$$

$$\bar{M}_h^* = \left(\sum_{j=1}^{K_h} \pi_{hj}^{*-1} \right)^{-1} \sum_{k=1}^{K_h} \pi_{hk}^{*-1} M_{hk} ,$$

$$\bar{y}_{h..}^* = \left(\sum_{j=1}^{K_h} \pi_{hj}^{*-1} \right)^{-1} \sum_{k=1}^{K_h} \pi_{hk}^{*-1} M_{hk}\bar{y}_{hk.} .$$

The estimator $\hat{S}_{MYh}$ is constructed as a degrees-of-freedom adjustment to a Horvitz-Thompson ratio estimator of the mean of the products $(M_{hk} - \bar{M}_h)(Y_{hk.} - \bar{Y}_{h..})$. The degrees-of-freedom adjustment, the factor $K_h(K_h - 1)^{-1}$, is introduced because it is necessary to replace the population means with sample means when constructing the product.

Substituting (3.15), (3.16), and (3.20) into (3.18) gives

$$\hat{V}\{\hat{Y}_r\} = N^2[\bar{M}_n^2\ \hat{V}\{\bar{y}_r\} + 2\bar{y}_r\hat{C}\{\bar{Y}_n,\ \bar{M}_n\} - \bar{y}_r^2\,\hat{V}\{\bar{M}_n\}] , \qquad (3.21)$$

where $\hat{V}[\bar{M}_n]$ is the variance estimator for a stratified mean. Equation (3.21) is a stratified double sampling estimator of the variance of the estimated total state production. Unlike the estimator $\hat{V}_*(\hat{Y})$ of (3.9), estimator (3.21) does not assume that the yield and acreage estimators are uncorrelated. Equation (3.21) also uses an unconditional estimator of the variance of yield.

3.4. A Monte Carlo Comparison of Estimators

A Monte Carlo study was performed to illustrate the differences among alternative estimators. Cotton acreage data from the 1983 June Enumerative Survey in California and data from the corresponding 1983 objective yield survey were used as a basis for the study. For purposes of the Monte Carlo study, 28 strata were considered to have cotton.

Table 2 shows the distribution of cotton among the 28 strata as observed in the 1983 June Enumerative Survey. Fecso and Johnson (1981) describe the six different land uses, where land use is the first two digits of the stratum identification, as follows:

1300 - 50% or more cultivated land, primarily general crops with less than or equal to 10% fruit or vegetables;
1700 - 50% or more cultivated land, primarily fruit, tree nuts, or grapes mixed with general crops;
1900 - 50% or more cultivated land, primarily vegetables mixed with general crops;
2000 - 15-50% cultivated land with extensive cropland and hay;
3100 - residential mixed with agricultural lands, more than 20 dwellings per square mile;
4100 - less than 15% cultivated land, primarily privately owned rangeland.

A population was simulated from the results of the 1983 June Enumerative Survey. Table 2 compares the characteristics of the simulated population to the results of the survey. In the simulated population, cotton was determined to be present in segment k ($k = 1, \ldots, N_h$) within stratum h ($h = 1, \ldots, 28$) if $X_{hk} = 1$, where X_{hk} is an independent Bernoulli (p_h) random variable and p_h is the observed proportion of segments in stratum h found to have cotton in the 1983 June Enumerative Survey.

The next step in the creation of the population was the assignment of cotton acres to the segments for which $X_{hk} = 1$. A set of 1983 observed ratios of segment cotton acreages to the average segment acreage was compiled for land use substrata having more than one segment with cotton in the 1983 June Enumerative Survey. This set of observed ratios was used to generate the number of cotton acres in segments having cotton. If $X_{hk} = 1$, then a ratio, r_{hk}, was drawn from the set of observed ratios such that each observed ratio in the set had an equal probability of selection. The number of acres of cotton in segment hk, M_{hk}, was defined by

$$M_{hk} = r_{hk}\bar{M}_{h.}, \tag{4.1}$$

where $\bar{M}_{h.}$ was the observed average number of cotton acres for segments with cotton in stratum h in the 1983 June Enumerative Survey. (See Table 2.)

Results of the 1983 objective yield survey for cotton were used to simulate yield observations within segments. Since estimated yields were not readily accessible, an alternative variable - a major component of yield estimates - was used. This variable is the number of plants per 100 square feet. The estimated overall population mean number of plants per 100 square feet was 79.6 for the 1983 objective yield survey. Table 3 shows the average number of plants

Table 2

Cotton Acreage Estimates from the 1983 June Enumerative Survey in California and Cotton Acreages in the Simulated Population

Stratum	Target Segment Size (Acres)	Number of Segments in Stratum	Number of Segments Sampled in 1983	Percentage of Segments with Cotton		Mean Acres Cotton in Segments with Cotton	
				1983	Simulated Population	1983	Simulated Population
1314	640	291	10	60	60	197	200
1315	640	291	10	100	100	354	348
1316	640	291	10	90	89	167	173
1317	640	291	10	90	92	149	148
1318	640	291	10	50	53	481	422
1319	640	291	10	20	19	249[1]	260
1320	640	291	10	90	91	154	155
1321	640	291	10	60	61	270	274
1322	640	291	10	70	71	205	210
1323	640	291	10	80	79	288	279
1713	320	432	10	30	28	125	122
1714	320	432	10	30	31	58	57
1715	320	432	10	20	22	86[2]	84
1716	320	432	10	10	8	86[2]	89
1717	320	432	10	40	38	26	27
1718	320	432	10	30	29	144	144
1719	320	432	10	30	31	65	67
1720	320	432	10	30	30	38	35
1721	320	432	10	30	29	133	138
1722	320	432	10	50	47	130	131
1723	320	432	10	40	40	76	76
1906	640	362	10	70	73	117	127
1907	640	362	10	70	74	192	194
1908	640	362	10	80	83	253	246
2010	640	649	10	30	31	303	306
2011	640	649	10	40	41	175	165
3107	160	1,847	5	20	22	25[3]	25
4110	2,560	1,044	10	10	10	178	165

[1] Number of segments sampled was less than or equal to 2. Average of all segments in substrata within land use stratum 13 is shown.

[2] Number of segments sampled was less than or equal to 2. Average of all segments in substrata within land use stratum 17 is shown.

[3] Number of segments sampled was less than or equal to 2. Approximate acreages for this agri-urban stratum are shown.

per 100 square feet. The estimated overall population mean number of plants per 100 square feet was 79.6 for the 1983 objective yield survey. Table 3 shows the average number of plants per 100 square feet by stratum for the 1983 survey. The average for each stratum is based on all secondary units within the stratum that were drawn as part of the probability-proportional-to-estimated-size sampling scheme.

An analysis of variance of the 1983 plant data (Table 4) shows that 28 percent of the total variation among secondary units was due to between-segment differences within strata ($s_b^2 = 378.0$), whereas 58 percent of the total variation was due to variation among secondary units within segments ($s_w^2 = 776.6$). If the stratum component is treated as fixed, 67 percent of the within-segment variation is due to variance among secondary units.

Table 3

Average Number of Plants per 100 Square Feet from the 1983 Objective Yield Survey for Cotton in California and in the Simulated Population

Stratum	Average Number of Plants per 100 Square Feet	
	1983 Objective Yield Survey	Simulated Population
1314	78	76
1315	80	80
1316	67	68
1317	72	73
1318	80	80
1319	93	93
1320	92	91
1321	70	69
1322	84	84
1323	72	71
1713	118	117
1714	96[1]	95
1715	96[1]	93
1716	96[1]	86
1717	96[1]	96
1718	139	140
1719	96[1]	97
1720	96[1]	97
1721	89	86
1722	79	79
1723	84	85
1906	98	98
1907	67	67
1908	53	53
2010	118	118
2011	47	47
3107	80[2]	79
4110	60	59

[1] Number secondary units observed was less than or equal to 2. Secondary unit average for land use stratum 17 is shown.

[2] Number secondary units observed was less than or equal to 2. Secondary unit average for all strata is shown.

Table 4

Analysis of Variance for the 1983 Objective Yield Survey Data

Source	Degrees of Freedom	Sum of Squares	Mean Square	Variance Component	Percent of total
Stratum	26	80,193	3,084.3	187.3	14
Segment within Stratum	85	124,086	1,459.8	378.0	28
Residual	103	79,991	776.6	776.6	58
Total	214	284,270		1,341.9	100

When a segment had cotton, the mean number of plants per 100 square feet for segment hk was simulated by

$$\bar{c}_{hk} = \bar{c}_{h.} + e_{hk}, \tag{4.2}$$

where $\bar{c}_{h.}$ is the average number of plants per 100 square feet for stratum h, e_{hk} is distributed $N(0, s_b^2)$, and $s_b^2 = 378.0$. In the event that the simulated segment mean $(\bar{c}_{hk})$ was less than 10% of the stratum mean, then c_{hk} was set equal to $(.10)\bar{c}_{h.}$. Table 3 compares the simulated stratum means with those from the 1983 objective yield survey. The overall mean in the simulated population was $\bar{\bar{y}}_N = 79.6$.

From the simulated population 500 June Enumerative Survey samples were drawn using stratified random sampling. A total of 275 segments were drawn for each of the simulated samples. The number of segments drawn from each stratum was the same as that for the 1983 June Enumerative Survey (see Table 2). For each of the simulated samples, estimates of the mean number of acres per segment in the population, as well as the conditional probabilities π_{hk}, from (3.12), that the segments in the sample would receive plots in a draw, were calculated. These conditional probabilities were used at the second stage of sampling in the single start probability-proportional-to-estimated-size systematic sampling described in Section 2. Objective yield survey samples were simulated by selecting 220 secondary units using this systematic sampling scheme. Two objective yield survey samples were simulated for each of the 500 simulated June Enumerative Survey samples.

When a segment was selected to receive a secondary unit, the yield (number of plants per 100 square feet) observed within a field was simulated under the assumption that the coefficient of variation within each segment was constant. The observed number of plants was defined as

$$y_{hk\ell} = \bar{c}_{hk} + s_w \bar{\bar{y}}_N^{-1} \, \bar{c}_{hk} f_{hk\ell}, \tag{4.3}$$

where $y_{hk\ell}$ is the estimated average number of plants per 100 square feet for the ℓ-th secondary unit in segment k of stratum h, and $f_{hk\ell}$ is distributed $N(0, 1)$. The within-segment standard error is the square root of the $s_w^2 = 776.6$ of Table 4, and $\bar{\bar{y}}_N$ is the overall mean number of plants per plot. In the event that $y_{hk\ell}$ was less than 10% of the stratum mean, then $y_{hk\ell}$ was set equal to $(.10)\bar{c}_{hk}$. Similarly, if $y_{hk\ell}$ was greater than 190% of the stratum mean, then $y_{hk\ell}$ was set equal to $(1.9)\bar{c}_{hk}$.

Results of the simulations for cotton acreages are summarized in Table 5. The estimated mean acres per segment is

$$\bar{A}_n = \sum_{h=1}^{L} W_h n_h^{-1} \sum_{k=1}^{n_h} A_{hk}, \tag{4.4}$$

with estimated variance

$$\hat{V}(\bar{A}_n) = \sum_{h=1}^{L} W_h{}^2 n_h^{-1} (n_h - 1)^{-1} \sum_{k=1}^{n_h} (A_{hk} - \bar{A}_h)^2. \tag{4.5}$$

Table 5

Estimated Cotton Acreages from 500 Simulated June Enumerative Survey Samples

	$\bar{A}_n$	$\hat{V}(\bar{A}_n)$
Average	9.93	0.64
Range	8.13 - 12.21	
Variance	0.66	0.016

The average cotton acres per segment in the simulated population was 9.94, while the average of the 500 sample estimates was 9.93. The actual variance of the stratified estimator $\bar{A}_n$ was 0.63, while the average estimated variance for the 500 simulated samples was 0.64. Because the variation in estimated cotton acreage is small, π^*_{hk} provides a stable estimate of the unconditional probability that segment k in stratum h is selected to receive at least one secondary unit.

In addition to the estimators discussed previously, random group estimators of the variance were constructed. Two sets of random groups were formed for each objective yield survey sample. One set contained five groups ($\gamma = 5$) and one set contained ten groups ($\gamma = 10$). Random groups were created by dividing the primary sampling units, the segments, into subsets within each land use substratum. The first group in each set of groups was obtained by drawing a simple random sample without replacement of size $K_{h(\gamma)} = n_h/\gamma$ from the sample of segments selected from each stratum ($h=1, \ldots, 28$) of the parent June Enumerative Survey sample. The second random group was obtained in the same fashion by selecting $K_{h(\gamma)}$ segments from the remaining $n_h - K_{h(\gamma)}$ segments in each stratum. The remaining random groups were formed in a like manner. One land use substratum, stratum number 3107, had a sample size of $n_h = 5$ segments. Acreage and yield values of the observed five segments were repeated to form the ten observations required to create ten groups when $\gamma = 10$.

Let D_α be the number of secondary units with positive acres which were selected during the objective yield survey in random group α where $\alpha = 1, \ldots, \gamma$. Let $\bar{y}_{(\alpha)}$ denote the yield estimator obtained from the α-th random group:

$$\bar{y}_{(\alpha)} = D_\alpha^{-1} \sum_{t=1}^{D_\alpha} Y_{t(\alpha)}, \tag{4.6}$$

where $\bar{y}_{(\alpha)}$ is the analogue of equation (3.3) for the α-th group. The random group estimator of the variance of $\bar{y}$ is then given by

$$\hat{V}_{g\gamma}(\bar{y}) = \gamma(\gamma - 1)^{-1} \sum_{\alpha=1}^{\gamma} (\bar{y}_{(\alpha)} - \bar{y})^2. \tag{4.7}$$

This estimator is slightly biased for the ten group estimator because one stratum contained only five observations, and these observations were repeated in the groups.

Similarly, let $\hat{Y}_{(\alpha)}$ denote the total production estimator obtained from the α-th random group:

$$\hat{Y}_{(\alpha)} = N \bar{M}_{n(\alpha)} \bar{y}_{(\alpha)}, \tag{4.8}$$

where

$$\bar{M}_{n(\alpha)} = \sum_{h=1}^{L} W_h K_{h(\alpha)}^{-1} \sum_{k=1}^{K_{h(\alpha)}} M_{hk(\alpha)},$$

$M_{hk(\alpha)}$ is the number of acres of cotton in segment k of stratum h for random group α and $K_{h(\alpha)}$ is the number of segments in stratum h for the α-th group. The random group estimator of the variance of $\hat{Y}$ is then given by

$$\hat{V}_{g\gamma}(\hat{Y}) = \gamma(\gamma - 1)^{-1} \sum_{\alpha=1}^{\gamma} (\hat{Y}_{(\alpha)} - \hat{Y})^2. \tag{4.9}$$

Tables 6 and 7 summarize the results of the Monte Carlo study for yield and production estimators. Average values of the estimates and their variance estimates across the 1,000 simulated objective yield survey samples are shown in the tables. Simulation of two objective yield survey samples for each June Enumerative Survey sample made the estimation of between - and within - June Enumerative Survey variance components possible.

The estimator (3.1) currently used, $\bar{y}$, and the combined ratio estimator (3.15), $\bar{y}_r$, which is based on the π_{hk}^* calculated from June Enumerative survey results, provide estimates with similar accuracy (see Table 6). The equal efficiency is partly due to the accuracy with which the unconditional selection probabilities are estimated in each sample.

As was shown in Section 3.2, the conditional variance $\hat{V}_2(\bar{y})$ is an underestimate of $V(\bar{y})$. For this simulated population, $\hat{V}_2(\bar{y})$ underestimated the observed variance of $\bar{y}$ by 38%. The observed variance of $\bar{y}$ was 11.57 as compared to an average of 7.21 for $\hat{V}_2(\bar{y})$. This underestimation of the variance was consistent across samples. The estimated variance of $\hat{V}_2(\bar{y})$ was 0.99, with $\hat{V}_2(\bar{y})$ ranging from a low of 3.85 to a high of 11.24 in the 1,000 observations. Thus, the maximum observed estimate of the conditional variance was less than the true variance.

Table 6

Monte Carlo Properties of Yield per Acre Estimates and Estimated Variances[1]

	Estimator					
	$\bar{y}$	$\hat{V}_2(\bar{y})$	$\hat{V}_{g5}(\bar{y})$	$\hat{V}_{g10}(\bar{y})$	$\bar{y}_r$	$\hat{V}(\bar{y}_r)$
Average	79.74	7.21	12.62	12.39	79.76	12.39
Total Variance	11.57	0.99	74.58	36.86	11.56	12.51
Between JES	7.60	0.48	6.10	4.56	7.64	7.61
Within JES	3.97	0.51	68.48	32.30	3.92	4.90

[1] Two objective yield survey samples were simulated from each of 500 simulated June Enumerative Survey samples.

Table 7

Monte Carlo Properties of Production Estimates and Estimated Variances[1]

	Estimator[2]					
	$\hat{Y}$	$\hat{V}_*(\hat{Y})$	$\hat{V}_{g5}(\hat{Y})$	$\hat{V}_{g10}(\hat{Y})$	$\hat{Y}_r$	$\hat{V}(\hat{Y}_r)$
Average	73.04	40.85	48.99	48.53	73.07	48.73
Total Variance	49.69	82.52	1245.10	608.80	49.58	222.96
Between JES	46.35	78.17	50.82	208.48	46.30	199.58
Within JES	3.34	4.35	1194.28	400.32	3.28	23.38

[1] Two objective yield survey samples were simulated from each of 500 simulated June Enumerative Survey samples. There were $N = 92{,}240$ segments in the simulated population.

[2] The estimator $\hat{Y}$ is in millions of units and variances are in the corresponding units.

Assuming probability-proportional-to-size sampling with replacement of segments at the second phase, $\hat{V}_2(\bar{y})$ was shown in Section 3.2 to be unbiased for the variance of $\bar{y}$ conditional on the sample of segments selected at the first stage of sampling. The estimate of the expected value of the conditional variance of $\bar{y}$, $V_2(\bar{y})$, from the Monte Carlo study is 3.97. This large discrepancy (3.97 versus 7.21) can be attributed to the fact that the estimator $\hat{V}_2(\bar{y})$ ignores the effects of stratification in the population (see Tables 2 and 3) and to the fact that $\hat{V}_2(\bar{y})$ was derived under the assumption that segments are selected with replacement at the second stage of sampling.

The estimator (3.9), $\hat{V}_*(\hat{Y})$, underestimates the unconditional variance of $\hat{Y}$. While the observed variance of $\hat{Y}$ from the Monte Carlo simulations is 49.69 (million)^2, the average of the $\hat{V}_*(\hat{Y})$ is only 40.85 (million)^2. This 18% underestimate of the true variance occurs for a number of reasons. As was shown previously, there is a negative bias in $\hat{V}_2(\bar{y})$ as an estimator of $\hat{V}(\bar{y})$; another important factor contributing to the bias is the failure of $\hat{V}_*(\hat{Y})$ to take into account the covariance between $\bar{M}_n$ and $\bar{y}$. In this example, the bias caused by omitting the covariance term partially balances the bias associated with $\hat{V}(\bar{y})$.

Using expression (3.16), $\hat{V}(\bar{y}_r)$, as an estimator of the variance of $\bar{y}_r$ and expression (3.21), $\hat{V}(\hat{Y}_r)$, as an estimator of the variance of $\hat{Y}_r$ provided results which are much more satisfactory than those of the estimators currently used. The Monte Carlo average of the estimates $\hat{V}(\bar{y}_r)$ was 12.51, which overestimates the observed variance of $\bar{y}_r$ (11.57) by about 7%. About one-third of the overestimate (2-4%) can be attributed to the use of sampling without replacement at the first two stages of sampling. The remaining difference of about 4% is small relative to the standard error of the estimated difference. The variance of the difference was estimated by estimating the variance of the mean of z_{tj}, where

$$z_{tj} = (\bar{y}_{n,r(tj)} - 79.76)^2 - \hat{V}(\bar{y}_{n,r(tj)}) , \tag{4.10}$$

for the j-th yield sample ($j = 1, 2$) within June Enumerative Survey sample t ($t = 1, \ldots, 500$). The estimated standard error of the difference was 0.58. Thus, the average value of $\hat{V}(\bar{y}_r)$ is within 1.5 standard errors of the estimated variance of $\bar{y}_r$. The average estimated variance of $\hat{Y}_r$ is within 2 percent of the variance observed in the Monte Carlo simulations.

Random group estimators of the variance of $\bar{y}$ displayed little bias. The Monte Carlo averages of estimators $\hat{V}_{g5}(\bar{y})$ and $\hat{V}_{g10}(\bar{y})$ were 9% and 7%, respectively, larger than the corresponding Monte Carlo variances. These differences are not significantly different from zero and are comparable to those obtained for the estimator $\hat{V}(\bar{y}_r)$. The variance estimator $\hat{V}(\bar{y}_r)$, however, is a much more stable variance estimator. The coefficient of variation for the estimator $\hat{V}(\bar{y}_r)$ is about 30%; it is 75% for $\hat{V}_{g5}(\bar{y})$. As expected (Wolter 1985), an increase in the number of random groups resulted in a decrease in the coefficient of variation of the random group variance estimator. The coefficient of variation for $\hat{V}_{g10}(\bar{y})$ was 50%. Differences among random groupings and yield samples within June Enumerative Surveys accounted for most of the variance in the random groups variance estimators.

4. CONCLUSIONS

Analyses show that the estimators of statewide average yield and total production currently used by the National Agricultural Statistics Service are satisfactory. However, the simple variance estimators $\hat{V}_2(\bar{y})$ and $\hat{V}_*(\hat{Y})$ were shown to have a negative bias, where the extent of the underestimation is a function of the within-segment variance and of the within-segment sampling rates. The estimator $\hat{V}_2(\bar{y})$ underestimated the true variance of $\bar{y}$ by nearly 40%, and $\hat{V}_*(\hat{Y})$ underestimated the true variance of $\hat{Y}$ by 18% for the simulated California cotton population.

The alternative estimators, $\bar{y}_r$ and $\hat{Y}_r$, were developed by viewing the yield sampling scheme as a two-phase process in which segments found to contain crop acreage during phase one (the June Enumerative Survey) are subsampled during phase two to estimate yield. The unconditional probability of selecting a segment to receive a secondary unit within a stratum, π^*_{hk}, is estimated by assuming that this probability is proportional to the conditional probability of selecting segments at the second phase of sampling. With this assumption, the unequal probability combined ratio estimator of the mean yield, $\bar{y}_r$, and the estimator of its variance, $\hat{V}(\bar{y}_r)$, were developed. The estimator of the total $\hat{Y}_r$ is a two-phase product estimator of the mean production per segment, where the estimator of the mean of the auxiliary variable (crop acreage) comes from the June Enumerative Survey (phase one of sampling). The variance estimator $\hat{V}(\hat{Y}_r)$ is a stratified double sampling (two-phase) estimator of the variance of $\hat{Y}_r$.

As shown by the Monte Carlo study, $\bar{y}_r$ and $\hat{Y}_r$ give estimates that are comparable to their currently used counterparts, $\bar{y}$ and $\hat{Y}$. Both $\hat{V}(\bar{y}_r)$ and $\hat{V}(\hat{Y}_r)$ are accurate variance estimators in samples of the size typically used by NASS. These results are due, in part, to the precision with which average crop acreages are estimated by the June Enumerative Survey. Precise acreage estimates produce estimates of selection probabilities that are close to the unconditional probabilities of selection. In addition, the ratio form of the estimator reduces the effect of replacing true unconditional probabilities with estimators.

Random group variance estimators are also essentially unbiased estimators of the variance of estimated yield and production. However, random group estimators are much less stable than $\hat{V}(\bar{y}_r)$ and $\hat{V}(\hat{Y}_r)$. Therefore, estimators $\hat{V}(\bar{y}_r)$ and $\hat{V}(\hat{Y}_r)$ are recommended over random group estimators.

The June Enumerative Survey forms phase one of the objective yield survey. Sampling procedures for the June Enumerative Survey are straightforward and, as was shown by the Monte Carlo study, provide accurate acreage estimates. Hence, no change in the overall design for phase one of the objective yield survey is recommended.

A number of modifications for phase two of the objective yield surveys should be investigated. The current procedure for estimating yield is a two phase procedure in which a combined ratio estimator is used. In states where the sample is relatively large, independent sampling at phase two within individual strata or for groups of strata, as well as the use of a separate ratio estimator should be considered.

Systematic sampling at phase two should be replaced if unbiased estimators of the variance are desired. Segments for yield sampling at phase two are now selected by computer at a national level so it should be relatively easy to change to a selection procedure with known joint selection probabilities. Estimators similar to those recommended for the current design would still be suitable if the same selection probabilities were retained. The scheme described by Fuller (1970) is one procedure that can be computerized, for which joint selection probabilities can be calculated, and which maintains specified selection probabilities and a degree of control similar to that of systematic sampling.

ACKNOWLEDGEMENTS

This research was supported in part by cooperative research agreement 58-319T-1-0054X with the National Agricultural Statistics Service, U.S. Department of Agriculture. We thank the referees for useful comments.

REFERENCES

FECSO, R. (1978). Cluster analysis as an aid in creating paper strata. Statistical Reporting Service, U.S. Department of Agriculture.

FECSO, R., and JOHNSON, V. (1981). The new California area frame: A statistical study. Statistical Reporting Service, U.S. Department of Agriculture.

FULLER, W.A. (1970). Sampling with random stratum boundaries. *Journal of the Royal Statistical Society*, Ser. B, 32, 209-226.

HOUSEMAN, E.E. (1975). Area frame sampling in agriculture. Statistical Reporting Service, U.S. Department of Agriculture.

PRATT, W.L. (1984). The use of interpenetrating sampling in area frames. Statistical Reporting Service, U.S. Department of Agriculture.

WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

Survey Methodology, June 1987
Vol. 13, No. 1, pp. 63-74
Statistics Canada

Current Issues on Seasonal Adjustment

ESTELA BEE DAGUM[1]

ABSTRACT

This paper discusses three problems that have been a major preoccupation among researchers and practitioners of seasonal adjustment in statistical bureaus for the last ten years. These problems are: (1) the use of concurrent seasonal factors versus seasonal factor forecasts for current seasonal adjustment; (2) finding an optimal pattern of revisions for series seasonally adjusted with concurrent factors; and (3) smoothing highly irregular seasonally adjusted data.

KEY WORDS: Concurrent vs forward seasonal factors; Revisions; Trend-cycle filters; Smoothing.

1. INTRODUCTION

During the last decade, within the domain of seasonal adjustment, statistical bureaus have focused their attention on three important issues: (1) the seasonal adjustment of a current value; (2) the revisions of concurrent seasonally adjusted data; and (3) the smoothing of highly irregular seasonally adjusted series.

The main purpose of this article is to discuss each of the above problems with respect to the X-11-ARIMA seasonal adjustment program developed by Dagum (1980) and which is applied by Statistics Canada and other statistical bureaus of the world.

The four modes in which the X-11-ARIMA computer package can be used to produce a current seasonally adjusted value are discussed in Section 2. In Section 3, the focus is on analysis of the revisions of concurrent seasonally adjusted data based on the linear filters of X-11-ARIMA. Section 4 deals with the nature and characteristics of the smoothing (trend-cycle) filters available in X-11-ARIMA.

2. SEASONAL ADJUSTMENT OF CURRENT VALUES

The seasonal adjustment of a current value can be done using either a "concurrent" seasonal estimate or a seasonal "forecast".

A "concurrent" seasonal estimate (factor or effect depending on whether a multiplicative or additive model is assumed) is obtained by seasonally adjusting, each time a new observation is available, all the data available up to and including that observation. On the other hand, a seasonal "forecast" is obtained from a series that ended in the previous year. A common practice is to generate these seasonal forecasts, say for year $t + 1$, from data that ended in December of the previous year t.

There are four modes in which the X-11-ARIMA computer program can be applied to produce a current (last observation) seasonally adjusted value. These four modes are: (i) using ARIMA extrapolations and concurrent seasonal factors; (ii) using ARIMA extrapolations and seasonal factor forecasts; (iii) using concurrent seasonal factors without the use of ARIMA extrapolations; and (iv) using seasonal factor forecasts without the use of ARIMA extrapolations.

[1] Estela Bee Dagum, Time Series Research and Analysis Division, Methodology Branch, Statistics Canada, 13th Floor, R.H. Coats Building, Ottawa, Ontario, Canada K1A 0T6.

While statistical bureaus use the four modes to obtain current seasonally adjusted values, not all of them do so with the same frequency. Thus, for example, the dominant mode in Statistics Canada is (i) followed by mode (iii) whereas in the U.S. Bureau of Labor, the dominant mode is (ii) followed by mode (iv). The current seasonally adjusted value produced by each type of seasonal adjustment varies and is subject to different degrees of error.

Under the assumption of an additive decomposition model, the seasonal adjustment of a current value X_t can be obtained by

$$\hat{X}_t^{(\ell)} = X_t - \hat{S}_t^{(\ell)}, \tag{1}$$

where $\hat{S}_t^{(\ell)}$ denotes a forward seasonal estimate; or by

$$\hat{X}_t^{(0)} = X_t - \hat{S}_t^{(0)}, \tag{2}$$

where $\hat{S}_t^{(0)}$ denotes a concurrent seasonal estimate.

The current seasonally adjusted value will become "final" in the sense that it will no longer be revised after m more observations are added. Thus,

$$\hat{X}_t^{(m)} = X_t - \hat{S}_t^{(m)}, \tag{3}$$

where $\hat{S}_t^{(m)}$ denotes a final seasonal estimate.

Therefore, the total revision of a concurrent and of a forward seasonal estimate can be written as

$$r_t^{(0,m)} = \hat{S}_t^{(0)} - \hat{S}_t^{(m)}, \ m > 0; \tag{4}$$

$$r_t^{(\ell,m)} = \hat{S}_t^{(\ell)} - \hat{S}_t^{(m)}, \ m > 0 > \ell. \tag{5}$$

Under the assumption of an additive decomposition and no replacement of extreme values, $\hat{S}_t^{(m)}$, the final seasonal estimate from a series $X_{t-m}, \ldots, X_t, \ldots, X_{t+m}$ can be expressed by

$$\hat{S}_t^{(m)} = \sum_{j=-m}^{m} h_{m,j} X_{t-j} = h^{(m)}(B) X_t, \tag{6}$$

where $h_{m,j} = h_{m,-j}$ are the symmetric moving average weights to be applied to the series. $h^{(m)}(B)$ denotes the corresponding linear filter using the backshift operator B, such that $B^n = X_{t-n}$. Young (1968) showed that the length of this symmetric filter $h^{(m)}(B)$, for monthly series, is 145 but that it can be well approximated by 85 weights because the values of the weights attached to distant observations are very small and, thus, $m = 42$.

Following equation (6) we can express a concurrent seasonal estimate $\hat{S}_t^{(0)}$ and a forward seasonal estimate $\hat{S}_t^{(\ell)}$ by:

$$\hat{S}_t^{(0)} = \sum_{j=-2m}^{0} h_{0,j} X_{t-j} = h^{(0)}(B) X_t, \ m = 42, \tag{7}$$

where $h^{(0)}(B)$ denotes the asymmetric *concurrent* seasonal filter; and

$$\hat{S}_t^{(\ell)} = \sum_{j=-2m}^{\ell} h_{\ell,j} X_{t-j} = h^{(\ell)}(B) X_t,\ m = 42, \tag{8}$$

where $h^{(\ell)}(B)$ denotes the asymmetric *forecasting* seasonal filter and $\ell = 1,2, ..., 12$ for a monthly series.

The revision of a concurrent seasonal estimate depends on the distance between the concurrent and the final filter, that is, $d[h^{(0)}(B), h^{(m)}(B)]$, and on the innovations of the new observations $X_{t+1}, X_{t+2}, ..., X_{t+m}$.

Similarly, the revision of a forward seasonal estimate depends on $d[h^{(\ell)}(B), h^{(m)}(B)]$ and on the new innovations introduced by $X_{t-\ell}, ..., X_t, X_{t+1}, ..., X_{t+m}$.

Theoretical studies by Dagum (1982a and 1982b) have shown that

$$d[h^{(0)}(B), h^{(m)}(B)] < d[h^{(\ell)}(B), h^{(m)}(B)] \text{ for } \ell = 1,2, ... 12. \tag{9}$$

The distance between the two filters is defined as the mean squared difference between the frequency response function of the filters over all the seasonal frequencies; a similar definition is given in the next section (equation (17)) using the root mean squared difference.

Relation (9) is true whether ARIMA extrapolations are used or not. Furthermore, the two studies also showed that

$$\begin{aligned} & d[h^{(0)}(B), h^{(m)}(B)] \quad \text{using ARIMA extrapolations} \\ < & d[h^{(0)}(B), h^{(m)}(B)] \quad \text{without ARIMA extrapolations,} \end{aligned} \tag{10}$$

and similarly

$$\begin{aligned} & d[h^{(\ell)}(B), h^{(m)}(B)] \quad \text{using ARIMA extrapolations} \\ < & d[h^{(\ell)}(B), h^{(m)}(B)] \quad \text{without ARIMA extrapolations,} \end{aligned} \tag{11}$$

$$\text{for } \ell = 1,2, ..., 12.$$

Studies by Dagum (1978), Bayer and Wilcox (1981), Kenney and Durbin (1982), McKenzie (1984), Dagum and Morry (1984), Pierce (1980) and Pierce and McKenzie (1985) have shown that

$$r^{(0,m)} < r^{(\ell,m)} \tag{12}$$

except in a few cases where

$$r^{(0,m)} > r^{(\ell,m)}. \tag{13}$$

The relationship (13) can be observed when the current observations of the latest year are strongly revised since X_t gets the largest weight in the estimations of $\hat{S}_t^{(0)}$.

From the viewpoint of the total revisions of the seasonal estimates, the results of the above empirical studies permit the ranking of the four modes as follows: mode (i) (ARIMA extrapolations with concurrent seasonal estimates) gives the smallest total revision; mode (iii) (no ARIMA extrapolations with concurrent seasonal estimates) ranks second; mode (ii) (ARIMA extrapolations with forward seasonal estimates) ranks third and mode (iv) (ARIMA extrapolations with forward seasonal estimates) ranks fourth.

3. REVISIONS OF CONCURRENT SEASONALLY ADJUSTED DATA

Statistics Canada's practice of using concurrent seasonal adjustment was first established in 1975 for the Labour Force Survey series. Gradually other foreign statistical agencies followed it. The use of concurrent seasonal factors for current seasonal adjustment poses the problem of how often should the series be revised. Kenny and Durbin (1982) recommended that revisions should be made after one month and thereafter each calendar year. Dagum (1982c) supported these conclusions and furthermore, recommended an additional revision at six months if the seasonal adjustment method is the X-11-ARIMA without the ARIMA extrapolation option.

For any two points in time $t + k$, $t + \ell$ $(k < \ell)$, the revisions of the seasonal estimates and consequently of the seasonally adjusted value is given by

$$r_t^{(\ell,k)} = \hat{X}_t^{(\ell)} - \hat{X}_t^{(k)}, \; k < \ell. \tag{14}$$

This revision reflects: (1) the innovations introduced by the new observations X_{t+k+1}, $X_{t+k+2}, \ldots, X_{t+k+\ell}$; and (2) the differences between the two asymmetric seasonal adjustment filters $Y^{(\ell)}(B)$ and $Y^{(k)}(B)$. If one fixes $k = 0$ and lets ℓ vary from 1 to m, then relation (14) gives a sequence of revisions of the concurrent seasonally adjusted values for different time spans or lags. The *total revision* of the concurrent estimate is given for $\ell = m$. If one fixes $\ell = k + 1$ and lets k take values from 0 to $m - 1$, then relation (14) gives the sequence of *single period revisions* of each estimated seasonally adjusted value and in particular, if one starts at $k = 0$ one obtains the $m - 1$ successive single period revisions of each estimated seasonally adjusted value before it becomes final. If one fixes $\ell = k + 12$ and lets k take values from 0 to $m - 12$, then equation (14) gives the sequence of annual revisions.

The revisions in which we are interested here are those introduced by filter discrepancies, and these can be studied by looking at the frequency response functions of the corresponding filters. Similarly to equation (6), we can approximate the seasonally adjusted value for recent years from the X-11-ARIMA program (with or without ARIMA extrapolations) by

$$\hat{X}_t^{(n)} = \sum_{j=n}^{m} Y_{n,j} X_{t-j} = Y^{(n)}(B)\, X_t. \tag{15}$$

Equation (15) represents a linear system where $\hat{X}_t^{(n)}(n)$ is the convolution of the input X_t and a sequence of weights $Y_{n,j}$ called the *impulse response function* of the filter. The properties of this function can be studied using its Fourier transform which is called the *frequency response function*, defined by

$$\Gamma^{(n)}(\omega) = \sum_{j=-n}^{m} Y_{n,j} e^{-2\pi\omega j}, \quad -\tfrac{1}{2} \leq \omega \leq \tfrac{1}{2}, \tag{16}$$

where ω is the frequency in cycles per unit time. $\Gamma^{(n)}(\omega)$ fully describes the effects of the linear filter on the given input. Monthly and annual revisions of the concurrent filter of X-11-ARIMA with and without the ARIMA extrapolations have been calculated by Dagum (1987) based on the mathematical distance between the various frequency response functions of the filters. The pattern is characterized by a rapid decrease in the size of the monthly revisions of the concurrent filter for $\ell = 1, 2$, and 3; and a slow decrease thereafter until $\ell = 11$; then a large increase occurs at $\ell = 12$ followed by a decrease at $\ell = 13$ and then another large increase at $\ell = 24$ followed by a decrease at $\ell = 25$. Dagum (1987) showed that this pattern of monthly revisions is the same whether ARIMA extrapolations are used or not.

The significant decreases for the first three consecutive revisions are due to the improvement of the Henderson (trend- cycle) filter weights. The reversal of direction in the size of the filter revisions at $\ell = 12$ and $\ell = 13$, is due to the improvements of the seasonal filter that becomes less asymmetrical from year to year until three full years are added to the series. The two largest revisions occur at $\ell = 1$ and $\ell = 12$. *Given the non-monotonicity of single monthly revisions, it is not advisable to revise the concurrent estimate any time a new observation is added to the series.*

A revision scheme often used by statistical bureaus for their concurrent seasonally adjusted series consists of keeping constant the concurrent estimate from the time it appears until the end of the year and then revising annually the current and earliest years. Therefore, first year revisions due to filter discrepancies are given by $R^{(0,0)}$, $R^{(1,0)}$, ..., $R^{(11,0)}$; second year revisions by $R^{(12,0)}$, $R^{(13,1)}$, ..., $R^{(23,11)}$ third-year revisions by $R^{(24,12)}$ $R^{(25,13)}$ and so on where $R^{(\ell,k)}$ is defined by

$$R^{(\ell,k)} = [2\int_0^{\frac{1}{2}} \| \Gamma^{(\ell)}(\omega) - \Gamma^{(k)}(\omega) \|^2 \, d\omega]^{\frac{1}{2}}, \tag{17}$$

$$\ell = 1, 2, \ldots, n, k = 0, 1, 2, \ldots, n - 12,$$

and $n = 42$ for the X-11-ARIMA seasonal adjustment filters.

Table 1 shows the first-, second- and third-year revisions of the concurrent seasonal adjustment filter for X-11-ARIMA without extrapolation and with extrapolations from one ARIMA model and two sets of parameter values (other cases are shown in Dagum 1987). The ARIMA model chosen is the classical (0,1,1) $(0,1,1)_{12}$ model that is $(1 - B)\ (1 - B^{12})\ X_t = (1 - \theta B)\ (1 - \Theta B^{12})a_t$ where X_t denotes the original series, B is the backshift operator such that $B^n X_t = X_{t-n}$, a_t is a purely random process that represents the innovations and θ and Θ are the non-seasonal and seasonal parameters, respectively.

Since the largest single period revisions occur at $\ell = 1$ and $\ell = 12$ as mentioned above, a better revision scheme would be to incorporate monthly and annual revisions. It is expected that (1) adjusting concurrently each month, say from January to November and revising only once when the next month is available, and (2) adjusting concurrently December when it first appears and then revising the first year and earlier years when January is added, should improve the reliability of the filter applied during the current year while maintaining simultaneously the filter's homogeneity for month-to-month comparisons.

The first-year revisions of the first-month revised filter would then be $R^{(1,1)}$, $R^{(2,1)}$, ..., $R^{(11,1)}$. Table 2 shows these revisions and although the pattern is very similar to that of the concurrent filter, *the size of the revisions are much smaller if no extrapolations are used.* On the other hand, *the improvement is less important if ARIMA extrapolations are used.* Similarly, no major differences were observed for the second- and third-year revisions.

3.1 Estimation of Trading Day Variations and ARIMA Models with Concurrent Seasonal Adjustment

Besides the type of revisions scheme to be applied, there are two other problems posed by concurrent seasonal adjustment associated with trading day variations and ARIMA modelling.

Table 1
First-, Second- and Third-Year Revisions of the Concurrent Seasonal Adjustment Filter of X-11-ARIMA

Revisions $R^{(\ell,k)}$	Without ARIMA Extrapolations	With ARIMA Extrapolations from a (0,1,1) $(0,1,1)_{12}$ Model	
		$\theta = .40$ $\Theta = .80$	$\theta = .80$ $\Theta = .80$
$R^{(1,0)}$	.12	.12	.06
$R^{(2,0)}$	.13	.13	.08
$R^{(3,0)}$	.13	.13	.08
$R^{(4,0)}$	.13	.13	.09
$R^{(5,0)}$	.15	.13	.09
$R^{(6,0)}$	.17	.13	.09
$R^{(7,0)}$	.16	.13	.09
$R^{(8,0)}$	.16	.13	.09
$R^{(9,0)}$	.16	.13	.09
$R^{(10,0)}$	.16	.14	.09
$R^{(11,0)}$	.16	.14	.09
$R^{(12,0)}$	.29	.28	.26
$R^{(13,1)}$	.27	.27	.26
$R^{(14,2)}$	.27	.27	.26
.	.	.	.
.	.	.	.
.	.	.	.
$R^{(23,11)}$	.27	.26	.26
$R^{(24,12)}$	.20	.16	.16
$R^{(24,13)}$	.18	.17	.16
$R^{(36,24)}$	.16	.17	.16
.	.	.	.
.	.	.	.
.	.	.	.

Table 2
First-Year Revisions of the First-Month Revised Seasonal Adjustment Filter

Revisions $R^{(\ell,1)}$	Without ARIMA Extrapolations	With ARIMA Extrapolations from a (0,1,1) $(0,1,1)_{12}$ Model	
		$\theta = .40$ $\Theta = .80$	$\theta = .80$ $\Theta = .80$
$R^{(2,1)}$	.07	.10	.06
$R^{(3,1)}$	'07	.10	.06
$R^{(4,1)}$	.07	.10	.07
$R^{(5,1)}$	.08	.10	.08
$R^{(6,1)}$	.10	.11	.08
$R^{(7,1)}$	.11	.11	.08
$R^{(8,1)}$	.11	.11	.08
$R^{(9,1)}$	.11	.11	.08
$R^{(10,1)}$	.12	.11	.08
$R^{(11,1)}$	.12	.12	.08

For series which are flows in the sense that they result from the accumulation of daily values over the calendar months, there is a systematic effect caused by trading day variations. Trading day variations arise mainly because the activity varies with the days of the week. Other sources are associated with accounting and reporting practices. For example, stores that do their bookkeeping activities on Friday tend to report higher sales in months with five Fridays than in months with four Fridays. The trading day effects are estimated in the X-11-ARIMA program using ordinary least squares on a simple deterministic regression model. Consequently, the weights estimated for each day change any time a new observation is added to the series. Since regression techniques are very sensitive to outliers, these changes can be sometimes unnecessarily large.

When the series are seasonally adjusted concurrently, the trading day estimates change all the time. In order to avoid unnecessary revisions, Statistics Canada's practice is to use the weights calculated by the program at the end of the previous calendar year or the weights provided by the users, as priors for the current year. The weights are then revised on an annual basis.

The effect of trading day variations must be removed from the series before ARIMA modelling, for these type of models cannot adequately handle trading day variations. In other words, if the X-11-ARIMA program is used with ARIMA extrapolations on series with trading day variations, these variations should be estimated *a priori* and if significant, they should be removed from the original series before the ARIMA modelling.

Another problem associated with concurrent seasonal adjustment refers to how often the ARIMA models should be identified. The current practice at Statistics Canada is to use the automatic ARIMA model selection option once a year and if the model is accepted, then it is kept constant for a whole year, letting only the parameters change when more observations are added. In order to keep the model constant, the user's supplied model option should be applied. Maintaining the ARIMA model constant avoids unnecessary revisions that may result from changing of models back and forth simply because of the presence of outliers.

4. SMOOTHING OF VOLATILE SEASONALLY ADJUSTED SERIES

One of the main purposes of the seasonal adjustment of economic time series is to provide information on current economic conditions, particularly to determine the stage of the cycle at which the economy stands. Since seasonal adjustment means removing only seasonal variations, thus leaving trend-cycle variations together with irregular fluctuations, it is often difficult to detect the short-term trend or cyclical turning points for series strongly affected with irregulars. In such cases, it may be preferable to smooth the seasonally adjusted series using trend-cycle estimators which suppress as much as possible the irregulars without affecting the cyclical component.

The use of trend-cycle values has been discussed by several writers and recently by Moore *et al* (1981), Kenny and Durbin (1982), Maravall (1986) and Dagum and Laniel (1987). Although not yet practised widely, some statistical agencies such as Statistics Canada and the Australian Bureau of Statistics smooth some of their seasonally adjusted series, particularly those series that are strongly affected by irregulars.

The combined linear filters applied to the original series to generate a central (symmetric) estimate of the trend-cycle component have been calculated by Young (1968) for Census Method II-X-11 variant. This filter is similar to that of X-11-ARIMA with and without ARIMA extrapolations. Dagum and Laniel (1987) extended Young's (1968) results to include the estimation of the asymmetric trend-cycle filters of X-11-ARIMA with and without the ARIMA extrapolations.

Figure 1 shows the gain functions of the central (symmetric) seasonal adjustment filters and smoothed seasonally adjusted data (trend-cycle) filters. It is apparent that the trend-cycle filters suppress all the noise present in the series, where the noise is defined as the power present in all frequencies $\omega \leq .166$. This frequency corresponds to the first harmonic of the fundamental seasonal frequency of a monthly series. This pattern results from the convolution of the seasonal adjustment filters with the 13-term Henderson trend-cycle filter.

Figure 2a shows the gain functions of the concurrent and first-month revised trend-cycle filters of X-11-ARIMA *without* ARIMA extrapolations. Figure 2b shows their corresponding phase-shift functions expressed in months instead of radians. We can observe that the gain for all $\omega \leq .166$ is much larger for these two asymmetric filters as compared with the central filter. Furthermore, there are large amplifications for frequencies near the fundamental seasonal. All this means that the concurrent and first revised smoothed seasonally adjusted values will have more noise than the final estimates. On the other hand, it is apparent that the phase shifts are very small, less than one month for the most important cyclical frequencies $0 < \omega < .055$ (i.e., cycles of periodicities equal to and longer than 18 months).

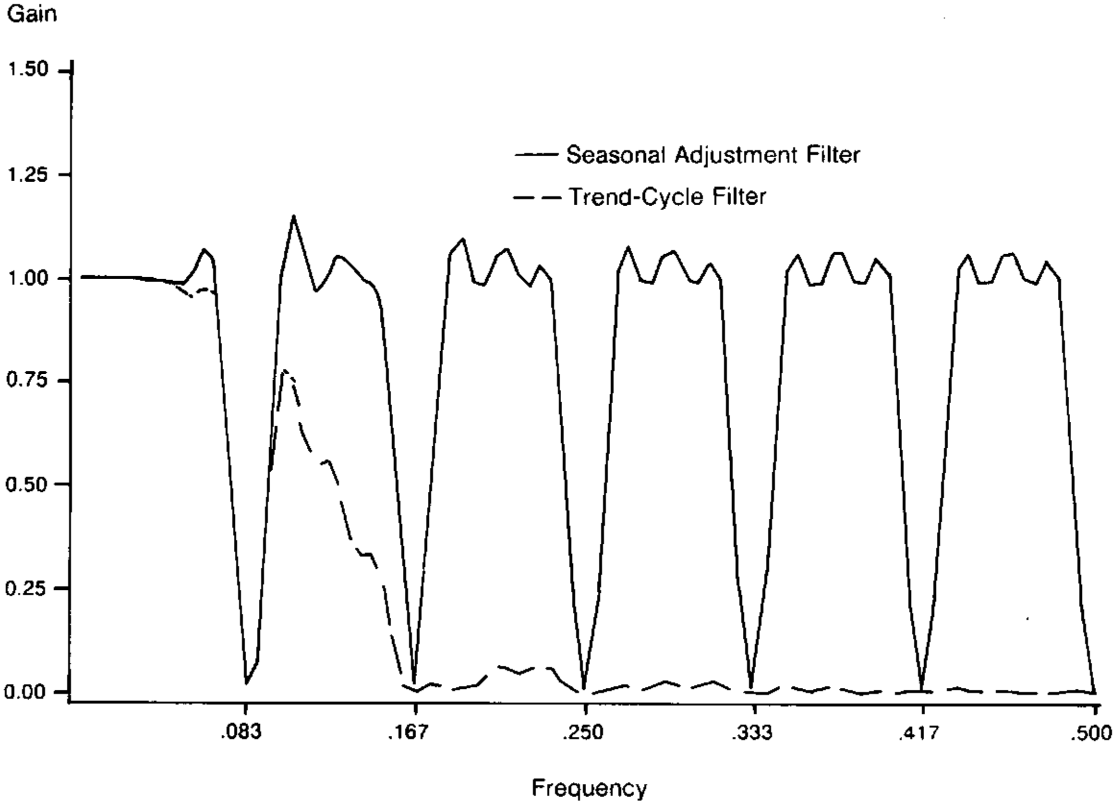

Figure 1. Gain Functions of the Central (Symmetric) Trend-Cycle and Seasonal Adjustment Filters of X-11-ARIMA.

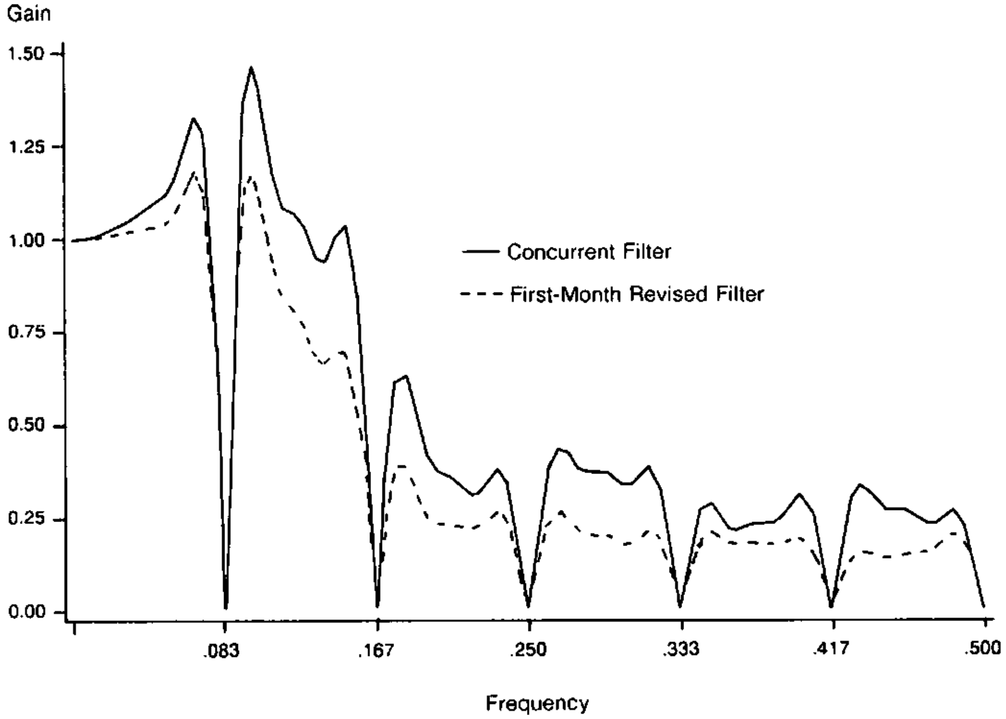

Figure 2a. Gain Functions of the Concurrent and First-Month Revised Filters of X-11-ARIMA without ARIMA Extrapolations.

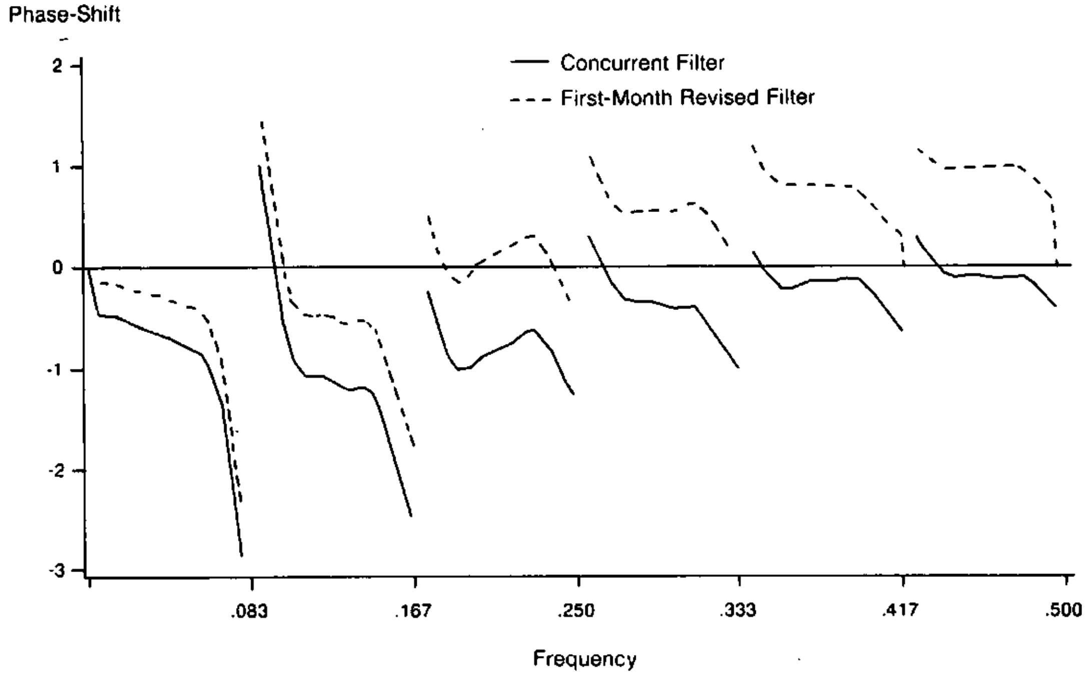

Figure 2b. Phase-Shift Functions of the Concurrent and First-Month Revised Filters of X-11-ARIMA without ARIMA Extrapolations.

Figures 3a and 3b show the gain and phase-shift functions of the concurrent and first-month revised trend-cycle filters of X-11-ARIMA with ARIMA extrapolations. The extrapolations are obtained from an IMA model $(0,1,1)(0,1,1)_{12}$ with $\theta = .40$ and $\Theta = .60$. The gain functions are closer to the symmetric (central) filter than those of X-11-ARIMA without the ARIMA extrapolations. There are no amplifications around the fundamental seasonal frequency and a similar attenuation of power at higher frequencies. On the other hand, there is more phase-shift (being near to one month) for low frequencies and less phase-shift for all high frequencies.

Dagum and Laniel (1987) studied the time path of the revisions of the trend-cycle filters and compared them with those of the seasonal adjustment filters. Their results, as summarized in Table 3, show that the total revisions of the trend-cycle asymmetric filters converge to zero much faster than those of the corresponding seasonal adjustment filters. In fact, the total revision of the trend-cycle filter three months after the concurrent filter is only .1, whereas a close value is achieved for the seasonal adjustment filter only after 24 months have been added to the series. Except for the total revisions of the concurrent filter which is larger for the trend-cycle filters compared with the corresponding seasonal adjustment filter, in all the other cases the total revisions are smaller for the trend-cycle filters. Furthermore, the trend-cycle filter revisions converge much faster to zero as compared with those of the seasonal adjustment filters.

Table 3
Time Path of the Total Revisions of the Trend-Cycle and the Seasonal Adjustment Asymmetric Filters of X-11-ARIMA

Revisions	Without Extrapolations		With Extrapolations from a $(0,1,1)\ (0,1,1)_{12}$ Model $\theta = .40 \quad \Theta = .60$	
$R^{(\ell,k)}$*	Trend-Cycle Filter	Seasonal Adjustment Filter	Trend-Cycle Filter	Seasonal Adjustment Filter
$R^{(48,0)}$	.45	.36	.41	.32
$R^{(48,1)}$	.27	.33	.26	.32
$R^{(48,2)}$	.15	.32	.15	.32
$R^{(48,3)}$	.11	.32	.11	.31
$R(48,4)$	.12	.32	.11	.31
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
$R^{(48,12)}$	.10	.23	.09	.20
$R^{(48,24)}$	.07	.13	.05	.10
$R^{(48,36)}$	.03	.05	.02	.04
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
$R^{(48,47)}$	.01	.01	.01	.01

* $\ell = 48$ for the "final" trend-cycle filter and $\ell = 42$ for the final seasonal adjustment filter. However, the values shown for the revision of the seasonal adjustment filters are also calculated for $\ell = 48$ since after $\ell = 42$ the values are final and, thus, do not change.

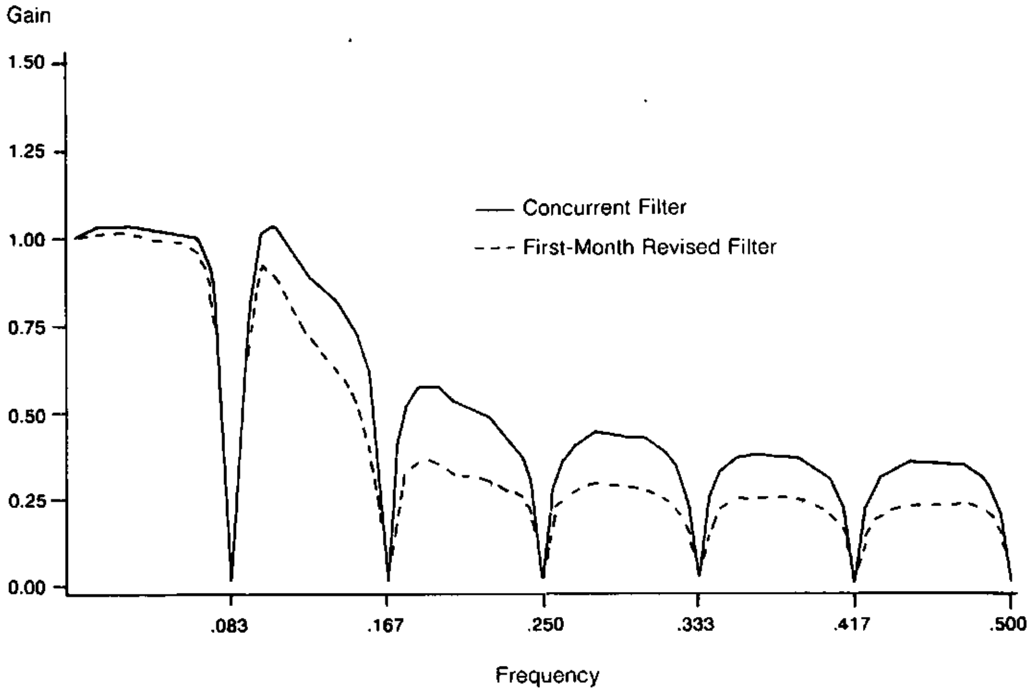

Figure 3a. Gain Functions of the Concurrent and First-Month Revised Filters of X-11-ARIMA with ARIMA Extrapolations (θ = .40, Θ = .60).

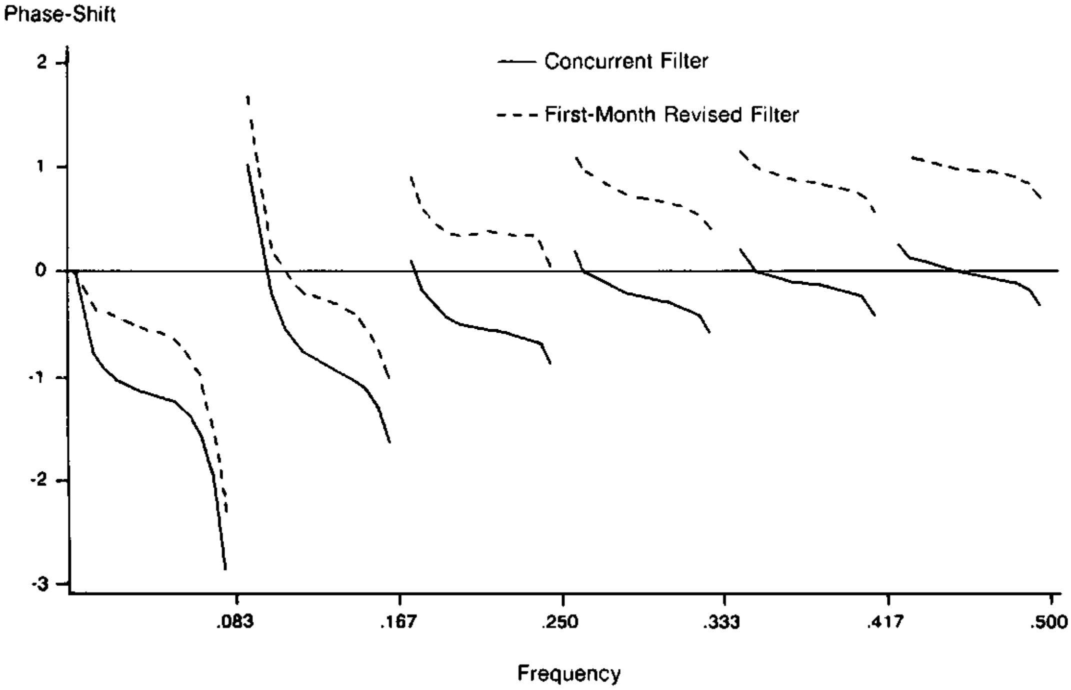

Figure 3b. Phase-Shift Functions of the Concurrent and First-Month Revised Filters of X-11-ARIMA with ARIMA Extrapolations (θ = .40, Θ = .60).

REFERENCES

BAYER, A., and WILCOX, D. (1981). An evaluation of concurrent seasonal adjustment. Technical Report, Board of Governors of the Federal Reserve System.

DAGUM, E.B. (1978). *Comparison and Assessment of Seasonal Adjustment Methods for Labour Force Series*. Stock No. 052-003-00603-1, U.S. Government Printing Office.

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

DAGUM, E.B. (1982a). Revisions of time varying seasonal filters. *Journal of Forecasting*, 1, 173-187.

DAGUM, E.B. (1982b). The effects of asymmetric filters on seasonal factor revisions. *Journal of the American Statistical Association*, 77, 732-738.

DAGUM, E.B. (1982c). Revisions of seasonally adjusted data due to filter changes. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 39-45.

DAGUM, E.B., and MORRY, M. (1984). Basic issues on the seasonal adjustment of the Canadian Consumer Price Index. *Journal of Business and Economic Statistics*, 2, 250-259.

DAGUM, E.B. (1987). Monthly versus annual revisions of concurrent seasonally adjusted series. In *Time Series and Econometric Modelling*, (Eds. I.B. MacNeill and G.J. Umphrey), New York: D. Reidel, 131-196.

DAGUM E.B., and LANIEL, N. (1987). Revisions of trend-cycle estimators of moving average seasonal adjustment method. *Journal of Business and Economic Statistics*, (forthcoming).

KENNY, P., and DURBIN, J. (1982). Local trend estimation and seasonal adjustment of economic time series. *Journal of the Royal Statistical Society,* Ser. A, 145, 1-41.

MARAVALL, A. (1986). An application of model-based estimation of unobserved components. *International Journal of Forecasting*, 2, 305-318.

MOORE, G.H., BOX, G.E.P., KAITZ, H.B., STEPHENSON, J.A., and ZELLNER, A. (1981). Seasonal adjustment of the monetary aggregates. In *Report of the Committee of Experts on Seasonal Adjustment Techniques*, Washington: Board of Governors of the Federal Reserve System.

McKENZIE, S. (1984). Concurrent seasonal adjustment with Census X-11. *Journal of Business and Economic Statistics*, 2, 235-249.

PIERCE, D.A. (1980). Data revisions with moving average seasonal adjustment procedures. *Journal of Econometrics*, 14, 95-114.

PIERCE, D., and McKENZIE, S. (1985). On concurrent seasonal adjustment. Special Studies Paper 164, Federal Reserve Board.

Survey Methodology, June 1987
Vol. 13, No. 1, pp. 75-83
Statistics Canada

On Efficient Estimation of Unemployment Rates from Labour Force Survey Data

S. KUMAR and A.C. SINGH[1]

ABSTRACT

The method of minimum $Q^{(T)}$ estimation for complex survey designs proposed by Singh (1985) provides asymptotically efficient estimates of model parameters analogous to Neyman's (1949) min X^2 estimation procedure for simple random samples. The $Q^{(T)}$ can be viewed as a X^2 type statistic for categorical survey data, and min $Q^{(T)}$ estimates provide a robust alternative to Weighted Least Squares estimates, which often display unstable behaviour for complex surveys. In this paper, the min $Q^{(T)}$ method is first described and then illustrated for the problem of estimating parameters of a logit model for survey estimates of unemployment rates which are obtained from the October 1980 Canadian LFS data cross-classified according to age-education covariate categories. It is seen that the trace efficiency of smoothed estimates obtained by Kumar and Rao (1986), who applied the method of pseudo maximum likelihood estimates (pseudo mle) to the same problem can be slightly improved by the min $Q^{(T)}$ method. Interestingly enough, pseudo mle for individual cells behave much the same way as the efficient min $Q^{(T)}$ estimates for the particular LFS example.

KEY WORDS: Pseudo mle; WLS estimator; Min $Q^{(T)}$ estimator; Asymptotic efficiency; Approximate likelihood; Generalized score statistic.

1. INTRODUCTION

Based on October 1980 Labour Force Survey (LFS) data, Kumar and Rao (1984, 1986) proposed and analysed a logistic regression (logit) model for unemployment rates. They used the theory developed by Roberts (1985) and Roberts, Rao and Kumar (1987) who generalized the Rao-Scott method (1981, 1984) of adjusting X^2 for impact of the underlying survey design to test the fit of the logit model. Kumar and Rao considered unemployment rates in various cells (or domains) that had been obtained by cross-classifying the population into a number of age and education categories. The logit model consisted of both linear and quadratic effects for the age variable, with only the linear effect for the education variable. The same LFS data were also analysed by Singh and Kumar (1986) using an alternative method, namely the $Q^{(T)}$ test proposed by Singh (1985). The test $Q^{(T)}$ is a X^2 type test based on a generalized score statistic of principal components. Results obtained by the $Q^{(T)}$ method were found to be in agreement with those arrived at by the adjusted X^2 method.

Whenever a suitable model is determined, it is of interest to find good estimates of model parameters. These, in turn, provide fairly good estimates of true rates for domains. Such estimates (often called "smoothed estimates") are especially useful for domains in which survey estimates lack precision because the number of observations is not sufficient. It may be noted that since smoothed estimates are obtained after a model is found to have a reasonable fit, the bias in the estimates is expected to be negligible. Kumar and Rao (1986) used the method of pseudo mle (pseudo maximum likelihood estimates) under the working form of the likelihood that corresponds to independent binomial samples for estimating parameters

[1] S. Kumar, Senior Methodologist, Social Survey Methods Division, Jean Talon Building, Tunney's Pasture, Statistics Canada, Ottawa, Ontario, K1A 0T6. A.C. Singh, Associate Professor, Department of Mathematics and Statistics, Memorial University of Newfoundland, St. John's, Newfoundland, Canada, A1C 5S7.

of a logit model after an adequate fit had been established for the October 1980 LFS data. They found a considerable gain in efficiency over survey estimates of unemployment rates in the particular LFS example.

Pseudo mle are known to be useful when the likelihood function is not available or when it is difficult to compute due to complexities of the survey design. Under suitable regularity conditions, the pseudo mle provide consistent and asymptotically normal estimates (Imrey, Koch and Stokes 1982). In this paper we consider the problem of finding asymptotically efficient (in a sense to be explained in Section 3) estimates of model parameters and therefore of domain estimates. We describe the min $Q^{(T)}$ estimator, proposed in Singh (1985), based on the generalized scores approach which can be viewed as analogous to Neyman's min X^2 estimator for simple random samples. It may be noted that the WLS (Weighted Least Squares) approach for complex survey designs (Koch, Freeman and Freeman 1975) also provides asymptotically efficient estimates. However, these estimates are usually unstable for moderate sample sizes due to near singularity of the estimated covariance matrix of survey cell estimates (see Imrey, Koch and Stokes 1982, Fay 1985). The min $Q^{(T)}$ estimates, on the other hand, are designed to guard against the instability problem mentioned above. It will be seen that the problem of instability can be overcome by the min $Q^{(T)}$ method by employing a modified version of the estimated covariance matrix in which the relatively very small eigenvalues from its spectral decomposition are trimmed.

The necessary notation along with a brief review of the test $Q^{(T)}$ are presented in Section 2. Next the min $Q^{(T)}$ estimator and its asymptotic behaviour are described in Section 3. The example using LFS data is given in Section 4 as an illustration. For this numerical example, an interesting finding was that over individual cells, the pseudo mle perform almost at par with efficient min $Q^{(T)}$ estimates. In terms of an overall measure as given by trace efficiency, pseudo mle are found to be only slightly inferior to min $Q^{(T)}$ estimates. Finally, Section 5 contains some concluding remarks.

2. THE TEST $Q^{(T)}$: A BRIEF REVIEW

We shall briefly describe the test $Q^{(T)}$ in order to motivate the min $Q^{(T)}$ method of estimation (for more details, see Singh 1985, Singh and Kumar 1986). Let I denote the number of disjoint domains and v_i denote the parameter of interest for the i-th domain. Consider a model for $v = (v_1, v_2, \ldots, v_I)'$ as

$$H_0: h(v) = X\theta \tag{2.1}$$

where X is a known $I \times r$ matrix of full rank r, θ is an r-vector of unknown parameters, and h is a continuously differentiable one-to-one function, for instance, log or logit.

Let $\hat{v}$ denote the I-vector of survey estimates. Assume that under a suitable central limit theorem

$$\hat{v} \stackrel{a}{\sim} MVN(v, \Gamma/n) \tag{2.2}$$

where " $\stackrel{a}{\sim}$ " means "asymptotically distributed as", n is the total sample size, and Γ is the asymptotic covariance matrix of $\sqrt{n}\,(\hat{v} - v)$.

Now, choose a small level $\epsilon (>0)$ of dimensionality reduction (eg., .01 or .005 can be taken as working values of ϵ). Find a number T such that with the eigenvalues $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_I$ of the estimated covariance matrix $\hat{\Gamma}$, we have

$$T = \max \left\{ t: t > r \text{ and } \sum_{i=t}^{I} \hat{\lambda}_i / \sum_{i=1}^{I} \hat{\lambda}_i \geq \epsilon \right\}. \tag{2.3}$$

The variable T, although random, can be regarded as fixed for our asymptotics. It may be noted that if there are no relatively very small eigenvalues (i.e. if $\hat{\Gamma}$ is not ill-conditioned), then there will usually be no effect of dimensionality reduction for small ϵ and T will coincide with I in those situations.

Consider the problem of testing H_0 against alternatives K_0: $h(v) \neq X\theta$ in the class of tests based on the first T principal components W of $\hat{v}$. Let the normalized eigenvector corresponding to $\hat{\lambda}_i$ be P_i (it need not be unique) and let M_T denote the $I \times T$ matrix of eigenvectors P_i's corresponding to the first T largest eigenvalues. Then

$$W = M_T'\hat{v} \overset{\cdot}{\sim} MVN(\mu, D_T/n), \tag{2.4}$$

where

$$\mu = M_T'v, \; D_T = diag(\lambda_1, \ldots, \lambda_T).$$

Based on W, the original testing problem concerning an I-dimensional v is reduced to testing a hypothesis about the T-dimensional parameter μ given by

$$H_0': \mu = M_T'h^{-1}(X\theta) \text{ vs } K_0': \mu \neq M_T'h^{-1}(X\theta). \tag{2.5}$$

The test statistic $Q^{(T)}$ can be obtained as a score statistic of principal components by employing the approximate likelihood of θ given by the limiting distribution (2.4) of W for computing the efficient scores (see Cox and Hinkley 1974, p. 321-324). We shall refer to $Q^{(T)}$ as a generalized score test that would reject H_0 for large values of the quadratic form

$$\begin{aligned} Q^{(T)}(\theta^o) &= Y(\theta^o)'\Delta_T Y(\theta^o) - Z_T(\theta^o)'\Lambda_T Z_T(\theta^o) \\ &\overset{\cdot}{\sim} \chi^2_{T-r} \end{aligned} \tag{2.6}$$

where

$$Y(\theta^o) = \hat{v} - v(\theta^o), \; \Delta_T = n \sum_{i=1}^{T} (P_i P_i' / \hat{\lambda}_i),$$

$$Z_T(\theta^o) = B'\Delta_T Y(\theta^o), \; B = (\partial v / \partial \theta), \; \Lambda_T = (B'\Delta_T B)^{-1},$$

and θ^o is some fixed point in the null parameter space. In computing $Q^{(T)}$, any root n-consistent estimate of θ under H_0 can be substituted for θ^o, such as pseudo mle of θ. Notice that $Q^{(T)}$ of (2.6) is in fact a quadratic form in W but is expressed in $\hat{v}$ for the sake of convenience.

For testing H_0 vs K_0 in the class of tests based on W, the asymptotic optimality of the test $Q^{(T)}$ follows from that of the score statistic. For small $\epsilon > 0$, $\hat{v}$ and W will be close in the sense that principal components provide the best possible way of dimensionality reduction with a minimum loss of information. Thus $Q^{(T)}$ (for small ϵ) is expected to be robust with respect to the test Q corresponding to no dimensionality reduction. However, Q may be unstable (in the sense of inflated Type I error rate) for finite samples due to possible near singularity of $\hat{\Gamma}$. The test $Q^{(T)}$ is expected to control this problem of instability at the cost of sacrificing some information in the data that gives rise to possibly unreliable components in Q in the directions of eigenvectors that correspond to relatively very small eigenvalues. The loss of information implies that the test $Q^{(T)}$ will lack power for alternatives in directions of (near) singularities. However, this loss of power is offset by the gain in control of Type I error rate. The instability control is further ensured by the fact that, since H_0 is a subset of H_0', $Q^{(T)}$ will be a conservative test for H_0.

A special asymptotically equivalent version of $Q^{(T)}$ (θ^o) which has a simpler expression similar to that of the standard Pearson-Fisher's X^2, is obtained by replacing θ^o with an estimator $\tilde{\theta}$ that minimizes the expression $(\hat{v} - v(\theta))'\Delta_T(\hat{v} - v(\theta))$. We then have

$$\begin{aligned} Q^{(T)}\ (\tilde{\theta}) &= Y(\tilde{\theta})'\Delta_T Y(\tilde{\theta}) \\ &= \sum_{i=1}^{T} [P_i'(\hat{v} - v(\tilde{\theta}))]^2/\hat{\lambda}_i \qquad (2.7) \\ &\dot{\sim}\ \chi^2_{T-r} \end{aligned}$$

Henceforth we assume that, for a given data vector $\hat{v}$, a model H_0 has been deemed appropriate based on the test $Q^{(T)}$ or some other test such as the adjusted X^2 test. In the next section, we give an asymptotically efficient method of estimating parameters θ under H_0, using the statistic $Q^{(T)}$. The θ estimates in turn provide a set of smoothed estimates of v corresponding to survey estimates $\hat{v}$.

3. THE MIN $Q^{(T)}$ ESTIMATOR

Consider the approximate likelihood for the mean μ of the first T principal components W of $\hat{v}$, given earlier by (2.4). Suppose the model H_0: $h(v) = X\theta$ is accepted. Then, the kernel function $K(\theta)$ of the approximate likelihood for $\mu(\theta)$ is given by

$$\begin{aligned} K(\theta) &= (W - \mu(\theta))'D_T^{-1}(W - \mu(\theta)) \\ &= (\hat{v} - v(\theta))'\Delta_T(\hat{v} - v(\theta)) \qquad (3.1) \end{aligned}$$

The value $\tilde{\theta}$ that minimizes $K(\theta)$ corresponds to the mle of θ for the approximate likelihood of μ under H_0. The estimator $\tilde{\theta}$ will be asymptotically efficient (or best asymptotically normal (BAN) in the sense of Neyman, 1949), in a restricted class, namely in the class of estimates based on W. Following the min X^2 estimator of Neyman (1949), the estimator $\tilde{\theta}$ was termed min $Q^{(T)}$ estimator in Singh (1985). Notice that the estimator $\tilde{\theta}$ depends on the level ϵ of dimensionality reduction via Δ_T. Thus $\tilde{\theta}$ varies if ϵ does.

The smoothed estimates of v under H_0 based on W can be obtained as follows. Find $\tilde{\theta}$ which minimizes $K(\theta)$, i.e. $\tilde{\theta}$ is the solution of r equations

$$B'\Delta_T(\hat{v} - v(\theta)) = 0 \qquad (3.2)$$

where both $B(=\partial v/\partial\theta)$ and v involve θ. An iterative procedure such as Newton-Raphson can be used to solve (3.2). Weighted least squares (WLS) estimates or pseudo mle can be used as possible initial choices for θ. We can then compute the min $Q^{(T)}$ estimator of v as

$$\tilde{v} = h^{-1}(X\tilde{\theta}). \tag{3.3}$$

The asymptotic behaviours of $\tilde{\theta}$ and $\tilde{v}$ are given by the following proposition.

Proposition 3.1 As before, let Λ_T denote $(B'\Delta_T B)^{-1}$. We have

$$\begin{aligned} &\text{(a)}\quad \tilde{\theta} - \theta \approx \Lambda_T B'\Delta_T(\hat{v} - v(\theta)) \overset{\cdot}{\sim} MVN(0,\Lambda_T) \\ &\text{(b)}\quad \tilde{v} - v \approx B\Lambda_T B'\Delta_T(\hat{v} - v(\theta)) \overset{\cdot}{\sim} MVN(0,B\Lambda_T B') \end{aligned} \tag{3.4}$$

where " $\approx$ " indicates that the difference between the two sides is negligible in probability.

The proof follows from the application of the δ-method to the functions $B'\Delta_T(\hat{v} - v(\theta))$ and $\tilde{v} - v(\theta)$, which gives

$$B'\Delta_T(\hat{v} - v(\theta)) - (B'\Delta_T B)(\tilde{\theta} - \theta) = o_p(1),$$

$$\tilde{v} - v(\theta) - B(\tilde{\theta} - \theta) = o_p(1).$$

From the above proposition it follows that the asymptotic covariance matrix of the min $Q^{(T)}$ estimator $\tilde{\theta}$ is the inverse of the information matrix $B'\Delta_T B$ for θ, which was obtained from the approximate likelihood of θ as given by (2.4). It can then be seen that in the absence of dimensionality reduction, the estimator $\tilde{\theta}$ will be asymptotically equivalent to the WLS estimator of Koch, Freeman and Freeman (1975). As mentioned in the Introduction, the WLS estimator generally shows unstable finite sample behaviour because of the inefficient estimation of Γ. In contrast, the estimator $\tilde{\theta}$ for a given $\epsilon > 0$ is expected to show stable finite sample behaviour in the sense that it can be approximated well by its asymptotic behaviour. This is achieved at the cost of compromising the asymptotic optimality of $\tilde{\theta}$ by restricting it to a smaller class, namely the class of estimates based on the first T principal components W. The WLS estimator, on the other hand, is asymptotically optimal in a wider class, namely the class of estimates based on the full data vector $\hat{v}$. If, for a small ϵ, the $Q^{(T)}$ test statistic indicates insignificance for H_0, then the corresponding min $Q^{(T)}$ estimator $\hat{v}$ will likely provide a robust alternative to the WLS estimator.

4. MIN $Q^{(T)}$ ESTIMATES OF UNEMPLOYMENT RATES

The Canadian labour force survey (LFS) data for October 1980 was analysed by Kumar and Rao (1984, 1986) and Roberts, Rao and Kumar (1987). Both sets of authors applied the extension of the Rao-Scott adjusted X^2 method to the case of logistic regression. They showed that the logit model given below provided an adequate fit to the survey estimates of employment rates $(v_{j\ell})$ for the table of 60 cells cross-classified by age (10 categories) and education (6 categories). The model is

$$log\,\frac{v_{j\ell}}{1 - v_{j\ell}} = \beta_0 + \beta_1 A_j + \beta_2 A_j^2 + \beta_3 E_\ell \tag{4.1}$$

where A_j represents the midpoint $12 + 5j$ for j-th age group $(j = 1, \ldots, 10)$, and $E_\ell(\ell = 1, \ldots, 6)$ represents the median years of schooling with values 7, 10, 12, 13, 14 and 16.

The model (4.1) can be expressed in the notation of Section 2 by numbering the sixty cells lexicographically. Thus, (4.1) can be rewritten as $h(v) = X\theta$, where v is the vector of employment rates, h is the logit function, X is a 60×4 matrix whose i-th row is $(1, \mathrm{A}_i, \mathrm{A}_i^2, \mathrm{E}_i)$, and θ is $(\beta_0, \beta_1, \beta_2, \beta_3)'$. We also have

$$H = (\partial h / \partial v) = D_v^{-1} D_{1-v}^{-1}, B = H^{-1} X, \tag{4.2}$$

where D_v and D_{1-v} are diagonal matrices with diagonal elements given by the subscripts.

The pseudo mle of θ for the model (4.1) were obtained by Kumar and Rao (1984) under the pseudo product-binomial likelihood as

$$\bar{\theta} = (-3.10, 0.211, -0.00218, 0.1509)'. \tag{4.3}$$

They also computed Rao-Scott's first order adjusted X^2 (G_c^2 in their notation) as 55.3, which shows acceptance of the model (2.1) when referred to the χ_{56}^2 distribution.

The $Q^{(T)}$ method was applied for testing (4.1) (see Singh 1985, and Singh and Kumar 1986) also resulting in the acceptance of the model (4.1). For $\epsilon = .01$, T turns out to be 51 using the estimated covariance matrix $\hat{\Gamma}$ as obtained by Kumar and Rao (1984). Now using the pseudo mle $\bar{\theta}$, we have

$$Q^{(51)}(\bar{\theta}) = 58.665 - 4.454 = 54.211 \tag{4.4}$$

When $\epsilon = .005$, T is found to be 54, and

$$Q^{(54)}(\bar{\theta}) = 67.774 - 2.343 = 65.431 \tag{4.5}$$

When $\epsilon = 0$, $T = 58$ because two cells had zero observed unemployment rates. In this case,

$$Q^{(58)}(\bar{\theta}) = 87.302 - 0.812 = 86.49 \tag{4.6}$$

By referring $Q^{(51)}$ to the χ_{47}^2 distribution, $Q^{(54)}$ to a χ_{50}^2 and $Q^{(58)}$ to a χ_{54}^2 distribution, it is clear that both $Q^{(51)}$ and $Q^{(54)}$ accept (4.1) while $Q^{(58)}$ does not. An instability check can be performed by considering the difference $Q^{(58)} - Q^{(T)}$ for $T = 51, 54$, which can be seen to be highly significant when referred to the χ_{58-T}^2 distribution. These indicate presence of the instability problem in the Q-test statistic that corresponds to no dimensionality reduction. It is clear that WLS test would also have an instability problem due to the difficulty involved in inverting the matrix $\hat{\Gamma}$ which is singular. Thus, min $Q^{(T)}$ method would be preferable to min Q or WLS methods. In the interests of reducing loss of information, the method with the largest value of T is recommended, providing of course that the corresponding $Q^{(T)}$ shows insignificance for the model.

We shall now compute asymptotically efficient estimates. Neither min Q nor WLS estimates were computed because $\hat{\Gamma}$ was singular. The min $Q^{(T)}$ estimates $\tilde{\theta}$ were computed for $\epsilon = .005$ and $\epsilon = .01$ by using the Newton-Raphson iterative procedure and $\bar{\theta}$ as the initial estimate of θ for solving (3.2). The values of $\tilde{\theta}_T$ and $Q^{(T)}(\tilde{\theta})$ (in this case the negative term in (2.6) drops out) for $\epsilon = .005$, $T = 54$ were obtained as

$$\tilde{\theta}_{54} = (-2.7112, 0.1944, -0.00196, 0.1432)', \text{ and}$$

$$Q^{(54)}(\tilde{\theta}_{54}) = 63.4737 \tag{4.7}$$

For $\epsilon = .01$, $T = 51$, we have

$$\tilde{\theta}_{51} = (-2.6739,\ 0.19702,\ -0.00202,\ 0.1364)', \text{ and}$$

$$Q^{(51)}(\tilde{\theta}_{51}) = 55.2518. \tag{4.8}$$

Conclusions based on the statistic $Q^{(T)}(\tilde{\theta})$ for both $T = 54$ and 51 agree with those obtained from $Q^{(T)}(\bar{\theta})$.

Table 1 gives efficiencies relative to survey estimates of unemployment rates $1 - v$ for all cells (excepts two with zero observed unemployment rates) corresponding to the three smoothed estimates. The three smoothed estimates are the pseudo mle, min $Q^{(51)}$, and min $Q^{(54)}$. The pseudo mle variances are taken from Kumar and Rao (1986), while those for min $Q^{(T)}$ estimates are obtained from the diagonal elements of $B \wedge_T B'$ of (3.4). As noted by Kumar and Rao (1986) for pseudo mle, smoothed estimates based on min $Q^{(T)}$ also lead to considerable efficiency gains over survey estimates. The relative trace efficiency of smoothed estimates over survey estimates is 17.9 for pseudo mle, 18.95 for min $Q^{(51)}$ and 19.88 for min $Q^{(54)}$ estimates. Thus the min $Q^{(T)}$ estimators provide a slight improvement in the

Table 1
Efficiencies of Smoothed Estimates of Unemployment rates relative to Survey Estimates[a]

Cell Number	Min $Q^{(51)}$	Min $Q^{(54)}$	Pseudo mle	Cell Number	Min $Q^{(51)}$	Min $Q^{(54)}$	Pseudo mle
1	5.87	5.74	5.44	31	9.01	9.32	8.65
2	3.62	3.62	3.28	32	8.76	9.46	10.68
3	3.45	3.55	3.12	33	36.93	42.93	51.59
4	52.45	51.65	43.46	34	51.55	60.23	81.12
5	104.77	114.30	96.21	35	69.76	79.93	98.37
7	5.33	5.14	4.38	36	9.17	11.01	15.07
8	9.36	9.53	8.09	37	3.48	3.01	3.45
9	6.85	7.16	6.70	38	13.74	15.91	18.00
10	25.65	28.40	26.31	39	66.87	80.98	97.30
11	13.34	14.13	17.73	40	154.81	187.73	221.50
12	27.74	30.85	30.85	41	49.14	67.56	80.61
13	8.64	8.84	7.15	42	17.32	21.73	24.98
14	13.84	13.84	12.37	43	8.57	9.28	8.49
15	8.20	8.49	9.47	44	27.42	31.65	30.74
16	23.14	24.09	27.75	45	58.55	70.67	75.72
17	18.20	18.20	21.49	46	94.11	114.13	121.49
18	9.87	11.14	12.51	47	82.12	112.65	108.52
19	15.87	16.03	13.66	48	26.54	39.41	41.22
20	11.44	11.98	12.56	49	4.95	5.37	4.41
21	12.39	12.39	15.53	50	12.11	14.10	11.17
22	24.83	24.83	32.02	51	6.75	8.61	7.50
23	16.43	18.16	21.55	52	8.83	11.45	9.90
24	6.98	7.83	10.06	53	52.64	71.49	61.14
25	7.49	7.74	6.99	55	3.59	3.93	3.03
26	10.33	11.33	12.32	56	7.33	8.96	8.23
27	6.47	7.18	8.69	57	23.50	29.83	22.11
28	125.81	140.57	172.91	58	221.23	294.59	208.77
29	33.88	38.13	52.00	59	6.45	8.82	6.62
30	14.89	15.24	20.43	60	38.90	52.84	41.96

[a] Cells 6 and 54 are omitted due to zero observed unemployment rates.

efficiency of smoothed estimates compared to pseudo mle. With regard to performance over individual cells Table 1 indicates that the pseudo mle behave very well as compared to efficient min $Q^{(T)}$ estimates for the example under consideration.

5. CONCLUDING REMARKS

For computing pseudo mle, the working form of the likelihood function corresponds to simple random samples (i.e. multinomial or product-multinomial sampling). The pseudo mle do provide consistent estimates of model parameters without requiring an estimate of the covariance matrix Γ. However, the pseudo mle are not asymptotically efficient for complex survey data. By contrast, the min $Q^{(T)}$ estimates are asymptotically efficient with respect to the class of estimates based on W (the first T principal components of the vector $\hat{v}$ of survey estimates). For investigating the relative performance of pseudo mle and min $Q^{(T)}$, it would be desirable to perform a simulation study for efficiency comparisons. The min $Q^{(T)}$ estimates do take into account of the underlying complex design by employing an appropriate $\hat{\Gamma}$. If $\hat{\Gamma}$ is not ill-conditioned, i.e. it has no relatively very small eigenvalues, then there is no instability problem with the well known WLS estimates which are of course asymptotically efficient. In this case, it will usually turn out that there is no dimensionality reduction for small ϵ, that T will coincide with I and that there will be no loss in efficiency of min $Q^{(T)}$ estimates in comparison with WLS estimates. However, given the instability problem common with cross-classified categorical survey data, the min $Q^{(T)}$ estimates are expected to provide a robust alternative to WLS estimates.

ACKNOWLEDGEMENT

The second author's research was supported by Statistics Canada and the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

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

FAY, R.E. (1985). A jackknifed chi-squared test for complex samples. *Journal of the American Statistical Association*, 80, 148-157.

IMREY, P.B., KOCH, G.G., and STOKES, M.E. (1982). Categorical data analysis: Some reflections on the log-linear model and logistic regression. Part II: Data analysis. *International Statistical Review*, 50, 35-63.

KOCH, G.G., FREEMAN, D.H. Jr., and FREEMAN, J.L. (1975). Strategies in the multivariate analysis of data from complex surveys. *International Statistical Review*, 43, 59-78.

KUMAR, S., and RAO, J.N.K. (1984). Logistic regression analysis of Labour Force Survey Data. *Survey Methodology*, 10, 62-81.

KUMAR, S., and RAO, J.N.K. (1986). On smoothed estimates of unemployment rates from labour force survey data. *In Small Area Statistics: An International Symposium '85* (Eds. R. Platek, and M.P. Singh), Ottawa: Carleton University.

NEYMAN, J. (1949). Contribution to the Theory of the X^2 test. *In Proceedings of the First Berkeley Symposium on Mathematical Statistics and Probability* (Ed. J. Neyman), Berkeley: University of California Press, 230-273.

RAO, J.N.K., and SCOTT, A.J. (1981). The analysis of categorical data from complex sample surveys: chi-squared tests for goodness of fit and independence in two way tables. *Journal of the American Statistical Association*, 76, 221-230.

RAO, J.N.K., and SCOTT, A.J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *Annals of Statistics*, 12, 46-60.

ROBERTS, G.R. (1985). *Contributions to chi-squared tests with survey data.* Ph.D. dissertation, Carleton University, Ottawa.

ROBERTS, G., RAO, J.N.K., and KUMAR, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, 74, 1-12.

SINGH, A.C. (1985). On optimal asymptotic tests for analysis of categorical data from sample surveys. Working Paper, Social Survey Methods Division, Statistics Canada.

SINGH, A.C., and KUMAR, S. (1986). Categorical data analysis for complex surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, (forthcoming).

Survey Methodology, June 1987
Vol. 13, No. 1, pp. 85-92
Statistics Canada

A Sampling Procedure with Inclusion Probabilities Proportional to Size

A. DEY and A.K. SRIVASTAVA[1]

ABSTRACT

A new unequal probability sampling scheme for selecting n (> 2) units without replacement from a finite population is proposed. This scheme ensures that the inclusion probabilities are proportional to sizes. It has the advantage of simplicity in selection and estimation and also provides a non-negative variance estimator. The variance of the Horvitz-Thompson (H-T) estimator under the proposed scheme is shown to be smaller than that of the customary estimator in probability proportional to size sampling with replacement. The proposed scheme also compares favourably with the without replacement scheme suggested by Sampford (1967) in an empirical study on a few natural populations.

KEY WORDS: Unequal probability sampling; Horvitz-Thompson estimator.

1. INTRODUCTION

In unequal probability sampling of n units without replacement from a finite population containing N units, if π_i denotes the inclusion probability of the i-th unit in the sample $i = 1, 2, \ldots, N$, the Horvitz and Thompson (1952) estimator (H-T estimator) of Y, the population total of the study variable y, is given by

$$\hat{Y} = \sum_{i \epsilon s} (y_i / \pi_i), \tag{1.1}$$

where y_i is the y-value for the i-th unit and the summation extends over the units included in the sample. The variance of $\hat{Y}$ is

$$Var\ (\hat{Y}) = \sum_{i=1}^{N} \sum_{j>i}^{N} (\pi_i \pi_j - \pi_{ij}) (y_i / \pi_i - y_j / \pi_j)^2_, \tag{1.2}$$

where π_{ij} denotes the joint inclusion probability of the i-th and j-th units in the sample $(i \neq j, i, j = 1, 2, \ldots, N)$.

Considerable reduction in the variance of $\hat{Y}$ can be expected if the sampling scheme ensures that π_i are proportional to a given measure of size, say, x_i for $i = 1, 2, \ldots, N$, where it is assumed that x_i are nearly proportional to y_i. Sampling schemes in which π_i are proportional to x_i are termed Inclusion Probability Proportional to Size (IPPS) schemes. For a comprehensive account of unequal probability sampling procedures, including IPPS sampling schemes, the reader is referred to the monograph of Brewer and Hanif (1983).

Some desirable properties of an unequal probability scheme without replacement in general, and IPPS schemes in particular, are simplicity in selection and estimation, availability of a non-negative variance estimator, and better efficiency than with the probability proportional to size (PPS) with replacement strategy. Unfortunately, for sample size greater than two, not many of the available procedures meet these requirements fully.

[1] A. Dey and A.K. Srivastava, Indian Agricultural Statistics Research Institute, Library Avenue, New Delhi 110012, India.

In this paper, an IPPS sampling scheme is suggested for arbitrary sample sizes, $n > 2$. The procedure is rather simple both in sample selection and at the estimation stage since compact expressions for π_{ij} are available. It has also been possible to provide a positive estimator of variance of the *H-T* estimator of Y. The performance of the *H-T* estimator under the proposed scheme is compared with the PPS with replacement strategy and a simple sufficient condition is derived under which the performance of the former strategy is superior to that of the latter. An empirical study on a few natural populations indicates that the proposed scheme compares favourably with that suggested by Sampford (1967).

2. THE SAMPLING PROCEDURE

Consider a population of N units with y as the study variable and x, an auxiliary variable, as the size. It is assumed that x-values are known for all the population units. A sample of size $n\,(> 2)$ is to be selected. To start with, it is assumed that n is *even*.

Divide the population into $m\,(> n/2)$ groups such that the i-th group contains $N_i\,(> 2)$ units $(i = 1, 2, \ldots, m)$ and, for each group,

$$X_i/X > (n - 2)/[n(m - 1)], \tag{2.1}$$

where

$$X_i = \sum_{u=1}^{N_i} x_{i_u},$$

x_{i_u} is the value of x for the u-th unit in the i-th group and $X = X_1 + X_2 + \ldots + X_m$.

Equation (2.1) is satisfied if the X_i $(i = 1, 2, \ldots, m)$ are made nearly equal. It has been seen in actual populations, considered by Rao and Bayless (1969) and others, that this condition is satisfied for quite a few values of m if the groups are so formed that their sizes, X_i, are nearly equal. Rao and Lanke (1984) suggested a grouping procedure in which N units are grouped into R groups such that group totals, X_i, are nearly equal and group sizes are either $[N/R]$ or $[N/R] + 1$, where $[x]$ is the largest integer contained in x. For the formation of groups, the Rao-Lanke procedure may also be tried.

Having formed the m groups, the suggested sampling procedure consists of the following steps:

Step 1. Select $n/2$ groups out of the m groups using Midzuno's (1951) sampling procedure with probabilities $\{P_i'\}$, that is, select one group with probability

$$P_i' = [n(m - 1)P_i - (n - 2)]/(2m - n), \text{ with } P_i = X_i/X,$$

and the remaining $(n/2) - 1$ groups with equal probabilities without replacement.

Step 2. From each of the selected groups, select two units by any IPPS procedure, say by Durbin's (1967) procedure, that is, in the i-th selected group $(i = 1, 2, \ldots, n/2)$ select one unit with probability

$$p_{i_u|i} = x_{i_u}/X_i,$$

and the second unit with revised probability

$$p_{i_u | i_v} = x_{i_v} [1/(X_i - 2x_{i_v}) + 1/(X_i - 2x_{i_u})] / D_i,$$

where

$$D_i = [1 + \sum_{u=1}^{N_i} x_{i_u} / (X_i - 2x_{i_u})].$$

For this sampling procedure, the inclusion probability for the i_u-th unit is evidently given by

$$\pi_{i_u} = n\, p_{i_u} \tag{2.2}$$

where

$$p_{i_u} = x_{i_u} / X.$$

Also, the joint inclusion probabilities for a pair of units are given by

$$\pi_{i_u i_v} = \frac{n\, p_{i_u} p_{i_v}\, (P_i - p_{i_u} - p_{i_v})}{D_i (P_i - 2\, p_{i_u})\ (P_i - 2\, p_{i_v})} \tag{2.3}$$

and

$$\pi_{i_u j_v} = \frac{n\, (n - 2)\, p_{i_u} p_{j_v}}{(m - 1)(m - 2)\, P_i P_j}\ [(m - 1)(P_i + P_j) - 1], \tag{2.4}$$

$$i \neq j,\ i, j = 1, 2, \ldots, m.$$

Thus we see that the proposed scheme is indeed an IPPS scheme.

As mentioned earlier, at step 2 of the proposed procedure, any IPPS scheme for selecting two units can be used. Since the procedure of Durbin (1967), which is equivalent to those of Rao (1963) and Brewer (1963), generally performs well, it has been adopted at step 2.

3. A VARIANCE ESTIMATOR

Two well-known unbiased estimators of $Var(\hat{Y})$ are due to Horvitz and Thompson (1952) and Yates and Grundy (1953). Both these estimators, however, suffer from the drawback that they sometimes assume negative values. In this section, a positive estimator of variance is proposed that utilizes the two-stage nature of the proposed sampling scheme.

Using a result due to Des Raj (1966), an unbiased estimator of $Var(\hat{Y})$ is given by

$$\hat{V}(\hat{Y}) = \sum_{i=1}^{n/2} \pi_i^{-1} \sum_{u < v}\sum \left[\frac{\pi_{i_u | i}\, \pi_{i_v | i}}{\pi_{i_u i_v | i}} - 1\right] \left[\frac{y_{i_u}}{\pi_{i_u | i}} - \frac{y_{i_v}}{\pi_{i_v | i}}\right]^2$$

$$+ \sum_{i <}^{n/2} \sum_{j}^{n/2} \left(\frac{\pi_i \pi_j}{\pi_{ij}} - 1\right) \left[\frac{\hat{Y}_i}{\pi_i} - \frac{\hat{Y}_j}{\pi_j}\right]^2, \tag{3.1}$$

where

$$\pi_i = nP_i/2,$$

$$\pi_{ij} = \frac{n\,(n-2)}{4\,(m-2)}\,\{(P_i + P_j) - 1/(m-1)\},$$

$$\pi_{i_u|i} = 2p_{i_u}/P_i,$$

$$\pi_{i_u i_v|i} = \frac{2p_{i_u}p_{i_v}(P_i - p_{i_u} - p_{i_v})}{D_iP_i(P_i - 2p_{i_u})(P_i - 2p_{i_v})},$$

and

$$\hat{Y}_i = \sum_{u=1}^{2} y_{i_u}/\pi_{i_u|i}, \tag{3.2}$$

y_{i_u} being the y-value of the u-th unit in the i-th group.

The two terms in the right side of (3.1) correspond to the Yates-Grundy variance estimator in Durbin's and Midzuno's procedures. Since under these two sampling procedures the Yates-Grundy estimator of variance is always positive, it follows that the variance estimator given by (3.1) is also positive. However, the estimator in (3.1) is neither the Horvitz-Thompson nor the Yates-Grundy variance estimator.

4. COMPARISON WITH PPS WITH REPLACEMENT STRAGEGY

In this section, we compare the efficiencies of the following two strategies:

Strategy 1. The proposed sampling scheme in conjunction with the Horvitz-Thompson estimator.

Strategy 2. PPS sampling with replacement in conjunction with the customary estimator.

Strategy 1 is more efficient than Strategy 2 if and only if

$$\begin{aligned} &\sum_{i=1}^{m} \sum_{u \neq v}^{N_i}\sum \pi_{i_u i_v}(y_{i_u}/p_{i_u} - Y)(y_{i_v}/p_{i_v} - Y) \\ &+ \sum_{i \neq j}^{m}\sum \sum_{u}^{N_i} \sum_{v}^{N_j} \pi_{i_u j_v}(y_{i_u}/p_{i_u} - Y)(y_{j_v}/p_{j_v} - Y) < 0. \end{aligned} \tag{4.1}$$

After some lengthy but routine algebra, the inequality (4.1) boils down to

$$\begin{aligned} &- \sum_{i=1}^{n} (n/D_i) \sum_{u=1}^{N_i} (y_{i_u} - Y_ip_{i_u}/P_i)^2/(P_i - 2p_{i_u}) \\ &- n(n-2)\left[\sum_{i=1}^{m} (Y_i/P_i - Y)\right]^2 / [(m-2)(m-1)] \\ &- n(m-2)^{-1} \sum_{i=1}^{m} [\{(2n - m - 2)P_i - (n-2)(m-1)^{-1}\}(Y_i/P_i - Y)^2] < 0, \end{aligned} \tag{4.2}$$

where $$Y_i = \sum_u y_{i_u}.$$

Obviously, (4.2) holds if

(i) $(2n - m - 2) > 0$, and

(ii) $P_i > (n - 2)/[(m - 1)(2n - m - 2)]$. (4.3)

Also, since we are using Midzuno's procedure at the first stage with revised probabilities $\{P_i'\}$, each P_i must satisfy (2.1), that is, each P_i must satisfy

$$P_i > (n - 2)/[n(m - 1)].$$

Thus, (4.2) holds if

$$m \leq (n - 2). \tag{4.4}$$

It appears, therefore, that for Strategy 1 to be superior to Strategy 2, m should be chosen such that

$$n/2 < m \leq (n - 2). \tag{4.5}$$

However, it is clear that (4.4) is merely a sufficient condition and is not necessary. For $n > 6$, condition (4.5) offers a somewhat wide choice for the value of m, while for $n = 6$, (4.5) implies that $m = 3$. For $n = 4$, (4.5) does not lead to a feasible value of m. Therefore, for $n = 4$, an investigation into the performance of Strategy 1 has been taken up for various values of m, not constrained by (4.5), on certain natural populations. A description of the populations appears in Table 1. Table 2 presents the relative efficiency of Strategy 1 compared to Strategy 2 for the populations in Table 1. The performance of the H-T estimator under Sampford's (1967) scheme (called Strategy 3) is also compared with that of Strategy 2.

It can be observed from Table 2 that the performance of the proposed strategy (Strategy 1) compares favourably with that of Sampford (Strategy 3) for most of the populations. Of course, both strategies are superior to Strategy 2.

To achieve the relative efficiency of Strategy 1, the units were grouped in an ad-hoc manner, ensuring only that requirement (2.1) was satisfied. The procedure of Rao and Lanke (1984) was also attempted in forming the groups. However, the Rao-Lanke procedure did not always result in a high efficiency. Further investigations are necessary to decide the 'best' choice of groups. For certain populations, suitable groups satisfying (2.1) could not be formed for higher values of m, and thus, for these cases, the relative efficiencies are not reported in Table 2.

In conclusion, a brief comment on cases in which the desired sample size, n, is *odd* is in order. An IPPS sample for odd n may be obtained by selecting $(n + 1)$ units by the suggested procedure and then randomly discarding one unit. The expressions for π_i and π_{ij} under this procedure are straghtforward. Obviously, when one of the sample units out of $(n + 1)$ is discarded at random, the resulting sample consists of two units from each of the $(n - 1)/2$ groups and just one unit from one of the groups. An unbiased and positive estimator of $Var(\hat{Y})$ can be obtained, analogous to (3.1), on the basis of the $(n - 1)/2$ groups, each containing two units in the sample.

Table 1

Description of the Populations

Pop. Number	Source	N	y	x
1.	Des Raj (1965)	20	Number of households	Eye-estimated number of households
2.	Rao (1963)	14	Corn acreage in 1960	Corn acreage in 1958
3.	Cochran (1963, p. 204)	10	Weight of peaches	Eye-estimated weight of peaches
4.	Hanurav (1967)	20	Population in 1967	Population in 1957
5.	Hanurav (1967)	19	Population in 1967	Population in 1957
6.	Hanurav (1967)	16	Population in 1967	Population in 1957
7.	Hanurav (1967)	17	Population in 1967	Population in 1957
8.	Cochran (1963, p. 325)	10	Number of persons per block	Number of rooms per block
9.	Cochran (1963, p. 156, cities 1-16)	16	Population in 1930	Population in 1920
10.	Cochran (1963, p. 156, cities 33-49)	17	Population in 1930	Population in 1920
11.	Sampford (1962, p. 61)	35	Oats acreage in 1957	Oats acreage in 1947
12.	Sukhatme and Sukhatme (1970, p. 256, circles 1-20)	20	Wheat acreage	Number of villages
13.	Sukhatme and Sukhatme (1970, p. 256, circles 21-40)	20	Wheat acreage	Number of villages
14.	Yates (1960, p. 163)	20	Volume of timber	Eye-estimated volume of timber

Table 2

Percent Relative Efficiencies of
Strategies 1 and 3 over Strategy 2 for the
Populations in Table 1 ($n = 4$)

Pop. Number	Strategy 1				Strategy 3
	$m = 3$	4	5	6	
1.	130.1	118.7	120.8	124.5	127.8
2.	132.6	130.2	–	–	127.1
3.	149.1	–	–	–	147.9
4.	120.7	120.6	122.7	129.7	117.8
5.	129.1	138.7	158.7	–	125.1
6.	158.0	173.1	–	–	139.5
7.	151.9	144.8	169.2	–	131.9
8.	168.5	–	–	–	145.5
9.	118.3	116.3	–	–	109.5
10.	126.6	–	–	–	112.2
11.	113.8	116.2	135.6	129.9	113.8
12.	117.4	128.0	119.0	–	119.3
13.	122.2	120.6	–	–	119.7
14.	124.8	123.1	115.4	113.2	116.3

ACKNOWLEDGEMENTS

The authors would like to thank the referee for making many useful suggestions on the first draft.

REFERENCES

BREWER, K.R.W. (1963). A model of systematic sampling with unequal probabilities. *Australian Journal of Statistics*, 5, 5-13.

BREWER, K.R.W. and HANIF, M. (1983). *Sampling with Unequal Probabilities*. Lecture Notes in Statistics, No. 15. New York: Springer-Verlag.

COCHRAN, W.G. (1963). *Sampling Techniques*, (2nd. ed.). New York: John Wiley.

DES RAJ (1965). Variance estimation in randomized systematic sampling with probability proportional to size. *Journal of the American Statistical Association*, 60, 278-284.

DES RAJ (1966). Some remarks on a simple procedure of sampling without replacement. *Journal of the American Statistical Association*, 61, 391-396.

DURBIN, J. (1967). Design of multi-stage surveys for the estimation of sampling errors. *Journal of the Royal Statistical Society*, Ser. C, 16, 152-164.

HANURAV, T. (1967). Optimum utilization of auxiliary information: πps sampling of two units from a stratum. *Journal of the Royal Statistical Society*, Ser. B, 29, 379-391.

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.

MIDZUNO, H. (1951). On the sampling system with probability proportionate to sum of sizes. *Annals of the Institute of Statistical Mathematics*, 2, 99-108.

RAO, J.N.K. (1963). On three procedures of unequal probability sampling without replacement. *Journal of the American Statistical Association*, 58, 202-215.

RAO, J.N.K. and BAYLESS, D.L. (1969). An empirical study of the stabilities of estimators and variance estimators in unequal probability sampling of two units per stratum. *Journal of the American Statistical Association*, 64, 540-559.

RAO, J.N.K. and LANKE, J. (1984). Simplified unbiased variance estimation for multistage designs. *Biometrika*, 71, 387-395.

SAMPFORD, M.R. (1962). *An Introduction to Sampling Theory*. Edinburgh: Oliver and Boyd.

SAMPFORD, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54, 499-513.

SUKHATME, P.V. and SUKHATME, B.V. (1970). *Sampling Theory of Surveys with Applications*, (2nd. ed.). Ames, Iowa: Iowa State University Press.

YATES, F. (1960). *Sampling Methods for Censuses and Surveys*, (3rd. ed.). London: Griffin.

YATES, F. and GRUNDY, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, Ser. B, 15, 253-261.

Survey Methodology, June 1987
Vol. 13, No. 1, pp. 93-108
Statistics Canada

Sample Design for the Health and Activity Limitation Survey

D. DOLSON, K. McCLEAN, J.-P. MORIN, and A. THÉBERGE[1]

ABSTRACT

The Health and Activity Limitation Survey is part of the program to establish a data base on the disabled population in Canada. The sample design used for the part of the survey covering the population not living in institutions is described. In addition, the methods used to determine the sizes of the samples and to select the samples are presented.

KEY WORDS: Disability; Stratified sampling; Two-stage sampling; Optimum allocation; Sampling without replacement.

1. INTRODUCTION

As part of the program to obtain more information about Canada's disabled population, the Health and Activity Limitation Survey (HALS) was conducted in the fall of 1986. It is designed to obtain information concerning the nature of the problems experienced by that population and, in general, their daily activities (at home, at work, at school, during travel, and so on). The survey is divided into two parts: one covers the population living in institutions and the other, which is the subject of this article, covers the non-institutional population.

Canada has been divided into 238 subprovincial areas (SPAs). All Quebec and Ontario municipalities with more than 125,000 residents and all municipalities in the other provinces with more than 75,000 residents are included as SPA's. The other areas are made up of groups of census subdivisions respecting geographical contiguity and the provincial boundaries. The number of these areas in each province is proportional to the square root of the population, minus the previously defined municipalities. One of the main objectives of the survey is to generate statistics on the disabled population at the SPA level so that the population's various needs can be analysed in detail. In addition, estimates will be produced for three age groups - namely, children (under 15 years of age), adults (15 to 64 years of age) and seniors (65 years of age and older).

The data was collected in two stages. The first stage involved a multipart question (question 20) included on form 2B of the 1986 Canadian Census of Population. This question asked about the respondents' limitations in various types of activities and their own assessments of their conditions. A copy of question 20 is given in the Appendix. The second stage was implemented some time after the census. It involves a screening questionnaire and follow-up to collect information on the problems and activities of disabled respondents.

The main purpose of the first stage is to separate respondents into two groups: those who answered "yes" to at least one part of question 20 and those who answered "no" to all parts. The aim is to identify beforehand a large part of the potential disabled population, in order to focus survey resources on the target group. However, previous surveys have shown that this question will not identify the entire target population. (See Dolson *et al.* 1984 and Dolson *et al.* 1986.)

[1] D. Dolson, K. McClean, J.-P. Morin, and A. Théberge, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6.

The second stage is HALS. Personal interviews are conducted for the "yes" stratum and telephone interviews are conducted for the "no" stratum. From an operational point of view, the interviews are in two parts - the screening questionnaire and the follow-up.

The screening questionnaire is designed to identify respondents for whom the follow-up questionnaire is relevant. The questionnaire for adults covers the seventeen activities of daily living (ADLs) used in the Canadian Health and Disability Survey in 1983 and 1984, repeats Part (a) of question 20 from the Census, and includes a few questions on mental illness and handicaps (see the Appendix). If an affirmative answer is given to at least one of these questions, the interviewer proceeds with the follow-up; if not, the interview is terminated. Part (a) of the Census question is asked again because there may have been a change in status, either because the response in the Census was given by a proxy, or because the respondent has reassessed his or her own condition.

The screening section in the questionnaire for children includes questions on special aids, activity limitations, attendance at a special school and health conditions or problems. A "yes" answer to at least one of these questions prompts a follow-up interview. The Census question is not repeated because all interviews regarding children require a proxy and the question on activity limitations is equivalent to Part (a) of Census question 20.

The second section of this article describes how the population of Canada has been divided into various subpopulations for estimation purposes. The third section covers the HALS sample design. The fourth section deals with the file of geographic information and projected demographic data for 1986 that was used to create the survey frame. The fifth section explains how the sampling was done.

2. POPULATIONS COVERED

Permanent residents of general and psychiatric hospitals, special care centres or institutions for the elderly or chronically ill, institutions for the physically handicapped and orphanages or children's homes are the subject of a distinct part of the survey - namely, HALS (Institutions). This article will look at the part of the survey covering that portion of the Canadian population not covered by HALS (Institutions) and not residing in jails, military camps, young offender facilities, naval vessels, penal or correctional institutions and collective dwellings in the "others" category (for example, circuses and non-religious communes).

Each enumeration area (EA) whose population is not totally excluded from the survey is classified in one of the following five survey frames:

1. Indian reserves where the 1981 Census was conducted using canvassers;
2. Other Indian reserves;
3. Canvasser EAs;
4. EAs in the Whitehorse, Yellowknife, Pine Point, Hay River and Fort Smith SPAs;
5. All other EAs.

The order of priority for belonging to a frame is 1-2-4-3-5. This means that an EA that is an Indian reserve and situated in the Whitehorse SPA is classified as an Indian reserve.

Each EA is divided in two, with the "yes" EA made up of those persons who would answer "yes" to the Census question, and the "no" EA made up of those who would answer "no" to it. A different sample design is used for each of the five survey frames: all of the "yes" EAs and none of the "no" EAs are selected in the first frame; all of the "yes" EAs and a sample of the "no" EAs are selected in the second frame; none of the "no" EAs and a sample of the "yes" EAs are selected in the third frame; all of the EAs are selected in the fourth frame; and a sample of the "yes" EAs and a sample of the "no" EAs are selected in the fifth frame.

3. SURVEY DESIGN

The sampling method presented in this section was used for survey frames three and five. Because our space is limited, the sample design used for the second survey frame will not be described in this article. (For more information on the HALS methodology, see Dolson *et al.* 1986.)

3.1 Sample Design

Each province is divided into subprovincial areas (SPAs), which are themselves divided into enumeration areas (EAs).

Each EA is divided into a "yes" EA and a "no" EA, the first containing those persons who would answer "yes" to Census question 20, the second containing those persons who would answer "no" to that question. In each SPA, the "yes" EAs are stratified into large and small EAs on the basis of the criterion explained in the fourth section of this paper. Persons belonging to a "yes" EA are associated with a stratum and an SPA in addition to their EA, while persons belonging to a "no" EA are associated only with their EA. In each province, the population is subdivided into three age groups: children (under 15 years of age), adults (15 to 64 years of age) and seniors (65 years of age and older).

The sampling method involves using a two-stage stratified sample design for the "yes" EAs in each SPA and a two-stage sample design for the "no" EAs in the province. The primary units are the EAs and the secondary units are the respondents.

All persons who completed Census form 2B in a "yes" EA selected for the sample are interviewed, along with a third of those in the "no" EAs selected.

3.2 Sample Allocation

This sample design must allow us to minimize sampling costs for a given maximum coefficient of variation of the estimates and a given variance for the estimator $\hat{B}$ of the relative bias B. We define B as the ratio of the number of "no" persons with a characteristic of interest in the province, T_0, to the number of "yes" persons with a characteristic of interest in the province, T_1. By "no" person, we mean an individual who would answer "no" to all parts of Census question 20, and by "yes" person, an individual who would answer "yes" to at least one part of the question.

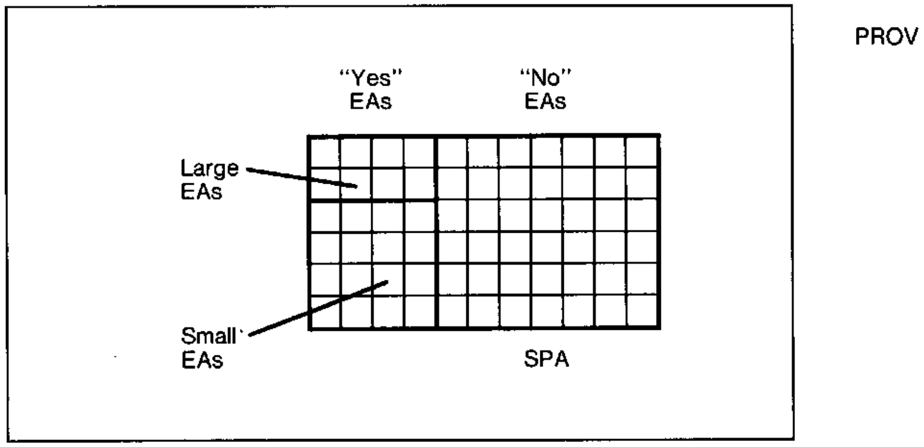

Figure 1. Illustration of Sample Design.

Let N_0 be the number of "no" EAs in the province; N_{jk}, the number of "yes" EAs in stratum j and SPA k in the province; n_0 and n_{jk}, the corresponding sample sizes; and c_0 and c_{jk}, the corresponding unit sampling costs. If we have an N_p SPAs in the province, we therefore want to minimize

$$\sum_{k=1}^{N_p} (c_{1k}n_{1k} + c_{2k}n_{2k}) + c_0 n_0$$

given

$$CV^2(y_k) \leq CV_*^2; \ \mathrm{Var}(\hat{B}) = \mathrm{Var}_*(\hat{B});$$

$$n_{jk} \leq N_{jk}; \ n_{2k} = \lambda_k n_{1k}; \ n_0 \leq N_0$$

$$(j = 1, 2; \ k = 1, \ldots, N_p)$$

where λ_k is the ratio of the expected number of disabled persons in the small EAs to the expected number of disabled persons in the large EAs of SPA k, y_k is the estimated number of "yes" persons who have a characteristic of interest in SPA k, and values marked with an asterisk are constants.

If the sampling fraction in the "yes" EAs is f_1, M_{ijk} is the number of "yes" persons in EA i of stratum j of SPA k in the province and p_{ijk} is the probability of a characteristic of interest for a "yes" person in EA i of stratum j in SPA k, then

$$E(y_k) = Y_k = \sum_{j=1}^{2} \sum_{i=1}^{N_{jk}} M_{ijk}p_{ijk},$$

$$\mathrm{Var}\ (y_k) = \sum_{j=1}^{2} \left\{ \frac{N_{jk}^2}{n_{jk}} \left(1 - \frac{n_{jk}}{N_{jk}}\right) S_{jk}^2 + \frac{N_{jk}}{n_{jk}} \left(\frac{1 - f_1}{f_1}\right) \sum_{i=1}^{N_{jk}} M_{ijk} S_{ijk}^2 \right\},$$

where

$$y_k = \sum_{j=1}^{2} \sum_{i=1}^{N_{jk}} \frac{N_{jk}}{n_{jk}} M_{ijk}p_{ijk},$$

$$S_{jk}^2 = \frac{1}{N_{jk} - 1} \sum_{i=1}^{N_{jk}} \left(M_{ijk}p_{ijk} - \left(\sum_{i=1}^{N_{jk}} \frac{M_{ijk}p_{ijk}}{N_{jk}} \right) \right)^2,$$

$$S_{ijk}^2 = \frac{M_{ijk}}{M_{ijk} - 1} p_{ijk} \left(1 - p_{ijk}\right).$$

After a few algebraic manipulations, we obtain

$$\mathrm{Var}\ (y_k) = \frac{1}{n_{1k}} \left\{ N_{1k}^2 S_{1k}^2 + \left(\frac{1 - f_1}{f_1}\right) N_{1k} \sum_{i=1}^{N_{1k}} M_{i1k}S_{i1k}^2 + \frac{N_{2k}^2 S_{2k}^2}{\lambda_k} \right.$$

$$\left. + \left(\frac{1 - f_1}{f_1}\right) \frac{N_{2k}}{\lambda_k} \sum_{i=1}^{N_{2k}} M_{i2k}S_{i2k}^2 \right\} - \sum_{j=1}^{2} N_{jk}S_{jk}^2.$$

We can therefore write $CV^2(y_k)$ as

$$CV^2(y_k) = \frac{\mathrm{Var}\,(y_k)}{Y_k^2} = \frac{A_k}{n_{1k}} - B_k. \tag{3.1}$$

Furthermore, B (the relative bias) and $\hat{B}$ (its estimator) are given by

$$B = \frac{T_0}{T_1} = \frac{\sum_{i=1}^{N_0} M_{i0} p_{i0}}{\sum_{k=1}^{N_p} Y_k},$$

$$\hat{B} = \frac{t_0}{t_1} = \frac{\sum_{i=1}^{n_0} \frac{N_0}{n_0} M_{i0} p_{i0}}{\sum_{k=1}^{N_p} y_k},$$

where M_{i0} is the number of "no" persons in EA i in the province and p_{i0} is the probability of a characteristic of interest for a "no" person in EA i.

Assuming that t_0 and t_1 are independent, then

$$\mathrm{Var}\,(\hat{B}) = B^2 \left(\frac{\mathrm{Var}(t_0)}{T_0^2} + \frac{\mathrm{Var}\,(t_1)}{T_1^2} \right). \tag{3.2}$$

After a few algebraic manipulations, if f_0 is the sampling fraction in the "no" EAs, we obtain

$$\mathrm{Var}\,(t_0) = \frac{1}{n_0} \left\{ N_0^2 S^2 + N_0 \left(\frac{1 - f_0}{f_0} \right) \sum_{i=1}^{N_0} M_{i0} S_i^2 \right\} - N_0 S^2$$

where

$$S^2 = \frac{1}{N_0 - 1} \sum_{i=1}^{N_0} \left(M_{i0} p_{i0} - \left(\sum_{i=1}^{N_0} \frac{M_{i0} p_{i0}}{N_0} \right) \right)^2, \; S_i^2 = \frac{M_{i0}}{M_{i0} - 1} p_{i0} (1 - p_{i0})$$

which can be written in the form

$$\mathrm{Var}\,(t_0) = \frac{A_0}{n_0} - B_0. \tag{3.3}$$

Furthermore, assuming that the y_k's are independent, we have

$$\mathrm{Var}\,(t_1) = \sum_{k=1}^{N_p} \mathrm{Var}\,(y_k).$$

Using equation (3.1), this expression can be written as

$$\operatorname{Var}(t_1) = \sum_{k=1}^{N_p} \left(\frac{A_k Y_k^2}{n_{1k}} - B_k Y_k^2 \right). \tag{3.4}$$

From (3.2), (3.3) and (3.4), we obtain

$$\operatorname{Var}(\hat{B}) = B^2 \left\{ \left(\frac{A_0}{n_0 T_0^2} - \frac{B_0}{T_0^2} \right) + \sum_{k=1}^{N_p} \left(\frac{A_k Y_k^2}{n_{1k} T_1^2} - \frac{B_k Y_k^2}{T_1^2} \right) \right\}$$

$$= \frac{1}{n_0} \left(\frac{B^2 A_0}{T_0^2} \right) + \sum_{k=1}^{N_p} \frac{1}{n_{1k}} \left(\frac{B^2 A_k Y_k^2}{T_1^2} \right) - B^2 \left(\frac{B_0}{T_0^2} + \sum_{k=1}^{N_p} \frac{B_k Y_k^2}{T_1^2} \right)$$

$$= \frac{X}{n_0} + \sum_{k=1}^{N_p} \frac{W_k}{n_{1k}} - Z.$$

The optimization problem can be re-expressed as the problem of minimizing

$$\sum_{k=0}^{N_p} c_k n_k$$

subject to

$$0 < a_k \leq n_k \leq b_k \quad (k = 0, 1, 2, \ldots, N_p) \tag{3.5}$$

and

$$\sum_{k=0}^{N_p} d_k / n_k = e \tag{3.6}$$

where, for $k = 1, 2, \ldots, N_p$,

$$n_k = n_{1k}, \ c_k = c_{1k} + c_{2k} \lambda_k, \ a_k = \frac{A_k}{CV_*^2 + B_k}, \ b_k = \min(N_{1k}, N_{2k} / \lambda_k).$$

In practice, rather than using $b_k = \min(N_{1k}, N_{2k}/\lambda_k)$ we define

$$b_k = \frac{N_{1k} N_{2k} (1 + \lambda_k)}{\lambda_k^2 N_{1k} + N_{2k}},$$

then, if $n_k > N_{1k}$, sample sizes are given by $n_{1k} = N_{1k}$ and

$$n_{2k} = \lambda_k n_k + \frac{(n_k - N_{1k}) N_{2k}}{N_{1k} \lambda_k},$$

while, if $\lambda_k n_k > N_{2k}$ sample sizes are given by $n_{2k} = N_{2k}$ and

$$n_{1k} = n_k + \frac{(\lambda_k n_k - N_{2k}) N_{1k} \lambda_k}{N_{2k}}.$$

Thus, we consider $N_{2k} / (N_{1k}\lambda_k)$ small EAs to be equivalent to one large EA. On average, there are as many disabled persons in one large EA as in $N_{2k} / (N_{1k}\lambda_k)$ small EAs.

Proceeding in this way, it is not always true that $n_{2k} = \lambda_k n_{1k}$. However, we avoid CVs higher than target values, when, for example, small EAs remain to be observed (even if all the large EAs have been selected).

For some values of k, it is possible that $a_k \geq b_k$. If this is the case, we set $n_k = b_k$. Let

$$E_1 = \{k = 0, 1, 2, \ldots, N_p \mid n_k = a_k\},$$

$$E_2 = \{k = 0, 1, 2, \ldots, N_p \mid n_k = b_k > a_k\},$$

$$E_3 = \{k = 0, 1, 2, \ldots, N_p \mid a_k < n_k < b_k\},$$

$$E_4 = \{k = 0, 1, 2, \ldots, N_p \mid n_k = b_k \leq a_k\}.$$

The solution exists if

$$\sum_{k=0}^{N_p} d_k / b_k \leq e,$$

and it takes the form

$$n_k = \begin{cases} a_k & (k \in E_1) \\ b_k & (k \in E_2 \cup E_4) \\ K (d_k / c_k)^{1/2} & (k \in E_3) \end{cases} \tag{3.7}$$

where

$$K = \frac{\sum_{k \in E_3} (d_k / c_k)^{1/2}}{e - \sum_{k \in E_1} d_k / a_k - \sum_{k \in E_2 \cup E_4} d_k / b_k}, \tag{3.8}$$

since the n_k $(k \in E_3)$ minimize $\sum_{k \in E_3} c_k n_k$, subject to the constraint

$$\sum_{k \in E_3} d_k / n_k = e - \sum_{k \in E_1} d_k / a_k - \sum_{k \in E_2 \cup E_4} d_k / b_k.$$

What are the sets E_1, E_2, E_3 and E_4 corresponding to the solution? Set E_4 is easy to determine. We must have

$$a_k < (d_k/c_k)^{1/2}K < b_k \quad (k \epsilon E_3), \qquad (d_k/c_k)^{1/2}K \geq b_k \quad (k \epsilon E_2),$$

$$(d_k/c_k)^{1/2}K \leq a_k \quad (k \epsilon E_1). \tag{3.9}$$

Determining the sets involves trying each of the possibilities for E_1, E_2 and E_3 until a value for k which satisfies (3.9) is obtained. To reduce the number of possibilities to be examined, note that, if for $k' \geq k$,

$$b_k' (c_k'/d_k')^{1/2} \geq b_k (c_k/d_k)^{1/2} \quad (k, k' \epsilon \{0, 1, \ldots, N_p\}), \tag{3.10}$$

then there is a k^* such that $E_2 = \{0, 1, 2, \ldots, k^*\}$, or $E_2 = \{\ \}$, while, if for $k' \geq k$,

$$a_k' (c_k'/d_k')^{1/2} \geq a_k (c_k/d_k)^{1/2} \quad (k, k' \epsilon \{0, 1, \ldots, N_p\}), \tag{3.11}$$

then there is a k^{**} such that $E_1 = \{k^{**}, k^{**} + 1, \ldots N_p\}$ or $E_1 = \{\ \}$.

3.3 Parameter Estimation

To calculate the optimum sample allocation, the following quantities must be determined:

P_1 = proportion of HALS screened-in individuals who replied "yes" to Census question 20,

P_2 = proportion of HALS screened-out individuals who replied "yes" to Census question 20, and

P_3 = proportion of HALS screened-in individuals who replied "no" to Census question 20.

Since these parameters cannot be computed directly using data from the Canadian Health and Disability Survey, a test called the "calibration study" was carried out in September and October 1985.

Census question 20 was included, without abbreviation, as a supplementary question in the September Labour Force Survey (LFS). It was asked to a sample of approximately 36,000 individuals. The questions on the 17 ADLs and a question on mental handicaps were added as a supplement to the October LFS and were asked of the same individuals.

For each five-year age group, the weighted values from the calibration study were used to estimate the probability of an affirmative response, $P(yes)$, to Census question 20. The HALS screening questionnaire differs from that used in the calibration study. In HALS, there are more questions on mental and psychological problems and part (a) of Census question 20 is asked again. Therefore, we did not depend on the calibration study alone to calculate the parameters.

4. 1986 GEOGRAPHIC AND DEMOGRAPHIC FILE

4.1 Description of Available Information

When the sample allocation was done in the spring of 1986, the following information was available for use in calculation of population projections by age group and EA:

1. population projections by age group and province in 1986;
2. estimated population by age group and CD in 1984;
3. population by age group and EA in 1981;
4. conversion file to establish the correspondence between the 1981 and 1986 EAs;
5. estimated numbers of dwellings by EA in 1986.

The conversion file is structured according to the concept of equivalent sets. Each equivalent set is the smallest region consisting of EAs that has not had its boundaries altered. For example, if three 1981 EAs were reorganized as two 1986 EAs, the group of three 1981 EAs (or the group of two 1986 EAs) is an equivalent set.

The four methods described in the next subsection are designed to produce population projections by age group and by equivalent set in 1986. If an equivalent set is made up of several 1986 EAs, the projected population for the equivalent set can be divided proportionally among the EAs using the estimated numbers of dwellings by EA in 1986.

4.2 Estimation Methods

For province p, let

$ES_{l,k}$ = the l-th equivalent set of the k-th CD ($l = 1, 2, \ldots, N_k; k = 1, 2, \ldots, N_p$),

$ES_{l,k;81}(j)$ = population of $ES_{l,k}$ in the j-th age group in 1981 ($j = 1, 2, \ldots, 16$),

$CD_{k;84}(j)$ = estimated population of the k-th CD in the j-th age group in 1984,

$\hat{P}_{86}(j)$ = projected population in the j-th age group in the province in 1986.

For the three methods that follow, the first step is to calculate $\hat{CD}_{k;86}(j)$, the projected population of the j-th age group in the k-th CD in 1986. We assume there exists K_j' ($j = 1, 2, \ldots, 16$) such that

$$\hat{CD}_{k;86}(j) = K_j'(\hat{CD}_{k;84}(j)) \quad (k = 1, 2, \ldots, N_p; j = 1, 2, \ldots, 16),$$

$$\sum_{k=1}^{N_p} \hat{CD}_{k;86}(j) = \hat{P}_{86}(j) \quad (j = 1, 2, \ldots, 16).$$

This implies that

$$\hat{CD}_{k;86} = \frac{\hat{P}_{86}(j)\, CD_{k;84}(j)}{\sum_{k=1}^{N_p} CD_{k;84}(j)}.$$

The first method of estimating $ES_{l,k;86}(j)$ involves assuming the existence of K_j $(j = 1, \ldots, 16)$ such that

$$\hat{ES}_{l,k;86}(j) = K_j ES_{l,k;81}(j) \quad (l = 1, \ldots, N_k;\ j = 1, \ldots, 16),$$

$$\sum_{l=1}^{N_k} \hat{ES}_{l,k;86}(j) = \hat{CD}_{k;86}(j) \quad (j = 1, 2, \ldots, 16).$$

We will say that this method uses the simple model. We obtain

$$\hat{ES}_{l,k;86}(j) = \frac{\hat{CD}_{k;86}(j)\, ES_{l,k;81}(j)}{\sum_{l=1}^{N_k} ES_{l,k;81}(j)} \quad (l = 1, \ldots, N_k;\ j = 1, \ldots, 16).$$

With this simple model, the estimated total population of $ES_{l,k}$ in 1986 is

$$\sum_{j=1}^{16} \frac{\hat{CD}_{k;86}(j)\, ES_{l,k;81}(j)}{\sum_{l=1}^{N_k} ES_{l,k;81}(j)}.$$

If one thinks that a better estimate, $\hat{ES}_{l,k;86}(tot)$ of this quantity can be produced by independent means (for example, using the estimated number of dwellings in $ES_{l,k}$ in 1986), then more elaborate models can be used to estimate $ES_{l,k;86}(j)$. The multiplicative model is specified by the following equations:

$$\hat{ES}_{l,k;86}(j) = K_j(ES_{l,k;81}(j)) + e_l' \quad (l = 1, \ldots, N_k;\ j = 1, \ldots, 16),$$

$$\sum_{l=1}^{N_k} \hat{ES}_{l,k;86}(j) = K(\hat{CD}_{k;86}(j)) \quad (j = 1, \ldots, 16),$$

$$\sum_{l=1}^{N_k} e_l' = 0,$$

$$\sum_{j=1}^{16} \hat{ES}_{l,k;86}(j) = \hat{ES}_{l,k;86}(tot) \quad (l = 1, \ldots, N_k).$$

One can interpret e_l as the net intra-CD migration for the l-th equivalent set.

The third model, called the additive model, is given by the following equations:

$$\hat{ES}_{l,k;86}(j) = ES_{l,k;81}(j) + e_l + f_j \quad (l = 1, \ldots, N_k; j = 1, \ldots, 16),$$

$$\sum_{l=1}^{N_k} \hat{ES}_{l,k;86}(j) = \hat{CD}_{k;86}(j) + D \quad (j = 1, \ldots, 16),$$

$$\sum_{l=1}^{N_k} e_l = D,$$

$$\sum_{j=1}^{16} \hat{ES}_{l,k;86}(j) = \hat{ES}_{l,k;86}(tot) \quad (l = 1, \ldots, N_k).$$

This model involves the assumption that the population increases (or decreases) for each age group in each of the equivalent sets in a CD can be decomposed into two terms - one which depends only on the equivalent set and not on age (e_l), and one which depends only on age and not on the equivalent set (f_j).

A final trivial model involves simply formulating

$$\hat{ES}_{l,k;86}(j) = ES_{l,k;81}(j) \quad (l = 1, 2, \ldots, N_k; j = 1, \ldots, 16).$$

4.3 Evaluation of Estimation Methods

The four methods were evaluated using data for the period 1976-1981. We used the 1976 projection of the population by age group and province in 1981, ($\hat{P}_{81}(j)$), the population by age group and EA in 1976, a 1976-1981 conversion file and the pre-Census estimate of the number of dwellings per EA in 1981. Since there are no estimates for population by age group and CD in 1979 (the equivalent of $CD_{k;84}(j)$), we set

$$\hat{CD}_{k;81} = \frac{\hat{P}_{81}(j)\,\hat{CD}_{k;84}(j)}{\sum_{k=1}^{N_p} \hat{CD}_{k;84}(j)}.$$

For $\hat{ES}_{l,k;81}(tot)$, which is needed for the multiplicative and additive models, we used

$$\hat{ES}_{l,k;81}(tot) = \frac{\sum_{j=1}^{16} ES_{l,k;76}(j) \quad \sum_{j=1}^{16} \hat{CD}_{k;81}(j)}{\sum_{l=1}^{N_k} \sum_{j=1}^{16} ES_{l,k;76}(j)}.$$

Table 1
Comparison of the Four Methods

Prov.	EFF_S	EFF_M	EFF_A
Nfld.	0.890	0.891	0.887
P.E.I.	0.903	0.914	0.919
N.S.	0.960	0.972	0.912
N.B.	0.870	0.868	0.884
Que.	0.778	0.764	0.818
Ont.	0.932	0.930	0.916
Man.	0.892	0.904	0.912
Sask.	0.732	0.749	0.801
Alta.	0.818	0.827	0.860
B.C.	0.713	0.716	0.775
Yukon	0.770	0.768	0.840
N.W.T.	1.252	1.246	1.157

For each province p, an efficiency measure was calculated for the simple, multiplicative and additive models relative to the trivial model:

$$EFF_m = \frac{\sum_{k=1}^{N_p} \sum_{l=1}^{N_k} \sum_{j=1}^{16} \left((\hat{ES}^{(m)}_{l,k;81}(j) - ES_{l,k;81}(j) \right)^2}{\sum_{k=1}^{N_p} \sum_{l=1}^{N_k} \sum_{j=1}^{16} \left((\hat{ES}^{(T)}_{l,k;81}(j) - ES_{l,k;81}(j) \right)^2} \quad (m = S, M, A),$$

where $\hat{ES}^{(m)}_{l,k;81}(j)$ with $m = S, M, A$ and T are the projections obtained by means of the simple, multiplicative, additive and trivial models respectively. Some values obtained are given in Table 1.

The simple model gives the worst results for one province and one territory, the multiplicative model for two provinces and the additive model for seven provinces and one territory.

The simple model is the best for five provinces, while the multiplicative model is best for two provinces and one territory and the additive model is best for three provinces and one territory.

Since the simple model also has, as its name implies, the advantage of simplicity, it is the one that was chosen.

4.4 Method of Stratification by Enumeration Area Size

If simple random sampling were used to select EAs within each subprovincial area (SPA), disabled persons belonging to an EA with many disabled residents would have less chance of being selected than those in a small EA - that is, an EA with few disabled persons. To avoid excessive differences in selection probabilities, the population of EAs in each SPA is stratified according to the number of disabled persons in the EAs, and then proportional allocation is used. With proportional allocation, the number of EAs selected is proportional to the number of disabled persons for each stratum.

Using the results of earlier surveys, a link was established between the age distribution of the population of an EA and the number of disabled persons expected in the EA. Since the number of disabled persons is unknown, the variable used for stratification and sample allocation is the expected number of disabled persons.

In the case under consideration here, there are only two strata - one for large EAs and one for small EAs. Since proportional allocation is being used, we employed a criterion found in Raj (1968) to determine the optimum dividing line between large and small EAs. This criterion gives the optimum dividing line as the average of the average size of the small EAs and the average size of the large EAs.

5. SAMPLE SELECTION

It was necessary to draw samples for the three populations (children, adults and seniors) among the large and small "yes" EAs of each SPA, both for frame three and for frame five, and among the "no" EAs of each province for frame five. When an SPA contained fewer than two large EAs or fewer than two small EAs, we selected all of the EAs in that SPA for the three populations. The "yes" and "no" samples were created independently, using the one-pass algorithm described by Bebbington (1975). The samples from the three populations for the "yes" and "no" components were nested to minimize the total number of EAs selected.

The following table shows the sizes obtained for the samples by province for each age group.

Table 2

Sample Sizes by Province and Age Group

Province	Children		Adults		Seniors	
	Number of "yes" EAs selected	Number of "no" EAs selected	Number of "yes" EAs selected	Number of "no" EAs selected	Number of "yes" EAs selected	Number of "no" EAs selected
Nfld.	880	136	405	154	476	173
P.E.I.	242	242	111	217	82	166
N.S.	1257	157	434	130	438	115
N.B.	1142	162	459	146	453	138
Que.	4749	153	1070	114	1488	133
Ont.	6085	158	1304	116	1542	120
Man.	1082	203	457	169	367	144
Sask.	2291	265	942	241	921	193
Alta.	2762	190	909	176	1389	222
B.C.	3117	170	752	125	948	119

6. DISCUSSION

The postcensal survey is a relatively new survey method that will no doubt undergo extensive development in the next few years. This type of survey allows for a great deal of flexibility in data collection and use of large samples scattered throughout the country, with reasonable costs and timeliness. The Health and Activity Limitation Survey is the first postcensal survey of its size in Canada.

The sample design presented in this article is an attempt to maximize use of the opportunities offered by the postcensal approach, with optimum use of the available resources. One of the major problems inherent in the proposed method is control of sample size. Sample allocation is determined before the census is taken; this means that all calculations must be done using projections based on the previous census. In this context, the actual size of a sample made up of a group of small areas selected on the basis of the projection results may vary considerably from its expected size.

Therefore, on the one hand, one may obtain a sample that is inadequate with respect to the quality requirements for the estimates. On the other hand, the resources allocated to data collection may be exceeded. In order to prevent these problems, we implemented the following strategy. A target number of interviews for each population was calculated for the "yes" sample. This number was based on the sample size required to produce estimates that would satisfy our quality criteria. However, for the reasons mentioned above, we selected more EAs than were necessary to obtain the target number of interviews. For reasons of cost, if the real number of interviews to be conducted, as calculated in the field, was higher than the target number, a subsample of EAs were excluded from the survey. Only for the Halifax Regional Office (covering Prince Edward Island, Nova Scotia and New Brunswick) was the number of interviews in the "yes" sample substantially higher than the target number. The decision was therefore made to exclude certain EAs from this part of the sample. In order to know which EAs would be excluded, it was necessary to know the target number and the real number of interviews for each EA. For 40 per cent of the EAs, the real number of interviews had to be imputed since this information was not available in time.

For this imputation, the total real number of interviews was known for each census commissioner district. The portion of this total not already allocated to EAs with known numbers of interviews was distributed among the EAs requiring imputation, in proportion to the target number of interviews.

We then calculated, for each population, the difference between the real number and the target number of interviews for each of the two strata of each SPA. A positive difference (real-target) indicated a population for which some EAs could be excluded from the survey. In each stratum, the EAs were divided into three groups (1, 2 and 3), in accordance with whether they had been selected for three, two or only one of the populations respectively. The EA file was then sorted by stratum and by group in ascending order, with the order of the EAs within each group being random. Each EA was considered successively and was suppressed for the three populations if:

1) a positive difference remained non-negative after suppression of the EA;

2) a negative difference was not further reduced.

In this way, each positive difference was reduced to a number as close as possible to zero, considering the random order of the EAs.

ACKNOWLEDGEMENTS

The authors would like to thank D.A. Binder for his contribution with regard to the sample design and P. Reed for the computer work involved in the sample allocation. The authors would also like to thank the referees for their useful comments.

APPENDIX

Question 20 of Census Form 2B

20. a) Are you limited in the kind or amount of activity that you can do because of a long-term physical condition, mental condition or health problem: (See Guide)

At home?

□ No, I am not limited

□ Yes, I am limited

At school or at work?

□ No, I am not limited

□ Yes, I am limited

□ Not applicable

In other activities, e.g., transportation to or from work, leisure time activities?

□ No, I am not limited

□ Yes, I am limited

b) Do you have any long-term disabilities or handicaps?

□ No

□ Yes

Screening Questions for HALS (Questionnaire for Adults)

1. Do you have any trouble hearing what is said in a normal conversation with one other person?
2. Do you have any trouble hearing what is said in a group conversation with at least three other people?
4. Do you have any trouble reading ordinary newsprint, with glasses if normally worn?
5. Do you have any trouble seeing clearly the face of someone from 12 feet/4 metres (example: across a room), with glasses if normally worn?
7. Do you have any trouble speaking and being understood?
8. Do you have any trouble walking 400 yards/400 metres without resting (about three city blocks)?
9. Do you have any trouble walking up and down a flight of stairs (about 12 steps)?
10. Do you have any trouble carrying an object of 10 pounds for 30 feet/5 kg for 10 metres (example: carrying a bag of groceries)?
11. Do you have any trouble moving from one room to another?
12. Do you have any trouble standing for long periods of time, that is, more than 20 minutes? Remember, I am asking about problems expected to last 6 months or more.
13. When standing do you have any trouble bending down and picking up an object from the floor (example: a shoe)?
14. Do you have any trouble dressing and undressing yourself?
15. Do you have any trouble getting in and out of bed?
16. Do you have any trouble cutting your own toenails?

17. Do you have any trouble using your fingers to grasp or handle?
18. Do you have any trouble reaching in any direction (example: above your head)?
19. Do you have any trouble cutting your own food?
20. Because of a long-term physical condition or health problem, that is, one that is expected to last 6 months or more, are you limited in the kind or amount of activity you can do ...

    (i) at home? (ii) at school or at work? (iii) in other activities such as travel, sports, or leisure?
21. Has a school or health professional ever told you that you have a learning disability?
22. From time to time, everyone has trouble remembering the name of a familiar person, or learning something new, or they experience moments of confusion. However, do you have any ongoing problems with your ability to remember or learn?
23. Because of a long-term emotional, psychological, nervous, or mental health condition or problem, are you limited in the kind or amount of activity you can do?

    (i) at home? (ii) at school or at work? (iii) in other activities such as travel, sports, or leisure?

REFERENCES

BEBBINGTON, A.C. (1975). A simple method of drawing a sample without replacement. *Applied Statistics*, 24, 136.

CARTER, R.G., GILES, P.D., and SHERIDAN, M.J. (1982). Description and rationale for the screen tests for the January 1983 Disability Survey. Disability Data Development Project, Health Division, Statistics Canada.

HOUSE OF COMMONS (1981). Obstacles, Report of the special committee on the disabled and handicapped. Ottawa.

COCHRAN, W.G. (1977). *Sampling Techniques*, (3rd ed.). New York: John Wiley.

DOLSON, D., GILES, P., and MORIN, J.-P. (1984). A Methodology for surveying disabled persons using a supplement to the Labour Force Survey. *Survey Methodology*, 10, 187-197.

DOLSON, D., McCLEAN, K., MORIN, J.-P., and THÉBERGE, A. (1986). Methodology report of HALS. Working Paper, Statistics Canada.

GRABOWIECKI, F. (1982). Discussion of the target population for the Disability Survey. Disability Data Development Project, Health Division, Statistics Canada.

GRABOWIECKI, F. (1983). Content of Statistics Canada's Disability Survey. Technical Report, Health Division, Statistics Canada.

LAZARUS, G., and NESICH, R. (1985). A report on the methodology of the Canadian Health and Disability Survey. Working Paper, Statistics Canada.

McDOWELL, I. (1981). An examination of the OECD survey questions in a Canadian Study. *Revue d'épidémiologie et de santé publique*, 29, 412-429.

MORIN, J.-P., and DOWLER, L. (1986). Proposition d'une méthodologie pour l'ESLA-institutions. Working Paper, Statistics Canada.

MORIN, J.-P. (1986). Comparaison initiale de l'ESIC et de l'ESG. Working Paper, Statistics Canada.

WORLD HEALTH ORGANIZATION (1980). International classification of impairments, disabilities and handicaps. Geneva, Switzerland.

RAJ, D. (1968). *Sampling Theory*. New York: McGraw-Hill.

WILSON, R.W., and McNEIL, J.M. (1981). Preliminary analysis of OECD disability on the pretest of the Post-censal Disability Survey. *Revue d'épidémiologie et de santé publique*, 29, 469-475.

Survey Methodology, June 1987
Vol. 13, No. 1, pp. 109-121
Statistics Canada

Comparison of Estimators of Population Total in Two-Stage Successive Sampling Using Auxiliary Information

F.C. OKAFOR[1]

ABSTRACT

Singh and Srivastava (1973) proposed a linear unbiased estimator of the population mean when sampling on successive occasions using several auxiliary variables whose known population means remain unchanged for all occasions. In this paper, three composite estimators T_1, T_2 and T_3, each utilising an auxiliary variable whose known population mean changes from one occasion to the next, are presented for the estimation of the current population total. The proposed estimators are compared with the ordinary estimator, T_0, and the usual successive sampling estimator, T', of the current population total without the use of auxiliary information. We find that using auxiliary information in conjunction with successive sampling does not always uniformly produce a gain in efficiency over T_0 or T'. However, when applied to a survey of teak plantations to estimate the mean height of teak trees, T_1, T_2 and T_3 proved more efficient than T_0 and T'.

KEY WORDS: Successive occasion; Partial matching; Auxiliary variate.

1. INTRODUCTION

The theory and practice of surveying the same population at different points in time - technically called repetitive sampling or sampling over successive occasions - have been given considerable attention by some survey statisticians. The main objective of sampling on successive occasions is to estimate some population parameters (total, mean, ratio, etc) for the most recent occasion as well as changes in these parameters from one occasion to the next.

The theory of successive sampling was initiated by Jessen (1942). Many authors have since contributed, especially in the estimation of population means. Among them are Singh (1968), Abraham et al (1969), Kathuria and Singh (1971), and Kathuria (1975), to mention but a few.

Singh (1968) was the first to extend the theory of unistage sampling to two-stage sampling on successive occasions. He considered the sampling scheme in which, on the second occasion, a fraction λ of the first stage units (FSUs) selected on the previous occasion is retained, along with their selected second stage units (SSUs), and a fraction μ ($\lambda + \mu = 1$) selected afresh. He then obtained a minimum variance unbiased estimator of the population mean on the current occasion.

Abraham et al (1969) considered the situation in which partial matching of units was carried out at both stages. Units were selected by simple random sampling without replacement (SRSWOR). Kathuria (1975) modified this by using probability proportional to size and with replacement (PPSWR) for selection of the FSUs, and proposed a linear composite estimator for the population mean on the current occasion.

[1] F.C. Okafor, Department of Statistics, University of Ibadan, Ibadan, Nigeria.

When an auxiliary variable is highly correlated with the characteristic under study, the estimate of the population mean (total) of this characteristic can be improved using the auxiliary variable. Singh and Srivastava (1973) used auxiliary information to improve on the estimator of Singh (1968). They obtained a linear unbiased estimator of the population mean on the most recent occasion using several auxiliary variables whose population means are known and are the same for all occasions. Kathuria (1978) developed this study further by assuming that the population mean of the auxiliary variate is not known. He used a double sampling technique to estimate first the population mean of the auxiliary variate and then the mean of the characteristic under study.

In their contributions, Singh and Srivastava (1973) and Kathuria (1978) assumed that the necessary information on the auxiliary variables can be obtained from the respondents or reporting units (SSUs). This is not generally the case. It may happen that the information on the auxiliary variable is too distorted to be useful because of the sensitive nature of the question, or the respondents may refuse outright to supply any information. Alternatively, the information on the auxiliary variate may not be collected because the required question is not included in the questionnaire.

Singh and Srivastava also assumed that the known population total of the auxiliary variable is the same for all occasions. This may not be true in practice. If the population total of the main characteristic changes from one occasion to the next, there is every likelihood that the population total of any other variable correlated with it will also vary.

In this paper three composite estimators of the population total using auxiliary information and a two-stage successive sampling scheme are proposed. The performances of the three estimators are compared empirically and they are also applied to a survey of teak plantations to estimate the mean height of teak trees.

2. SAMPLING FOR TWO OCCASIONS

For all three proposed estimators, we assume that the population total of the auxiliary variable changes on the second occasion.

The estimators of the population total (mean) based on the partial matching scheme are better than the ordinary estimators of the population total (mean) without partial matching. Therefore, it is expected that the proposed estimators T_1, T_2 and T_3 will perform better than the ordinary population total estimator, T_o, and the estimator based on the partial matching scheme without the use of auxiliary information, T'.

In deriving these estimators, we assume that:

(i) the sample size is constant on each occasion;
(ii) the normed size measure P_i for the i^{th} first stage unit (FSU) is fixed for each occasion;
(iii) N and M_i, population sizes for the FSUs and the second stage units (SSUs) within the i^{th} FSU respectively, are constant for the two occasions;
(iv) the population total (mean) of the auxiliary variate is known.

Assumptions (i) – (iii) apply to T', T_1, T_2 and T_3; (iv) applies to T_1, T_2 and T_3, but not to T' and T_o.

On the first occasion, a sample S_1 of n FSUs is selected with probability proportional to size and with replacement (PPSWR) using P_i as normed size measure for the i^{th} $(i = 1, 2, \ldots, N)$ unit. For the selection of SSUs, we adopt the method due to Cochran

(1977, p. 306), which stipulates that if the i^{th} FSU in S_1 is drawn θ_i times ($i = 1, 2, \ldots, n$), we select θ_i independent subsamples of size m_i from the M_i SSUs.

On the second occasion, we select a sample of λn ($0 < \lambda < 1$) FSUs from S_1 by simple random sampling without replacement (SRSWOR). The SSUs selected on the first occasion are retained for each of these λn matched FSUs. Then, a fresh sample of μn ($\mu = 1 - \lambda$) FSUs is selected independently from the N FSUs by PPSWR, with P_i as normed size measure for the i^{th} FSU. In each of the μn FSUs, the SSUs are selected as on the first occasion.

3. NOTATION

We define y_{ij} (x_{ij}) as the value of the study variate for the j^{th} SSU in the i^{th} FSU on the current (previous) occasion. In addition, z_{hij} is defined as the value of the auxiliary variate for the j^{th} SSU in the i^{th} FSU on the h^{th} occasion ($h = 1, 2$). The sample means for SSUs in the i^{th} FSU are

$$\bar{x}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} x_{ij}, \; \bar{y}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij} \text{ and } \bar{z}_{hi} = \frac{1}{m_i} \sum_{j=1}^{m_i} z_{hij}.$$

The population total for the i^{th} FSU and the overall population total for the auxiliary variate are

$$Z_{hi} = \sum_{j=1}^{M_i} z_{hij} \text{ and } Z_h = \sum_{j=1}^{N} Z_{hi}.$$

We define additional notation as follows:

$S_b^2\,(y) = \sum_{i=1}^{N} P_i \left(\frac{Y_i}{P_i} - Y\right)^2$ is the between - FSU variance;

$S_w^2\,(y) = \sum_{i=1}^{N} \frac{M_i^2}{P_i} \left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_{wi}^2\,(y)$ is the variance among SSUs within the FSUs;

$S_{wi}^2\,(y) = \frac{1}{M_i - 1} \sum_{j=1}^{M_i} (y_{ij} - \bar{y}_i)^2$ is the variance among the SSUs in the i^{th} FSU;

$S^2\,(y) = S_b^2\,(y) + S_w^2\,(y);$

$C_b\,(x,y) = \rho_b S_b(x)\, S_b(y)$ is the between-FSU covariance of x and y;

$C_w(x,y) = \rho_w S_w(x)\, S_w(y)$ is the covariance of x and y among SSUs within the FSUs;

$C(x,y) = C_b(x,y) + C_w(x,y).$

The between- and within-FSU correlation coefficients between x and y are respectively ρ_b and ρ_w.

4. ESTIMATORS FOR THE POPULATION TOTAL AND THEIR OPTIMUM VARIANCES

4.1 Case (i)

The first estimator of the population total, Y, on the second occasion is used when information on the auxiliary variable is not available but the FSU population total of the auxiliary variable is available for the selected FSUs. It is given as

$$T_1 = \theta(1)\, T_m(1) + (1 - \theta(1))\, T_u(1) \tag{4.1}$$

$\theta(1)$ is a constant chosen so that the variance of T_1, $V(T_1)$, attains a minimum; while

$$\begin{aligned} T_m(1) = {} & \frac{1}{\lambda n} \sum_{i=1}^{\lambda n} \left\{ \frac{M_i \bar{y}_i}{P_i} - k(1) \left(\frac{Z_{2i}}{P_i} - Z_2 \right) \right\} \\ & - b(1) \left[\frac{1}{\lambda n} \sum_{i=1}^{\lambda n} \left\{ \frac{M_i \bar{x}_i}{P_i} - k(1) \left(\frac{Z_{1i}}{P_i} - Z_1 \right) \right\} \right. \\ & \left. - \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{M_i \bar{x}_i}{P_i} - k(1) \left(\frac{Z_{1i}}{P_i} - Z_1 \right) \right\} \right] \end{aligned}$$

is the difference estimator of Y based on the matched sample;

$$T_u(1) = \frac{1}{n\mu} \sum_{i=1}^{n\mu} \left\{ \frac{M_i \bar{y}_i}{P_i} - k(1) \left(\frac{Z_{2i}}{P_i} - Z_2 \right) \right\}$$

is the estimator for Y based on the unmatched sample; and $k(1)$ and $b(1)$ are known constants.

For this estimator, it is assumed that the population total of the auxiliary variate, Z_i, is available for each selected FSU on each occasion. The overall population total, Z, is also available on each occasion. No additional information on the auxiliary variate is obtained from the respondents (SSUs).

Now by minimizing $V(T_1)$ with respect to $\theta(1)$ and solving, the optimum value of $\theta(1)$ becomes

$$\theta_0(1) = \lambda A_2(1)/\Delta(1)$$

where

$$A_2(1) = S^2(y) + k^2(1)\, S_b^2(z_2) - 2k(1)\, C_b(z_2, y),$$

$$\Delta(1) = A_2(1) + \mu^2 \{ b^2(1)\, A_1(1) - 2b(1)\beta(1) \}.$$

The optimum value of $k(1)$ is obtained by minimizing $V(T_u(1))$ with respect to $k(1)$. This gives $k_0(1) = C_b(z_2, y) / S_b^2(z_2)$.

It can be shown that the optimum $V(T_1)$ for a given λ, following the method adopted by Jessen (1942), is

$$V_0(T_1) = \frac{1}{n}\,[A_2(1) + \mu\{b^2(1)A_1(1) - 2b(1)\,\beta(1)\}]\,A_2(1)/\Delta(1) \tag{4.2}$$

where

$$A_1(1) = S^2(x) + k^2(1)\,S_b^2(z_1) - 2k(1)\,C_b(z_1,x),$$

$$\beta(1) = C(x,y) + k^2(1)\,C_b(z_1,z_2) - k(1)\{C_b(x,z_2) + C_b(z_1,y)\},$$

$$\Delta(1) = A_2(1) + \mu^2\{b^2(1)\,A_1(1) - 2b(1)\,\beta(1)\}.$$

Minimizing the variance of $T_m(1)$, the optimum $b(1)$ is

$$b_0(1) = \beta(1)/A_1(1).$$

If $b_0(1)$ is substituted in (4.2), the optimum variance becomes

$$V_0(T_1) = \frac{1}{n}\left[\frac{A_1(1)\,A_2(1) - \mu\beta^2(1)}{A_1(1)\,A_2(1) - \mu^2\beta^2(1)}\right]A_2(1). \tag{4.3}$$

By minimizing $V_o(T_1)$ in (4.2) with respect to μ, the optimum matching fraction boils down to $\lambda_0 = 1 - \mu_0$ where

$$\mu_0 = A_2(1)\,[A_2(1) + \{A_2^2(1) + A_2(1)\,(b^2(1)A_1(1) - 2b(1)\beta(1))\}^{1/2}]^{-1}. \tag{4.4}$$

If $A_2(1) = A_1(1)$, i.e. the population variability is the same on both occasions, the expression in (4.3) yields

$$V_0(T_1) = \frac{1}{n}\left[\frac{A^2(1) - \mu\beta^2(1)}{A^2(1) - \mu^2\beta^2(1)}\right]A(1) \tag{4.5}$$

while the optimum matching fraction, μ_0 (given in (4.4)), with $b_0(1)$ substitued for $b(1)$ becomes

$$\mu_0 = A(1)\,[A(1) + \{A^2(1) - \beta^2(1)\}^{1/2}]^{-1}. \tag{4.6}$$

When μ_0 is substituted in (4.5) the variance works out as

$$V_0(T_1) = \frac{1}{2n}\,[A(1) + \{A^2(1) - \beta^2(1)\}^{1/2}]. \tag{4.7}$$

4.2 Case (ii)

The second estimator is the usual one in which information is obtained on both the main and auxiliary characteristic from the reporting units and the population total of the auxiliary characteristic is known.

It is written as

$$T_2 = \theta(2)\, T_m(2) + (1 - \theta(2))\, T_u(2), \tag{4.8}$$

where

$$T_m(2) = \frac{1}{\lambda n} \sum_{i=1}^{\lambda n} \left\{\frac{M_i \bar{y}_i}{P_i} - k(2) \left(\frac{M_i \bar{z}_{2i}}{P_i} - Z_2\right)\right.$$
$$- b(2) \left[\frac{1}{\lambda n} \sum_{i=1}^{\lambda n} \left\{\frac{M_i \bar{x}_i}{P_i} - k(2) \left(\frac{M_i \bar{z}_{1i}}{P_i} - Z_1\right)\right\}\right.$$
$$\left.\left. - \frac{1}{n} \sum_{i=1}^{n} \left\{\frac{M_i \bar{x}_i}{P_i} - k(2) \left(\frac{M_i \bar{z}_{1i}}{P_i} - Z_1\right)\right\}\right]\right.,$$

and

$$T_u(2) = \frac{1}{n\mu} \sum_{i=1}^{n\mu} \left\{\frac{M_i \bar{y}_i}{P_i} - k(2) \left(\frac{M_i \bar{z}_{2i}}{P_i} - Z_2\right)\right\}.$$

Here the overall population total of the auxiliary variate is known on both occasions. In addition, information on the auxiliary variate, z_{ij}, is obtained for every SSU in the sample. This is the usual way of using the auxiliary information in a two- stage design described in the literature. It can be shown that the optimum variance of T_2 is

$$V_0(T_2) = \frac{1}{n} [A_2(2) + \mu\{b^2(2) A_1(2) - 2b(2)\beta(2)\}] A_2(2) / \Delta(2) \tag{4.9}$$

and the optimum weight is

$$\theta_0(2) = \lambda A_2(2) / \Delta(2)$$

where

$$A_2(2) = S^2(y) + k^2(2)\, S^2(z_2) - 2k(2)\, C(z_2,y),$$

$$A_1(2) = S^2(x) + k^2(2)\, S^2(z_1) - 2k(2)\, C(z_1,x),$$

$$\beta(2) = C(x,y) + k^2(2)\, C(z_1,z_2) - k(2)\, \{C(z_1,y) + C(x, z_2)\},$$

$$\Delta(2) = A_2(2) + \mu^2\, \{b^2(2)\, A_1(2) - 2b(2)\, \beta(2)\}.$$

The optimum value of $k(2)$ is $k_0(2) = C(z_2,y)/S_2(z_2)$.

By substituting the optimum regression coefficient $b_0(2) = \beta(2)/A_1(2)$, obtained by minimizing the variance of $T_m(2)$, in (4.9) and assuming that $A_2(2) = A_1(2) = A(2)$ we have

$$V_0(T_2) = \frac{1}{n}\left[\frac{A^2(2) - \mu\beta^2(2)}{A^2(2) - \mu^2\beta^2(2)}\right] A(2). \tag{4.10}$$

If the optimum μ is substituted in (4.10), the variance becomes

$$V_0(T_2) = \frac{1}{2n}\,[A(2) + \{A^2(2) - \beta^2(2)\}^{1/2}]. \tag{4.11}$$

4.3 Case (iii)

The third way of utilising available auxiliary information to improve the estimate of the current population total, Y, under the given sampling scheme is similar to the second. The only difference is that the population total of the auxiliary characteristic is not known; however, its FSU population mean is known for the selected FSUs.

This is given as

$$T_3 = \theta(3)\,T_m(3) + (1 - \theta(3))\,T_u(3), \tag{4.12}$$

where

$$\begin{aligned} T_m(3) = {} & \frac{1}{\lambda n}\sum_{i=1}^{\lambda n}\frac{M_i}{P_i}\{\bar{y}_i - k(3)\,(\bar{z}_{2i} - \bar{Z}_{2i})\} \\ & - b(3)\left[\frac{1}{\lambda n}\sum_{i=1}^{\lambda n}\frac{M_i}{P_i}\,[\bar{x}_i - k(3)\,(\bar{z}_{li} - \bar{Z}_{li})\} \right. \\ & \left. - \frac{1}{n}\sum_{i=1}^{n}\frac{M_i}{P_i}\,[\bar{x}_i - k(3)\,(\bar{z}_{1i} - \bar{Z}_{1i})]\right], \end{aligned}$$

and

$$T_u(3) = \frac{1}{n\mu}\sum_{i=1}^{n\mu}\frac{M_i}{P_i}\{\bar{y}_i - k(3)\,(\bar{z}_{2i} - \bar{Z}_{2i})\}.$$

For this estimator, we suppose that the values of both the main variate and the auxiliary variate are obtained for every SSU in the sample on both occasions. We also assume that the population mean, $\bar{Z}_i$, of the auxiliary variate is known for the selected FSUs.

The optimum variance of T_3 for a given λ is given as

$$V_0(T_3) = \frac{1}{n}\,[A_2(3) + \mu\{b^2(3)A_1(3) - 2b(3)\beta(3)]]\,A_2(3)/\Delta(3) \tag{4.13}$$

while the optimum weight is as usual obtained as

$$\theta_0(3) = \lambda A_2(3)/\Delta(3),$$

where

$$A_2(3) = S^2(y) + k^2(3)\ S_w^2(z_2) - 2k(3)\ C_w(z_2,y),$$

$$A_1(3) = S^2(x) + k^2(3)\ S_w^2(z_1) - 2k(3)\ C_w(z_1,x),$$

$$\beta(3) = C(x,y) + k^2(3)\ C_w(z_1,z_2) - k(3)\ \{C_w(z_1,y) + C_w(z_2,x)\},$$

$$\Delta(3) = A_2(3) + \mu^2\ \{b^2(3)A_1(3) - 2b(3)\ \beta(3)\}.$$

The optimum value of $k(3)$ is $k_0(3) = C_w(z_2,y)/S_w^2(z_2)$.

If the optimum regression coefficient is substituted in (4.13), and it is assumed that population variances are the same on both occasions, then (4.13) works out as

$$V_0(T_3) = \frac{1}{n}\left[\frac{A^2(3) - \mu\beta^2(3)}{A^2(3) - \mu^2\beta^2(3)}\right] A(3). \tag{4.14}$$

When the optimum μ is substituted in (4.14), the variance is

$$V_0(T_3) = \frac{1}{2n}\ [A(3) + \{A^2(3) - \beta^2(3)\}^{½}]. \tag{4.15}$$

4.4 Efficiency of the Proposed Estimators

The variances given in (4.7), (4.11) and (4.15) will be used to compare the efficiencies of T_1, T_2 and T_3 with respect to

$$T_0 = \frac{1}{n}\sum_{i=1}^{n}\frac{M_i\bar{y}_i}{P_i}.$$

T_0 is the estimator for y when there is no partial matching of units and no auxiliary information used. In addition, the efficiency of T_0 compared to the usual partial matching estimator T', which uses no auxiliary information, will be presented to assist in understanding the performance of the proposed estimators.

The usual partial matching estimator is defined as

$$T' = \theta' T_m' + (1 - \theta')\,T_u', \tag{4.16}$$

where

$$T'_m = \frac{1}{\lambda n} \sum_{i=1}^{\lambda n} \frac{M_i \bar{y}_i}{P_i} - b' \left\{ \frac{1}{\lambda n} \sum_{i=1}^{\lambda n} \frac{M_i \bar{x}_i}{P_i} - \frac{1}{n} \sum_{i=1}^{n} \frac{M_i \bar{x}_i}{P_i} \right\},$$

and

$$T'_u = \frac{1}{n\mu} \sum_{i=1}^{n\mu} \frac{M_i \bar{y}_i}{P_i}.$$

The optimum variance of T', obtained using the optimum value of b', $b'_0 = C(x,y) / S^2(x)$, and assuming $S^2(y) = S^2(x)$ is

$$V_0(T') = \frac{1}{n} \left[\frac{S^2(y) - \mu C(x,y)}{S^2(y) - \mu^2 C(x,y)} \right] S^2(y). \tag{4.17}$$

Substituting the optimum value of μ in (4.17), the variance of T' becomes

$$V_0(T') = \frac{1}{2n} [S^2(y) + \{S^4(y) - C^2(x,y)\}^{1/2}]. \tag{4.18}$$

To calculate the efficiencies, the following assumptions about the correlation coefficients and the constant k were made:

$$\rho_b(x,z_2) = \rho_b(z_1,y) = \rho_b(z_1,z_2) = \rho_b;$$

$$\rho_w(x,z_2) = \rho_w(z_1,y) = \rho_w(z_1,z_2) = \rho_w;$$

$$k(1) = k(2) = k(3) = 1.$$

The efficiencies have been presented for only the positive values of ρ_b and ρ_w, and a set of values of

$$\delta = S^2_w(y) / S^2_b(y), \; R_b = S^2_b(z) / S^2_b(y) \text{ and } R_w = S^2_w(z) / S^2_b(y).$$

Looking at Table 2, we observe that none of the strategies T_1, T_2 and T_3 (sampling design and estimator) is uniformly more efficient than strategy T_0. The contrary is true of T', which is always more efficient than T_0; at worst, its gain over T_0 is small (see Table 1).

The results in Tables 1 and 2 show T_1 is to be preferred to T' only when $R_b = 0.05$; and when $\rho_b = 0.8$ and $R_b = 0.5$.

Table 1

The Efficiency of T' with Respect to T_0

ρ_b	δ	$\rho_w = 0.2$	$\rho_w = 0.8$
0.2	0.05	1.01	1.01
	0.5	1.01	1.04
	5.0	1.01	1.17
	0.05	1.22	1.25
0.8	0.5	1.11	1.25
	5.0	1.02	1.25

T_2 is better than T' when:

(i) $\rho_w = 0.2,\ R_b = R_w = 0.05;$

(ii) $\rho_b = \rho_w = 0.8,\ R_b = R_w = 0.05,\ 0.5;$

(iii) $\delta = 0.5,\ 5.0,\ R_w = R_b = 0.05,\ \rho 0.5,\ \rho_b = 0.2$ and $\rho_w = 0.8$.

T_3 is generally more efficient than T' when:

(i) $\delta = 5.0,\ \rho_w = 0.8;$

(ii) $\delta = 0.5,\ \rho_w = 0.8$ and $R_w = 0.05,\ 0.5$.

The maximum gain in efficiency of T' over T_0 is 25% (see Table 1). In Table 2, the maximum gain of T_1 over T_0 is 155%, which occurs when $\rho_b = \rho_w = 0.8, \delta = 0.05,\ R_b = 0.5$. The maximum gain in efficiency of T_2 over T_0 is 172%; this happens when $\rho_b = \rho_w = 0.8, \delta = R_w = 0.05$. We also observe that when $\rho_b = \rho_w = 0.8, \delta = R_w = 5.0$, the maximum gain of T_3 over T_0 is 104%. It is therefore evident that the use of an auxiliary variate has tremendously improved the efficiency of partial matching of units.

If we now take the three strategies T_1, T_2 and T_3, and compare them among themselves, we conclude that none of the strategies is uniformly better than the other, even though the maximum gain in efficiency of T_2 over T_0 is higher than that of T_1, which in turn is higher than the maximum gain of T_3. In general T_1 is superior to T_2 when $\rho_w = 0.2$, while T_2 is better than T_1 when $\rho_w = 0.8$. T_1 is preferred to T_3 when $\rho_b = 0.8,\ \rho_w = 0.2$ and $R_b = 0.05,\ 0.5$, and also when $\rho_b = \rho_w = 0.8$ and $\delta = R_b = 0.05$. Finally T_3 is better than T_2 when $\rho_w = 0.8,\ R_b = 5.0$, and when $\rho_b = \rho_w = 0.2$ with $R_b = 0.5,\ 5.0$.

5. APPLICATION

The proposed estimators were applied to a survey of teak plantations. The aim was to estimate the average height of teak trees using the girth as the auxiliary information.

Table 2

The Efficiency of T_1, T_2, and T_3 with Respect to T_0

		$\rho_w = 0.2$								
		$R_b = 0.05$			$R_b = 0.5$			$R_b = 5.0$		
			R_w			R_w			R_w	Strategy
ρ_b	δ	0.05	0.5	5.0	0.05	0.5	5.0	0.05	0.5	
0.2	0.05	1.04 1.01 0.98	1.04 0.73 0.71	1.04 0.18 0.18	0.83 0.81 0.98	0.83 0.62 0.71	0.83 0.17 0.18	0.20 0.20 0.98	0.20 0.19 0.71	T_1 T_2 T_3
	0.5	1.03 1.04 1.02	1.03 0.85 0.84	1.03 0.26 0.26	0.87 0.88 1.02	0.87 0.74 0.84	0.87 0.25 0.26	0.27 0.27 1.02	0.27 0.25 0.84	T_1 T_2 T_3
	5.0	1.02 1.04 1.03	1.02 1.03 1.03	1.02 0.67 0.67	0.97 0.99 1.03	0.97 0.99 1.03	0.97 0.65 0.67	0.60 0.60 1.03	0.60 0.60 1.03	T_1 T_2 T_3
0.8	0.05	1.62 1.53 1.16	1.62 0.94 0.77	1.62 0.19 0.18	2.53 2.35 1.16	2.53 1.23 0.77	2.53 0.20 0.18	0.45 0.45 1.16	0.45 0.38 0.77	T_1 T_2 T_3
	0.5	1.34 1.34 1.11	1.34 1.03 0.88	1.34 0.27 0.26	1.74 1.76 1.11	1.74 1.28 0.88	1.74 0.29 0.26	0.45 0.54 1.11	0.45 0.48 0.88	T_1 T_2 T_3
	5.0	1.07 1.10 1.05	1.07 1.09 1.03	1.07 0.69 0.67	1.13 1.16 1.05	1.13 1.15 1.03	1.13 0.72 0.67	0.83 0.84 1.05	0.83 0.83 1.03	T_1 T_2 T_3

		$\rho_w = 0.8$										
		$R_b = 0.05$				$R_b = 0.5$			$R_b = 5.0$			
				R_w			R_w			R_w		Strategy
ρ_b	δ	5.0	0.05	0.5	5.0	0.05	0.5	5.0	0.05	0.5	5.0	
0.2	0.05	0.20 0.11 0.18	1.05 1.07 1.04	1.05 0.85 0.83	1.05 0.23 0.23	0.83 0.85 1.04	0.83 0.70 0.83	0.83 0.21 0.23	0.20 0.19 1.04	0.20 0.19 0.83	0.20 0.12 0.23	T_1 T_2 T_3
	0.5	0.27 0.15 0.26	1.06 1.21 1.18	1.06 1.30 1.26	0.89 0.41 0.41	0.89 1.00 1.18	0.89 1.06 1.26	0.89 0.38 0.41	0.27 0.28 1.18	0.27 0.28 1.26	0.27 0.19 0.41	T_1 T_2 T_3
	5.0	0.60 0.46 0.67	1.17 1.31 1.30	1.17 1.64 1.63	1.17 2.03 2.00	1.09 1.22 1.30	1.09 1.51 1.63	1.09 1.87 2.00	0.62 0.67 1.30	0.62 0.76 1.63	0.62 0.84 2.00	T_1 T_2 T_3
0.8	0.05	0.45 0.15 0.18	1.65 1.70 1.27	1.65 1.22 0.98	1.65 0.25 0.24	2.55 2.72 1.26	2.55 1.64 0.98	2.55 0.27 0.24	0.46 0.46 1.27	0.46 0.42 0.98	0.46 0.18 0.24	T_1 T_2 T_3
	0.5	0.45 0.21 0.26	1.50 1.75 1.40	1.50 1.83 1.43	1.50 0.46 0.43	1.88 2.34 1.40	1.88 2.65 1.43	1.88 0.50 0.43	0.56 0.59 1.40	0.56 0.61 1.43	0.56 0.31 0.43	T_1 T_2 T_3
	5.0	0.83 0.85 0.67	1.30 1.46 1.39	1.30 1.85 1.74	1.30 2.25 2.04	1.35 1.53 1.39	1.35 1.98 1.74	1.35 2.53 2.04	0.95 1.03 1.39	0.95 1.22 1.74	0.95 1.38 2.04	T_1 T_2 T_3

Table 3

Estimated Efficiency of the Proposed Estimators with Respect to T_0 in the Estimation of the Average Height of Teak Trees

Estimators	Mean height (m)	Variance (m^2)	Estimated % Efficiency
T_0 (no matching)	20.04	6.3118	100
T' Partial matching	18.06	4.0680	155
T_1	17.86	0.0718	8791
T_2	17.31	0.0651	9635
T_3	17.99	4.0183	157

The teak trees used in this study were planted in 1965 with different spacings, producing plantations with the following number of trees per hectare: 2,000, 800, 400 and 250 trees. To measure the trees, an area of 40 metres by 40 metres was mapped out in each plantation after a sample of 8 plantations (FSUs) had been selected from 16 plantations, using the PPSWR scheme. The number of trees in each plantation was used as a measure of size. All the trees in the 40m by 40m area constituted the second stage units and the girth of each tree at breast height was measured. For the height measurements, a subsample of the trees was selected from the 40m by 40m area in each selected FSU. The first measurements were carried out in 1981 and the second in 1983. The sampling scheme used was the same as the one described in Section 2, with 50% matching of the FSUs.

The estimated efficiencies are given in Table 3. The sample estimates of the variance and covariance terms were used to obtain the optimum variances of T', T_1, T_2 and T_3 because the population values of these variances and covariances were not known. Therefore, the low values of the estimated optimum variances of T_1 and T_2 can be attributed partly to the nature of the sample data and partly to the nature of the estimators.

We observe that the estimator T_2 is more efficient than either T_1 or T_3, while T_1 is more efficient than T_3 in the estimation of the average height of teak trees using the girth as the auxiliary information.

ACKNOWLEDGEMENTS

I am grateful to the referee and the associate editor for their useful comments for the improvement of this paper. I thank Dr. O. Abe of the Department of Statistics, University of Ibadan, Ibadan, for going over the draft of the revised paper.

REFERENCES

ABRAHAM, T.P., KHOSLA, R.K., and KATHURIA, O.P. (1969). Some investigations of the use of successive sampling in pest and disease surveys. *Journal of the Indian Society of Agricultural Statistics*, 21, 43 - 57.

COCHRAN, W.G. (1977). *Sampling Techniques*, (3rd. ed.). New York: John Wiley.

JESSEN, R.J. (1942). Statistical investigations of a sample survey for obtaining farm facts. *Iowa Agricultural Experimental Station Research Bulletin*, 304, 54-59.

KATHURIA, O.P. (1975). Some estimators in two-stage sampling on successive occasions with partial matching at both stages. *Sankhyā*, Ser. C, 37, 147 - 162.

KATHURIA, O.P. (1978). Double sampling on successive occasions using a two-stage design. *Journal of the Indian Society of Agricultural Statistics*, 30, 49 - 64.

KATHURIA, O.P., and SINGH, D. (1971). Relative efficiencies of some alternative procedures in two-stage sampling on successive occasions. *Journal of the Indian Society of Agricultural Statistics*, 23, 101 - 114.

SINGH, S., and SRIVASTAVA, A.K. (1973). Use of auxiliary information in two-stage successive sampling. *Journal of the Indian Society of Agricultural Statistics*, 25, 101 - 114.

SINGH, D. (1968). Estimates in successive sampling using a multistage design. *Journal of the American Statistical Association*, 63, 99 - 112.

Corrigendum

'Some Optimality Results in the Presence of Nonresponse' by V.P. Godambe and M.E. Thompson, Survey Methodology (1986), 12, 29-36.

Formula (2.6), the definition of the optimum estimating function in $\bar{H}''$ $(p, \mathbf{q})$, should be

$$h''^{*} = \sum_{i \epsilon s'} (y_i - \theta x_i)\alpha_i / \pi_i q_i .$$

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.