12-001

c.3

Robert Lussier

Statistics Canada Statistique Canada

SURVEY METHODOLOGY

December 1983

Volume 9

Number 2

STATISTICS STATISTIQUE
CANADA CANADA

JUL
JUIL 30 2010

LIBRARY
BIBLIOTHÈQUE

A Journal produced by
Statistics Canada

Canada

SURVEY METHODOLOGY

December 1983 Vol. 9 No. 2

A Journal produced by Statistics Canada

CONTENTS

8-3200-501
Reference No.
Z - 079

ISSN: 0714-0045

SURVEY METHODOLOGY

December 1983 Vol. 9 No. 2

A Journal produced by Statistics Canada

Editorial Board:

Editorial Board:	R. Platek	- Chairman
	M.P. Singh	- Editor
	G.J.C. Hole	
	C. Patrick	
	P.F. Timmons	
	H. Lee	- Assistant Editor

Editorial Policy:

The objective of the Survey Methodology Journal is to provide a forum in a Canadian context for publication of articles on the practical applications of the many aspects of survey methodology. The Survey Methodology Journal will publish articles dealing with all phases of methodological development in surveys, such as, design problems in the context of practical constraints, data collection techniques and their effect on survey results, non-sampling errors, sampling systems development and application, statistical analysis, interpretation, evaluation and interrelationships among all of these survey phases. The emphasis will be on the development strategy and evaluation of specific survey methodologies as applied to actual surveys. 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 the Department.

Submission of Papers:

The Journal will be issued twice a year. Authors are invited to submit their papers, in either of the two Official Languages, to the Editor, Dr. M.P. Singh, Census and Household Survey Methods Division, Statistics Canada, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, K1A 0T6. Two copies of each paper, typed space-and-a-half, are requested.

COST MODELS FOR OPTIMUM ALLOCATION IN MULTI-STAGE SAMPLING

William D. Kalsbeek, Ophelia M. Mendoza
and
David V. Budescu[1]

Cost models to determine an optimum allocation of the sample among stages in cluster samples are considered. Results from a proposed cost model, which directly considers the implications of follow-up visits to sample clusters as well as other travel to and from the field by data collectors, are compared with results from existing cost models. The proposed model generally calls for fewer clusters with more elements selected per cluster than the existing models.

1. INTRODUCTION

One of the first issues in designing a multi-stage cluster sample is how to best allocate the sample among stages. In a two-stage design this amounts to deciding on the number of clusters to be selected in the first stage of sampling and the average sample size among selected clusters in the second stage. One normally wishes to choose that allocation of the sample among individual stages which will yield the best possible precision of estimates for the amount of funds available to conduct the survey. In the sequel, we will refer to this issue as the problem of determining an "optimum stage allocation".

The theory of optimum stage allocation requires both a variance and a cost model. The variance model is a mathematical formula for the precision of a survey estimator, written as a function of the sample sizes in each stage and certain measures of the components of unit variance attributable to each stage. Similarly, the cost model is a mathematical formula for the total cost

[1] William D. Kalsbeek and Ophelia M. Mendoza, Department of Biostatistics, University of North Carolina at Chapel Hill, and David V. Budescu, Department of Psychology, University of Haifa.

of conducting the survey, expressed as a function of the same stage-specific sample sizes but also various per-unit costs for each stage of the sampling design.

Variance models for many common multi-stage sampling designs have been produced, when the objective of the survey is to estimate the population means per element (see, for example, Hansen, Hurwitz, and Madow, 1953, and Cochran, 1977). Furthermore, important parameters of these variance models are readily estimable and can often be obtained from published reports. For example, Kish, Groves, and Krotki (1976) present estimates of one such parameter, the intraclass correlation coéfficient for several national fertility surveys.

The variance model used in this paper is a simple but common one. Suppose that the sample is selected in two stages from a population consisting of equal-sized clusters. If simple random sampling (with replacement) is used to first select a sample of n clusters and next a sample of m elementary units within selected clusters, then the variance of the estimated population mean per element $\bar{y}$ is simply

$$\mathrm{Var}(\bar{y}) = \sigma^2[1 + \rho(m - 1)]/nm, \tag{1.1}$$

where ρ is the intraclass correlation and σ^2 is the variance among all elementary units in the population. The result of (1.1) may also serve as a reasonable approximation even when clusters are of unequal size and selection procedures other than simple random sampling are used (see Kish 1965, Section 5.4). In this case we may view m as the average within-cluster sample size.

The development of reasonable cost models has received relatively little attention in the survey literature despite the fact that existing models contain parameters of survey cost which, though clearly defined, are difficult to compute. One such parameter is the cost of adding a cluster to the sample. Computing a reasonable measure of this per-unit cost is complicated by the difficulty in determining the impact of data collector travel which depends on such things as the size of the area being covered, the number of clusters assigned to each data collector, and the pattern of travel followed by the data collector in completing the survey. Some consolation can be derived from the known robustness of optimum stage allocation when imperfect cost measures are used (see Kish, 1976), although nontrivial departures from the best attainable precision may result when severely misinterpreted cost measures are used.

Two well known cost models have been applied to the survey setting in which data collection required a visit to each cluster by a data collector (or in some surveys a team of data collectors). We call the first of these models the simple model in which total non-overhead costs can be expressed as

$$C_0^{(\underset{\sim}{S})} = nC_1^{(\underset{\sim}{S})} + nmC_2^{(\underset{\sim}{S})}, \tag{1.2}$$

where $C_0^{(\underset{\sim}{S})}$ is the total nonoverhead cost, $C_1^{(\underset{\sim}{S})}$ is the average cost of adding a cluster to the sample, and $C_2^{(\underset{\sim}{S})}$ is the average cost of adding an elementary unit to the sample. The simple model, combined with the variance model of (1.1), yields (see Cochran 1977, Section 10.6)

$$m_{opt}^{(\underset{\sim}{S})} = \left\{ \left(\frac{1-\rho}{\rho}\right) \frac{C_1^{(\underset{\sim}{S})}}{C_2^{(\underset{\sim}{S})}} \right\}^{\frac{1}{2}} \tag{1.3}$$

as the optimum value of m.

The costs of travel during data collection often contribute significantly to total survey costs. Data collector travel and accompanying costs may be considered to be of two types. The first is between-cluster travel which refers to movement among clusters during a data collection trip. The second is positioning travel which refers to travel to the first cluster visited from the data collector's home base and then back to the home base from the last cluster visited during the data collection trip. The importance of the second cost model, suggested by Hansen, Hurwitz, and Madow (1953) and called the HHM Model here, is that it isolates between-cluster cost from the rest of the survey's total nonoverhead costs. This is done by assuming that the n clusters are uniformly arranged in a rectangular qeographic area of size A and that associated with each unit of distance travelled is a unit cost (U) consisting of two components: the mileage allowed for travel (e.g., in dollars per mile) and the ratio of hourly wages to the average rate of travel (e.g., in miles per hour).

In many surveys, data collection may require multiple visits to sample clusters. We incorporate the concept of follow-up visits into the HHM model by assuming the data collection is completed in H phases with np^{h-1} clusters being visited in the h-th phase; $0 < p < 1$. The cost of cluster follow-up is determined for the HHM model by summing the between-cluster travel cost over all phases. The HHM model as adapted here thereby takes the form,

$$C_0^{(\underset{\sim}{H})} = nC_1^{(\underset{\sim}{H})} + nmC_2^{(\underset{\sim}{H})} + n^{\frac{1}{2}}C_3^{(\underset{\sim}{H})} \tag{1.4}$$

where $C_3^{(\underset{\sim}{H})} = UA^{\frac{1}{2}}\ (1 - p^{H/2})/(1 - p^{\frac{1}{2}})$ is the cost parameter of the term isolating the cost of between-cluster travel with follow-up visits considered. The cost of adding a cluster $(C_1^{(\underset{\sim}{H})})$ and the cost of adding an element $(C_2^{(\underset{\sim}{H})})$ in the HHM model include positioning travel cost but exclude all remaining between-cluster travel costs which are covered by the term, $n^{\frac{1}{2}}C_3^{(\underset{\sim}{H})}$. The new HHM model, combined with the variance model once again, yields (see Hansen, et al., 1953, Vol. II, Section 6.11)

$$m_{\text{opt}}^{(\underset{\sim}{H})} = \left\{ \left(\frac{1-\rho}{\rho}\right) \frac{C_1^{(\underset{\sim}{H})} + C_3^{(\underset{\sim}{H})}/(2n^{\frac{1}{2}})}{C_2^{(\underset{\sim}{H})}} \right\}^{\frac{1}{2}} \tag{1.5}$$

which must be solved iteratively to determine the optimum value of m.

The intent of this paper is to extend the thinking about cost models used for optimum stage allocation and to produce a new model which more explicitly reflects actual survey costs. In so doing, we develop a cost model which: (1) isolates the increasingly important component of total survey costs due to data collector travel, (2) can easily accommodate follow-up visits to clusters, and (3) can be expressed as a relatively simple function of a number of readily interpretable measures.

2. PROPOSED MODEL

The cost model discussed in this section isolates from other survey costs the cost of both between-cluster and positioning travel for data collectors. This is contrasted by the HHM model where only between-cluster travel costs are isolated and by the simple model where isolation of travel costs does not occur at all. The proposed model can therefore be viewed as an attempt to avoid the difficulty in existing models of having to allocate unisolated travel costs among other per-unit costs, e.g., in the simple model data collector travel costs must be appropriated to $C_1^{(\underset{\sim}{S})}$ and $C_2^{(\underset{\sim}{S})}$. As with the HHM model, assumptions made for the proposed model regarding the location of clusters and the route of between-cluster travel are needed to express the survey's total travel cost as a function of n.

We shall see that assumptions concerning the spatial arrangement of clusters and travel by the data collectors are kept simple and admittedly somewhat naive. Less restrictive and presumably more realistic assumptions could be made, but the effect would be to add prohibitive complexity to the problem. We shall also see that the assumptions made in developing the proposed model allow one to express survey costs in terms of simple, well-known parameters of a survey operation. Thus, optimum stage allocation using the proposed model can be determined by specifying several easily understood measures characterizing a survey protocol.

2.1 Spatial Configuration of Sample Clusters

We now describe the spatial configuration of sample clusters as assumed for the proposed cost model and illustrated in Figure A. The object of the assumed configuration is for the uniformly scattered clusters to be arranged so that distances for reasonable travel routes can be expressed simply as a function of several readily obtained parameters. One assumes that the expressions will hold true for all possible parameter values.

Suppose that we have a survey population with land area of geographical size A and that the population is divided into t equal and nonoverlapping subareas, each of size A/t and containing $v = n/t$ sample clusters. One data collector is assigned to do the survey work in each subarea, which is shaped as a square with a number of evenly spaced concentric circles contained therein. The data collector's home base, assumed to be one of the clusters in the sample, lies in the center of the subarea in order to assure adequate accessibility to clusters during data collection. The distance from the home base to the outermost circle in each subarea is r. Thus, since the size of each subarea is $4r^2$, we have $r = (A/t)^{\frac{1}{2}}/2$. Moving from the home base in a subarea, the k-th circle $(k = 1,\ldots,K)$ contains 6k clusters. Assuming a multiple of six clusters on each concentric circle allows clusters to be almost uniformly spaced in the subarea, except for the square corners.

2.2 Data Collection Protocol

Using the spatial configuration of clusters just described, we now discuss a protocol for data collection which one might expect to observe in certain kinds of surveys with two or more stages of sampling. Comparison of results from existing cost models is later made within the context of this protocol.

Data collection in a subarea is assumed to require multiple phases of activity since work in most clusters usually involves several visits, some to make arrangements for data collection in the cluster and others to actually collect the data. As mentioned earlier, we let H denote the number of phases required to complete data collection in a subarea. This parameter can also be interpreted as the maximum number of required visits to individual clusters. In the h-th phase of data collection $(h = 1,2,\ldots,H)$, we assume that vp^{h-1} clusters (where $0 \leqslant p < 1$) are visited in a series of trips before proceeding with the next phase. Each trip involves a visit to ℓ neighboring clusters not previously visited during that phase of data collection. The cluster located in the home base is included in all phases of data collection.

Several assumptions are now made regarding movement of the data collectors among clusters. First the travel route followed in each trip proceeds from that data collector's home base, to each of the ℓ clusters (without backtracking), and then back again to the home base. Second, data collector travel is assumed to proceed in a straight line except between neighboring clusters on a circle where travel follows the arc of the circle. The choice of the arc distance over the straight-line is thought to be feasible since the formula for the former is simpler and since travel in surveys seldom follows a straight line.

Third, movement between two neighboring circles follows the shortest possible straight-line distance. This means that the cluster of departure from one circle and the cluster of destination on a neighboring circle are in line with the home base. The alignment of clusters 7 and 8 in Figure A illustrates this assumption. Fourth, travel within clusters and between data collector subareas is assumed to be negligible and is therefore not specifically isolated in the proposed model.

One final important assumption in the proposed model concerns the problem of the spatial configuration of clusters when $h > 1$; i.e., when the number of clusters visited during a phase of data collection is a subset of the v clusters originally selected in the subarea. To retain the simplicity of the concentric circle arrangement through all phases of data collection, we allow the number of concentric circles (K_h) at the h-th phase to vary according to the size of vp^{h-1} while fixing the size of the interviewer subarea at A/t. Thus, we have $K_h = (\alpha_h - 1)/2$, where $\alpha_h = \{1 + \frac{4}{3}(vp^{h-1} - 1)\}^{\frac{1}{2}}$.

2.3 Cost Formulation

Total travel cost in the proposed model is calculated as the product of U and the total distance travelled (D). Formulations for D, expressed alternatively as a function of the cluster workload per data collector (v) and the number of data collector subareas (t), are given below. Although the two formulations are functionally similar (since v = n/t), developing both solutions is thought to be important because either v or t may be specified in designing a survey. Details of the derivations for (2.1) - (2.5) are appended.

Assuming the above data collection protocol, the total distance travelled over all phases, expressed as a function of v, will be

$$D^{(\underset{\sim}{P})} = \delta_3^{(\underset{\sim}{P})} n^{\frac{1}{2}}, \tag{2.1}$$

where

$$\delta_3^{(\underset{\sim}{P})} = (A/v)^{\frac{1}{2}}[\frac{4}{3} \{v(1 - p^H)/(1 - p) - H\} + \{1 + (\ell - 1)\pi/2\}\{\sum_{h=1}^{H} \alpha_h + H\}]/2\ell.$$

This leads to a cost model which has the same general form as the HHM model of (1.4) but where the coefficient of the $n^{\frac{1}{2}}$ term is $U\delta_3^{(\underset{\sim}{P})}$ and the optimum value can be obtained from (1.5).

The total distance travelled, obtained as a function of t, can be written as

$$D^{(\underset{\sim}{P})} = \delta_0^{(\underset{\sim}{P})} + n\delta_1^{(\underset{\sim}{P})} + \sum_{h=1}^{H} \alpha_h \delta_4^{(\underset{\sim}{P})}, \tag{2.2}$$

where

$$\delta_0^{(\underset{\sim}{P})} = H(At)^{\frac{1}{2}}\{3(\ell - 1)\pi - 2\}/12\ell,$$

$$\delta_1^{(\underset{\sim}{P})} = 2\{(1 - p^H)/(1 - p)\}(A/t)^{\frac{1}{2}}/3\ell,$$

$$\delta_4^{(\underset{\sim}{P})} = (At)^{\frac{1}{2}}\{(\ell - 1)\pi + 2\}/4\ell.$$

The distance model of (2.2) leads to a cost model of the general form

$$C_0 = nC_1 + nmC_2 + \sum_{h=1}^{H} \alpha_h C_4. \tag{2.3}$$

Obtaining optimum values for n and m from (2.3) is an excessively cumbersome process which can be simplified by substituting a first-order Taylor series approximation (in n) for α_h, evaluated at t/p^{h-1} for simplicity. By so doing we have

$$\alpha_h \doteq (2p^{h-1}/3t)n + \frac{1}{3}, \tag{2.4}$$

which, when applied to (2.3), reduces the proposed cost model to

$$C_0^{(\underset{\sim}{P})} = nC_1^{(\underset{\sim}{P})} + nmC_2^{(\underset{\sim}{P})}, \tag{2.5}$$

where

$$C_0^{(\underset{\sim}{P})} = \underset{\sim}{C}_0 - U\{\delta_0^{(\underset{\sim}{P})} + H\delta_4^{(\underset{\sim}{P})}/3\},$$

$$C_1^{(\underset{\sim}{P})} = \underset{\sim}{C}_1 + U\{\delta_1^{(\underset{\sim}{P})} + 2\delta_4^{(\underset{\sim}{P})}(1 - p^H)/3t(1 - p)\},$$

$$C_2^{(\underset{\sim}{P})} = \underset{\sim}{C}_2.$$

$\underset{\sim}{C}_0$ is the total prespecified nonoverhead cost of the survey, $\underset{\sim}{C}_1$ is the prespecified average cost of adding a cluster to the sample (excluding all costs of data collector travel), and $\underset{\sim}{C}_2$ is the prespecified average cost of adding an element to the sample (excluding, once again, all data collector travel costs). We note from (2.5) that using the approximation for α_h has reduced the proposed model to the form which, except for the three cost parameters, resembles the simple cost model of (1.2). Optimum values of m and n are obtained from (1.3) and by solving for n in (2.5).

3. COMPARISON OF PROPOSED MODEL WITH EXISTING MODELS

In this section we compare results obtained from the proposed cost model (expressed as a function of v) with results from the simple and HHM cost models. We consider the situation where a two-stage survey of the United States is being planned, and the variance model of (1.1) is assumed in all comparisons. Measures used as the basis for comparisons among models are as follows: (1) optimum value of n, (2) optimum value of m, and (3) the variance of the survey estimate given the optimum allocation.

Optimum values of n and m for the simple HHM models are obtained from (1.3) and (1.5), respectively. To make comparisons with these models more realistic, adjustment factors are calculated to account for those travel costs not specifically isolated by the models. The adjustment procedure is similar to the approach mentioned earlier and suggested by Hansen, et al. (1953, Vol. 1, Section 6.13). To account for positioning travel costs in the HHM model we specify that

$$C_3^{(\underset{\sim}{H})} = \lambda^{(\underset{\sim}{H})}\underset{\sim}{C}_1,$$

$$C_2^{(\underset{\sim}{H})} = \lambda^{(\underset{\sim}{H})}\underset{\sim}{C}_2,$$

$$C_3^{(\underset{\sim}{H})} = \lambda^{(\underset{\sim}{H})}(A)^{\frac{1}{2}}U(1 - p^{H/2})/(1 - p^{\frac{1}{2}}),$$

where

$$\lambda^{(\underset{\sim}{H})} = \underset{\sim}{C}_0/\{n_{opt}^{(\underset{\sim}{P})}\underset{\sim}{C}_1 + n_{opt}^{(\underset{\sim}{P})}m_{opt}^{(\underset{\sim}{P})}\underset{\sim}{C}_2 + (n_{opt}^{(\underset{\sim}{P})}A)^{\frac{1}{2}}U\} \tag{3.1}$$

is the adjusting factor, $n_{opt}^{(\underset{\sim}{P})}$ is the corresponding optimum value for n under the proposed model, and $m_{opt}^{(\underset{\sim}{P})}$ is the corresponding optimum value for m under the proposed model. Using $\lambda^{(\underset{\sim}{H})}$ in this way has the effect of assuming that positioning travel costs contribute to each cost parameter of the HHM model by the same relative amount. In similar fashion, we account for all costs of data collector travel in the simple model by setting $C_1^{(\underset{\sim}{S})} = \lambda^{(\underset{\sim}{S})}\underset{\sim}{C}_1$ and $C_2^{(\underset{\sim}{S})} = \lambda^{(\underset{\sim}{S})}\underset{\sim}{C}_2$, where the adjustment factor is

$$\lambda^{(\underset{\sim}{S})} = \underset{\sim}{C}_0/(n_{opt}^{(\underset{\sim}{P})}\underset{\sim}{C}_1 + n_{opt}^{(\underset{\sim}{P})}m_{opt}^{(\underset{\sim}{P})}\underset{\sim}{C}_2). \tag{3.2}$$

We must acknowledge the synthetic nature of the adjustment factors, $\lambda^{(\underset{\sim}{H})}$ and $\lambda^{(\underset{\sim}{S})}$, used for our comparisons. In each case the adjustment factor is a function of the optimum values of n and m obtained from the corresponding proposed model. In reality, these factors would be calculated for the HHM and simple models by estimating the proportion of the survey's budget not spent on those travel costs left unaccounted for by the model. One might suspect that this estimated proportion would, at best, amount to a rough approximation which would probably differ from the adjustments produced from (3.1) and

(3.2). Thus, we suspect that using these factors may contribute to making the simple and HHM models seem more comparable to the proposed model than they in fact are.

3.1 Assumed Parameter Values

Producing the findings of the comparison study required several numerical values for the various statistical and cost parameters of the models. First, we consider national surveys in the United States, A=3,042,265 square miles, the land area of the United States, excluding Alaska and Hawaii. We also arbitrarily set $\underset{\sim}{C}_0$ = \$500,000, the total nonoverhead cost of the survey, and U = \$0.45, the unit cost per mile travelled. The latter figure is obtained by assuming a mileage allowance of \$0.25 per mile, an interviewer salary of \$6.00 per hour, and an average travel rate of 30 miles per hour. All combinations of the following groups of parameters are considered in our comparisons:

$(\underset{\sim}{C}_1, v)$: (\$50, 20); (\$250, 5)

$(\underset{\sim}{C}_2, p, H)$: (\$10, 0.3, 5); (\$25, 0.8, 20)

ℓ: 1; 2

ρ: 0.05; 0.15

Parameters were grouped in this manner since many of the combinations resulting from individual parameters were thought to be unrealistic.

The parameters $\underset{\sim}{C}_1$ and v are grouped together to indicate the degree of difficulty that data collectors would have in setting up and maintaining participation among clusters in the survey. For example, in a one-time survey or the first installment of an ongoing survey, one might expect to find cluster set-up costs to be high and the set-up activities to be sufficiently burdensome so that the average number of clusters assigned per data collector would of necessity be low. Thus, for present purposes we designate $\underset{\sim}{C}_1$ = \$250 and v = 5 to indicate cluster set-up and maintenance which is "difficult". Activities such as obtaining endorsements, making initial visits to solicit cooperation, and constructing the frame for selecting the second stage would all contribute toward the determination of these values. We designate $\underset{\sim}{C}_1$ = \$50 and v = 20 to indicate cluster set-up and maintenance activities which are "easy". This situation might be observed in surveys in which set-up activities are relatively simple. One example would be a subsequent installment of the ongoing survey while another would be a survey in which arrangements can be made by mail or telephone. The parameters $\underset{\sim}{C}_2$, p, and H are used to jointly indicate the level of difficulty in the data collection protocol. When $\underset{\sim}{C}_2$ = \$10, p = 0.3, and H = 5, the average number of times a cluster will be visited is 1.4 and data collection is assumed to be "easy". This may occur, for example, in a survey where the protocol requires only that a small amount of readily accessible data be extracted for each element in a cluster. When less accessible data are extracted or when follow-up of selected elements is required, data collection might be called "difficult" in which case we assume that $\underset{\sim}{C}_2$ = \$25, p = 0.8, and H = 20, thus implying that the average number of times a cluster will be visited is 5.6.

The parameter indicating the number of clusters visited per trip (ℓ) assumes the values 1 or 2 in these comparisons. Allowing $\ell \geq 2$ is thought to be unrealistic in national surveys since distances would preclude visiting a large number of clusters on a single trip. Two moderate values of intraclass correlation (ρ) are assumed.

3.2 Findings

Tables 1-3 contain the results of the comparison study involving the proposed model and the versions of the simple model and of the HHM model where $\lambda^{(\underset{\sim}{S})}$ and $\lambda^{(\underset{\sim}{H})}$ are applied, respectively. Optimum values of n and m, as determined under the proposed model, are presented in Table 1. As expected, optimum values of n tend to be lower when cluster set-up and maintenance is difficult, and optimum values of m tend to be lower when data collection is difficult.

The major focus of the comparison study is the difference between optimum results under the proposed model and comparable results under the simple and HHM models. Optimum results for the proposed and simple models are compared in Table 2 in which one notes that differences are generally substantial. Optimum values for n under the proposed model are found to be between 2.4 and 60.0 percent lower than under the simple model, while optimum values for m are between 7.6 and 198.2 percent higher under the proposed model. These large differences are thought to be attributable to the ability of the proposed model to isolate between-cluster and positioning travel costs. This results in greater per-cluster costs and a smaller optimum number of sample clusters. The greatest differences in optimum variances, computed by applying the optimum values of n and m to (1.1), occur in surveys with easy cluster set-up and maintenance and difficult data collection. One might speculate that the magnitude of these variance differences is largely due to the relatively heavy cluster workload (i.e., $v = 20$) assumable when cluster set-up and maintenance is deemed easy. However, when this workload is lightened (i.e., $v = 5$) and considered with the same combination of parameters, the relative difference among optimum variances is reduced but remains substantial at 11-16 percent, as opposed to the 18-27 percent figures presented in Table 2.

The effects of the number of clusters visited per trip (ℓ) and the intraclass correlation (ρ) are also readily apparent in Table 2. Larger differences appear when $\ell = 1$ than when $\ell = 2$. This effect can be attributable to the greater importance that travel costs would play when only a single cluster can be visited per trip to the field. Furthermore, when $\rho = 0.05$, relative differences for n and m are somewhat greater than when $\rho = 0.15$; however these differences are an artifact due in part to the iterative approach which is used to obtain $m^{(\underset{\sim}{P})}_{opt}$. From (1.3) and (1.5) we would expect relative differences on optimum values of m to be identical.

The relative differences between the proposed and HHM models presented in Table 3 remain notable but are generally smaller than the differences reported in Table 2. We suspect that the greater similarity between results under the proposed and HHM models can be attributable to the fact that the HHM model represents a more realistic reflection of survey costs than does the simple model. However, as with comparisons involving the simple model, optimum values of n are smaller and optimum values of m are higher under the proposed model in Table 3. These comparisons also reveal once again that the largest differences in optimum variance occur in surveys with easy cluster set-up and maintenance and difficult data collection. Variance differences in other instances are negligible.

3.3 Discussion

We have proposed a cost model where the important component of travel during data collection can be completely set apart to improve one's ability to accurately reflect survey costs in determining an optimum stage allocation. In addition, a study designed to compare optimum results of this proposed model with two existing cost models has indicated substantial differences. However, aside from these differences, perhaps the most important practical implication of the proposed cost model is that the optimum stage allocation can be produced by specifying measures which are intuitively simple. These measures are of two types: fiscal and nonfiscal characteristics of the survey design. The required fiscal characteristics (i.e., $\underset{\sim}{C}_0$, $\underset{\sim}{C}_1$, and $\underset{\sim}{C}_2$) can be determined by estimating the costs of certain components of the survey. For example, we might determine $\underset{\sim}{C}_1$ from a recent similar survey as the average per-cluster cost of choosing the sample of clusters, soliciting among clusters for participation in the survey (excluding travel costs), and constructing the sampling frame for sampling units within selected clusters. The required nonfiscal characteristics of the survey (i.e., A, v or t, p, H, ℓ, and ρ) can be obtained as factual information from prior surveys. For example, knowledge of the maximum and average number of visits required per cluster in a recent similar survey would determine p and H.

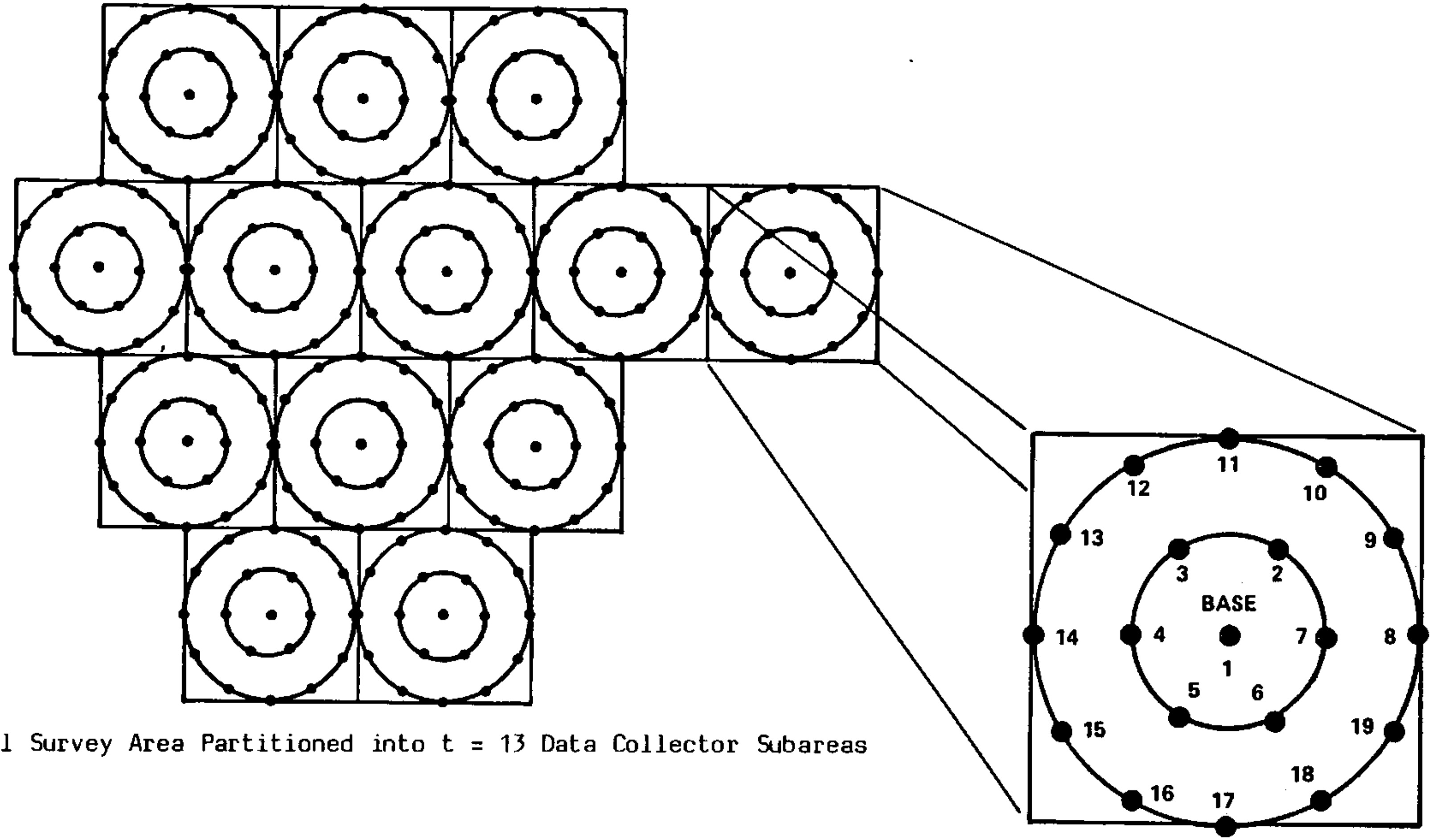
We conclude by briefly examining the robustness and artificiality of the proposed cost model. Robustness is considered on the one hand by determining (from stated assumptions) the types of surveys for which the model is likely to be useful. Assumptions of the model imply that the sample points are clustered rather than randomly scattered in the population and that during data collection a group of these clusters is assigned to each data collector. This arrangement of sample points and data collection assignments will occur in certain types of household and institutional samples. An example of one such arrangement is the National Survey of Nursing Homes (see National Center for Health Statistics, 1968) which is selected in two stages with nursing homes designated as clusters.

The arrangement might also appear in household surveys where the sample is chosen in two or more stages and where data collectors complete interviews within several small area segments (see, for example, the Virginia Health Survey conducted by the Statistical Sciences Group, Research Trianqle Institute, 1978). A household sample chosen in three or more staqes can be accommodated by treating A as the size of the land area occupied just by selected primary sampling units (PSU's) and then considering samplinq units from the second or subsequent stages to be the clusters that follow a concentric configuration within each data collector subarea (i.e., consider Figure A with t scattered rather than contiguous subareas). Procedurally, one would substitute t$\bar{A}$ for A in (2.2), where $\bar{A}$ is the average land area to be covered by each data collector in the planned survey. Given this adaptation, it is important to note that the number of sample PSU's would be prespecified and thereby not optimized, that n in the cost and variance models would be the number of sample <u>clusters</u> (i.e., not PSU's), and that m would be the average number of elementary units per cluster. Treating the number of sample PSU's to be fixed and then determining the optimum allocation for subsequent staqes would be reasonable for certain surveys where the ultimate sample is chosen from a sample of PSU's which is used repeatedly for different surveys. The variance model of (1.1) may have to be modified to reflect the additional sampling staqes (see Hansen, et al., 1953, Vol. II, Section 6.9). Some institutional samples selected in three or more stages (e.q., the Hospital Discharqe Survey of the National Center for Health Statistics, 1970) could be considered for the multi-stage adaptation as well. However, the proposed model would be less practical for those surveys where cluster sizes are so large that each data collector is assigned only one or two clusters or where selected clusters are not likely to be uniformly scattered about within subareas.

Another facet of the robustness issue is the generalizability of the findings. Clearly, any conclusions drawn from our findings must be limited by the parameter values we have assumed. Rather than using values from existing surveys in which case inferences would be limited to those surveys, our strategy was to create several prototype surveys based upon combinations of unit costs and other parameters thought to reflect current survey practice. Values used to create the prototype were often taken directly or inferred from recent surveys known to the authors.

Finally, a certain degree of impracticality is the price one pays to keep things simple since realism and simplicity seem to be indirectly related in building survey cost models. Thus, while the intent of our research has been to find a more realistic yet simple model, we must acknowledge a substantial amount of remaining artificiality in our assumptions. For example, clusters are more likely to be randomly scattered than to exist as multiples of six lying on concentric circles. Moreover, travel between neighboring clusters would follow winding, circuitous routes rather than arcs or straight lines, and return visits to clusters would have more haphazard schedules than well-established phases of follow-up with the number of clusters per phase decreasing each time by a factor of p. While the proposed model reflects the orderliness which one hopes for in most survey field operations, it, like other existing models, fails to capture the unpredictability of things which tends to blend into the orderliness. Stochastic events can be used to create unpredictability but adding them tends to complicate the model to the point of being less useful mathematically. Until more realistic assumptions can be tied to simplicity, we are faced with the need to settle for cost models which fall short of the realism we seek.

Figure A. Proposed Spatial Configuration of Clusters in a Survey Population



Total Survey Area Partitioned into t = 13 Data Collector Subareas

Spatial Arrangement of Clusters
in Each Data Collector Subarea

TABLE 1. Optimum Values for n and m Under the Proposed Cost Model
(A = 3,042,265 square miles; $\underset{\sim}{C}_0$ = $500,000)

Prototype Survey	Parameters				Optimum Values	
	Cluster set-up and maintenance	Data collection	ℓ	ρ	$n^{(\underset{\sim}{p})}_{opt}$	$m^{(\underset{\sim}{p})}_{opt}$
[1]	Easy	Easy	1	0.05	1673	14.1
[2]				0.15	2319	7.4
[3]			2	0.05	1910	13.1
[4]				0.15	2669	6.9
[5]		Difficult	1	0.05	385	18.4
[6]				0.15	518	9.4
[7]			2	0.05	489	16.0
[8]				0.15	675	8.2
[9]	Difficult	Easy	1	0.05	871	23.6
[10]				0.15	1095	12.8
[11]			2	0.05	847	23.9
[12]				0.15	1065	12.9
[13]		Difficult	1	0.05	426	19.0
[14]				0.15	560	10.1
[15]			2	0.05	378	20.2
[16]				0.15	493	10.6

Cluster set-up and maintenance	$(\underset{\sim}{C}_1, v)$	Easy	($50, 20)
		Difficult	($250, 5)
Data collection	$(\underset{\sim}{C}_2, p, H)$	Easy	($10, 0.3, 5)
		Difficult	($25, 0.8, 20)

TABLE 2. Relative Differences Between the Proposed Model and the Simple Model

($\underset{\sim}{C}_0$ = \$500,000)

	Parameters				Relative difference: proposed vs simple model (in percent)		
Prototype Survey	Cluster set-up and maintenance	Data collection	ℓ	ρ	n_{opt}	m_{opt}	Optimum Variance
[1]	Easy	Easy	1	0.05	-22.7	44.3	3.1
[2]				0.15	-16.6	38.5	2.7
[3]			2	0.05	-18.5	34.4	2.0
[4]				0.15	-13.3	29.7	1.7
[5]		Difficult	1	0.05	-60.0	198.2	24.4
[6]				0.15	-52.9	179.3	26.9
[7]			2	0.05	-54.6	159.2	18.1
[8]				0.15	-47.2	142.3	19.6
[9]	Difficult	Easy	1	0.05	-3.8	8.4	0.2
[10]				0.15	-2.4	7.6	0.1
[11]			2	0.05	-4.3	9.7	0.2
[12]				0.15	-2.7	8.7	0.2
[13]		Difficult	1	0.05	-18.0	37.8	2.6
[14]				0.15	-12.4	33.6	2.1
[15]			2	0.05	-21.2	46.2	3.6
[16]				0.15	-15.1	41.3	3.0

Relative difference is computed as the measure under the proposed model minus the measure under the simple model divided by the measure under the simple model, and multiplied by 100.

Cluster set-up and maintenance ($\underset{\sim}{C}_1$, v) $\left\{\begin{array}{ll}\text{Easy} & (\$50,\ 20)\\ \text{Difficult} & (\$250,\ 5)\end{array}\right.$

Data collection ($\underset{\sim}{C}_2$, p, h) $\left\{\begin{array}{ll}\text{Easy} & (\$10,\ 0.3,\ 5)\\ \text{Difficult} & (\$25.\ 0.8,\ 20)\end{array}\right.$

TABLE 3. Relative Differences Between the Proposed Model and the HHM Model

($\underset{\sim}{C}_0$ = $500,000)

	Parameters				Relative difference: proposed vs HHM model (in percent)		
Prototype Survey	Cluster set-up and maintenance	Data collection	ℓ	ρ	n_{opt}	m_{opt}	Optimum Variance
[1]	Easy	Easy	1	0.05	-12.5	22.9	0.9
[2]				0.15	-8.9	20.2	0.8
[3]			2	0.05	-8.7	15.2	0.4
[4]				0.15	-6.0	13.4	0.4
[5]		Difficult	1	0.05	-22.8	49.6	3.5
[6]				0.15	-17.5	46.2	3.3
[7]			2	0.05	-17.2	34.2	1.8
[8]				0.15	-13.0	31.7	1.7
[9]	Difficult	Easy	1	0.05	-1.3	2.9	0.0+
[10]				0.15	-0.8	2.6	-0.0
[11]			2	0.05	-1.8	4.0	0.0+
[12]				0.15	-1.1	3.6	0.0+
[13]		Difficult	1	0.05	-4.0	7.9	0.2
[14]				0.15	-2.6	7.2	0.1
[15]			2	0.05	-6.5	13.5	0.4
[16]				0.15	-4.4	12.2	0.3

Relative difference is computed as the measure under the proposed model minus the measure under the HHM model, divided by the measure under the HHM model, and multiplied by 100.

$$\text{Cluster set-up and maintenance}\quad (\underset{\sim}{C}_1,\ v)\quad \begin{cases} \text{Easy} & (\$50,\ 20) \\ \text{Difficult} & (\$250,\ 5) \end{cases}$$

$$\text{Data collection}\quad (\underset{\sim}{C}_2,\ p,\ h)\quad \begin{cases} \text{Easy} & (\$10,\ 0.3,\ 5) \\ \text{Difficult} & (\$25.\ 0.8,\ 20) \end{cases}$$

APPENDIX

Details of the derivations for (2.1) - (2.5) in the text are presented here. Using the assumed spatial configuration of clusters and data collection protocol for the proposed model as discussed in Sections 2.1 and 2.2, respectively, the total distance travelled ($D^{(\underset{\sim}{P})}$) is first expressed as a function of the number of sample clusters assigned to each data collector (v). Given the configuration of clusters as illustrated in Figure A, note that the positioning and between-cluster travel distances for each data collector during the h-th phase of data collection are $\{12r/K_h\ell\} \sum_{k=1}^{K_h} k^2$ and $\{2\pi r(\ell - 1)/K_h\ell\} \sum_{k=1}^{K_h} k$, respectively. Summing these two distances, recalling that $K_h = (\alpha_h - 1)/2$, where $\alpha_h = \{1 + \frac{4}{3}(vp^{h-1} - 1)\}^{\frac{1}{2}}$, and multiplying times the number of data collectors (t), we have the total positioning and between-cluster travel distance for the h-th phase expressed as:

$$D_h = rt[2K_h(K_h + 1)(2K_h + 1) + (\ell - 1)\pi K_h(K_h + 1)]/K_h\ell$$

$$= rt[(\alpha_h - 1)^2 + \{6 + (\ell - 1)\pi\}(\alpha_h - 1)/2 + 2 + (\ell - 1)\pi]/\ell. \qquad (A.1)$$

Noting that $\sum_{h=1}^{H} p^{h-1} = (1 - p^H)/(1 - p)$, we sum D_h over all phases to obtain:

$$D^{(\underset{\sim}{P})} = \sum_{h=1}^{H} D_h = (rt/\ell) \sum_{h=1}^{H} [\{1 + \frac{4}{3}(vp^{h-1} - 1)\} + \alpha_h + \{\alpha_h + 1\}\{(\ell - 1)\pi/2\}]$$

$$= rt[\frac{4}{3}\{v(1 - p^H)/(1 - p) - H\} + \{1 + (\ell - 1)\pi/2\}\{\sum_{h=1}^{H} \alpha_h + H\}]/\ell. \qquad (A.2)$$

Recalling that $r = (A/t)^{\frac{1}{2}}/2$ and $t = n/v$ and substituting these identities into (A.2) leads to (2.1).

To express $D^{(\underset{\sim}{P})}$ as a function of the number of data collectors (t), first note that we must use $\alpha_h = \{1 + \frac{4}{3}(np^{h-1}/t - 1)\}^{\frac{1}{2}}$ as opposed to the earlier expression for α_h. Using the new expression complicates things a bit since α_h is now a function of both t and the number of sample clusters (n), which is one of the parameters to be optimized. Using the new expression for α_h and recalling once again that $r = (A/t)^{\frac{1}{2}}/2$, a bit of algebra allows us to recast (A.1) as:

$$\begin{aligned} D_h &= rt[\alpha_h{}^2 + \{1 + (\ell - 1)\pi/2\}\alpha_h + (\ell - 1)\pi/2]/\ell \\ &= rt[\{3(\ell - 1)\pi - 2\}/6 + (\frac{4}{3}\, p^{h-1}/t)n + \{1 + (\ell - 1)\pi/2\}\alpha_h]/\ell \\ &= (At)^{\frac{1}{2}}\{3(\ell - 1)\pi - 2\}/12\ell + n\{2p^{h-1}(A/t)^{\frac{1}{2}}/3\ell\} \\ &\quad + \alpha_h(At)^{\frac{1}{2}}\{(\ell - 1)\pi + 2\}/4\ell. \end{aligned} \tag{A.3}$$

Summing D_h from (A.3) over all phases leads us to the total distance given in (2.2),

$$D^{(\underset{\sim}{P})} = \delta_0^{(\underset{\sim}{P})} + n\delta_1^{(\underset{\sim}{P})} + \sum_{h=1}^{H} \alpha_h \delta_4^{(\underset{\sim}{P})}, \tag{A.4}$$

where

$$\delta_0^{(\underset{\sim}{P})} = H(At)^{\frac{1}{2}}\{3(\ell - 1)\pi - 2\}/12\ell,$$

$$\delta_1^{(\underset{\sim}{P})} = 2\{(1 - p^H)/(1 - p)\}(A/t)^{\frac{1}{2}}/3\ell,$$

$$\delta_4^{(\underset{\sim}{P})} = (At)^{\frac{1}{2}}\{(\ell - 1)\pi + 2\}/4\ell.$$

The total travel distance given by (A.4) leads to an overall survey cost model given by:

$$\underset{\sim}{C}_0 = n\underset{\sim}{C}_1 + nm\underset{\sim}{C}_2 + U\delta_0^{(\underset{\sim}{P})} + Un\delta_1^{(\underset{\sim}{P})} + U\sum_{h=1}^{H}\{1 + \frac{4}{3}\,(np^{h-1}/t - 1)\}^{\frac{1}{2}}\delta_4^{(\underset{\sim}{P})}. \tag{A.5}$$

Where $\underset{\sim}{C}_0$ is the total prespecified nonoverhead cost of the survey, $\underset{\sim}{C}_1$ is the prespecified average cost of adding a cluster to the sample (excluding all costs of data collector travel), and $\underset{\sim}{C}_2$ is the prespecified average cost of adding an element to the sample (excluding, once again, all data collector travel costs).

Using the cost model given by (A.5) to obtain optimum values for n and m is disadvantageous because the final righthand term of (A.5) is a complex function of n. To circumvent this difficulty we suggest substituting a first-order Taylor series approximation in n for $\alpha_h = \{1 + \frac{4}{3}(np^{h-1}/t - 1)\}^{\frac{1}{2}}$, which is arbitrarily evaluated at t/p^{h-1} to simplify the approximation. By so doing we have

$$\alpha_h = f(n) = \{1 + \frac{4}{3}(np^{h-1}/t - 1)\}^{\frac{1}{2}} \doteq f(t/p^{h-1}) + f'(t/p^{h-1})(n - t/p^{h-1})$$

where f'(.) is the first partial derivative of f(.) with respect to n. Since $f(t/p^{h-1}) = 1$ and $f'(t/p^{h-1}) = 2p^{h-1}/3t$, we have

$$\alpha_h \doteq (2p^{h-1}/3t)n + \frac{1}{3} \qquad \text{(A.6)}$$

which is a linear function of n. Applying the approximation of (A.6) to (A.5) reduces the proposed model to the form,

$$C_0^{(\underset{\sim}{P})} = nC_1^{(\underset{\sim}{P})} + nmC_2^{(\underset{\sim}{P})}, \qquad \text{(A.7)}$$

where

$$C_0^{(\underset{\sim}{P})} = \underset{\sim}{C}_0 - U\{\delta_0^{(\underset{\sim}{P})} + H\delta_4^{(\underset{\sim}{P})}/3\},$$

$$C_1^{(\underset{\sim}{P})} = \underset{\sim}{C}_1 + U\{\delta_1^{(\underset{\sim}{P})} + 2\delta_4^{(\underset{\sim}{P})}(1 - p^H)/3t(1 - p)\},$$

and

$$C_2^{(\underset{\sim}{P})} = \underset{\sim}{C}_2.$$

The result of (A.7) corresponds to (2.5) in the main text.

REFERENCES

[1] Cochran, W.G. (1977). *Sampling Techniques.* Third Edition. New York: John Wiley and Sons.

[2] Hansen, M.H., Hurwitz, W.H., and Madow, W.G. (1953). *Sample Survey Methods and Theory*. Vols. I and II. New York: John Wiley and Sons.

[3] Kish, L., Groves, R.M., and Krotki, K.P. (1976). *Sampling Errors in Fertility Surveys*. World Fertility Survey Occasional Paper No. 17. London. World Fertility Survey.

[4] Kish, L. (1976). "Optima and Proxima in Linear Sample Designs". *Journal of Royal Statistical Society*, Series A. 139: 80-95.

[5] National Center for Health Statistics (1968). "Design and Methodology for a National Survey of Nursing Homes". *Vital and Health Statistics*. PHS Pub. No. 1000, Series 1, No. 7. Washington: U.S. Government Printing Office.

[6] National Center for Health Statistics (1970). "Development of the Design of the NCHS Hospital Discharge Survey". *Vital and Health Statistics*. PHS Pub. No. 1000, Series 2, No. 39. Washington: U.S. Government Printing Office.

[7] Statistical Sciences Group (1978). *Virginia Health Survey: Volume I --- Methodological Report*, Report No. RTI/1546/00-00F, Research Triangle Institute, Research Triangle Park, North Carolina.

[8] U.S. Bureau of the Census (1978). *The Current Population Survey: Design and Methodology*. Technical Paper No. 40. Washington: U.S. Government Printing Office.

EVALUATION OF COMPOSITE ESTIMATION FOR THE CANADIAN LABOUR FORCE SURVEY[1]

S. Kumar and H. Lee[2]

This study considers the suitability of composite estimation techniques for the Canadian Labour Force Survey. The performance of a class of AK composite estimators introduced initially by Gurney and Daly is investigated for several characteristics. While the ordinary composite estimate has a large bias, the AK composite estimate is capable of reducing the bias. Composite estimates having minimum variance and minimum mean square error are compared.

1. INTRODUCTION

The Canadian Labour Force Survey (LFS) is conducted each month by Statistics Canada and is designed to produce estimates for various labour force characteristics. The LFS sample design follows a rotation scheme that permits the replacement of one-sixth of the households in the sample each month (see [7]). The sample is composed of six panels or rotation groups. A panel remains in the sample for a period of six consecutive months.

As pointed out in Bailar [1], one of the major drawbacks of composite estimation currently in use for the U.S. Current Population Survey (CPS) is its bias as compared to the simple ratio estimator for estimates of level. This bias stems from rotation group differences: the phenomenon that estimates based on data from different panels relating to the same time period do not have the same expected value. This phenomenon, often referred to as the rotation group bias, has been studied for LFS (see [2] and [6]). Recently, Huang and Ernst [4] have reported results in the context of the CPS on the performance of AK composite estimator introduced initially by Gurney and Daly [3]. A and K are

[1] Presented at the Joint Statistical Meetings of the American Statistical Association, the Biometric Society, the Institute of Mathematical Statistics and the Statistical Society of Canada in Toronto, August 1983.

[2] S. Kumar and H. Lee, Census and Household Survey Methods Division, Statistics Canada.

constants in the equation defining the composite estimator. Their results show improvement over the composite estimates currently in use for CPS as regards variance and bias.

The objective of this investigation is to study the suitability of composite estimation techniques for LFS. In this study the performance of different composite estimators of level and change will be investigated for the following five characteristics; in labour force, employed, employed agriculture, employed non-agriculture, and unemployed. These composite estimators are compared with the simple ratio estimator which is presently in use for LFS. The study is based on the province of Ontario data for 1980-81.

2. DEFINITIONS AND NOTATION

We are interested in estimating Y_m the number of persons in the population with a certain characteristic for the month m. Let

$y_{m,i}$ = A simple ratio estimator of Y_m based on the i-th panel (i = 1,2,...,6). Here the i-th panel refers to the sub-sample (rotation group) that is in the sample for the i-th time. It will be referred to as the i-th panel estimator.

$d_{m,m-1}$ = estimator of change $(Y_m - Y_{m-1})$ from the month (m - 1) to the month m based on five panels that are common to the months m and (m - 1)

$$= \sum_{j=2}^{6} (y_{m,j} - y_{m-1,j-1})/5. \tag{2.1}$$

y'_m = AK composite estimator of Y_m defined as

$$y'_m = (1 - K + A)y_{m,1}/6 + (1 - K - \frac{A}{5}) \sum_{j=2}^{6} y_{m,j}/6$$

$$+ K(y'_{m-1} + d_{m,m-1}) \qquad (2.2)$$

where K and A are constants, and $0 \leq K < 1$.

The equation (2.2) defines a class of estimators referred to as AK composite estimators. The estimators obtained by taking $A = 0$ in (2.2) are referred to as K composite estimators. The simple ratio estimator, to be denoted by $\bar{y}_m$, the mean of six panel estimators can be obtained by taking $A = 0$ and $K = 0$ in (2.2). We investigate the relative performance of the optimal (minimum variance or minimum mean square error) AK composite, K composite and simple ratio estimators.

We assume the rotation group bias $E(y_{m,i}) - Y_m$ is independent of m and is a function of i. We denote this bias by α_i. Formally

$$\alpha_i = E(y_{m,i}) - Y_m. \qquad (2.3)$$

The expression for the bias of the composite estimator is given in Appendix I.

3. ASSUMPTIONS

The rotation system in the LFS is schematically described in Table 1, where the current (month m) panel i (= 1,2,...,6, denoting interview month no.) is the same as panel i - j in month m - j, provided i - j lies between 1 and 5. The immediate predecessor to panel i of month m as of month m - j is given by (6 + i - j) provided (6 + i - j) lies between 1 and 6. Likewise, the second predecessor to panel i as of month m - j is given by (12 + i - j) provided (12 + i - j) lies between 1 and 6. In general, the r-th predecessor to panel i of month m is given by (i - j + 6r) in month m - j. Note that the 0-th predecessor to a panel means the same panel in earlier months.

The expression for the variance of y'_m, i.e. $V(y'_m)$ involves the variances and covariances of various panel estimators (see Appendix II). The following variance-covariance structure for various panel estimators is assumed. The assumptions conform to the LFS rotation pattern, illustrated in Table 1.

(i) $V(y_{m,i}) = \sigma^2$ for all m and i = 1,2,...,6,

(ii) $\mathrm{Cov}(y_{m,i}, y_{m-j,i-j+6r}) = \gamma_j^{(r)}\sigma^2$, where i = 1,2,...,6, j > 0 and r ≥ 0, such that $6 \geq i - j + 6r \geq 1$. Here r denotes the number of predecessors to the current panel.

For r = 0, i.e., $6 > i - j \geq 1$, let $\gamma_j^{(r)} = \rho_j$ (based on overlapping panels of months m and m - j).

For r = 1, i.e., $6 \geq i - j + 6 \geq 1$, let $\gamma_j^{(r)} = \gamma_j$ (based on the current panel and its immediate predecessor j months back).

For r ≥ 2, i.e., $6 \geq i - j + 6r \geq 1$, let $\gamma_j^{(r)} = 0$ (based on the current panel and its r-th predecessor j months back).

(iii) Of interest to the development of the variance of the composite estimator y_m' are the correlation coefficients ρ_j and γ_j, both of which are assumed to be stationary; i.e. they are functions of j and not of m. It is reasonable to assume that both ρ_j's and γ_j's are positive since ρ_j's are based on characteristics of largely common households while γ_j's are based on the characteristics of households in the current month and those of their near (in many cases next door) neighbours j months back (apart from cluster rotation).

(iv) The expression for $V(y_m')$ contains covariance terms not included in the assumptions (ii) and (iii). Some of these are:

$\mathrm{Cov}(y_{m,i}, y_{m,j})$ for i ≠ j, $\mathrm{Cov}(y_{m,i}, y_{m-1,j})$ for i = 1, j ≠ 6, and i ≠ 1, j ≠ i - 1, and $\mathrm{Cov}(y_{m,i}, y_{m-g,j})$ for g ≥ 12. These and all other covariances not defined above, including those with $\gamma_j^{(2)}$ and existing in the expression for $V(y_m')$ are assumed to be zero.

Following these assumptions, a variance expression for the AK composite estimator was derived in terms of the above parameters. The mathematical details for derivation of the expression for the bias and variance of y'_m, and the variance of $y'_m - y'_{m-1}$ are given in the appendices.

4. RESULTS AND DISCUSSION

The quantities σ^2, ρ_j and γ_j in the expresssion for $V(y'_m)$ were replaced by their estimates (For details of the methodology for estimating ρ's and γ's, see [5]). Note that, in the Canadian LFS ρ_j's do not exist for $j \geq 6$ because of no overlapping panels. Nor do γ_j's exist for $j \geq 12$ because for $j \geq 12$, there exist 2nd or higher order predecessors to the current panel and the correlation may be taken as 0 in the developments. Estimates of ρ_j, $\hat{\rho}_j$, are given in Table 2. The estimate of ρ_5 has been obtained by extrapolating other ρ_j's as it was not possible to estimate it directly from the sample. Note that $\hat{\rho}_j$ $(j = 1,2,...,5)$ is a decreasing function of j for all the five characteristics. This is consistent with what we expect intuitively about the behaviour of ρ_j's. Also ρ_j's are high for all the characteristics except "unemployed".

Table 3 gives the estimates $\hat{\gamma}_j$ of γ_j. The estimates $\hat{\gamma}_5$ and $\hat{\gamma}_{11}$ were obtained respectively by interpolating and extrapolating other $\hat{\gamma}_j$'s. Intuitively, we expect γ_j's to decrease with j for each characteristic. We observe that this is not the case with $\hat{\gamma}_j$'s. Although $\hat{\gamma}_j$'s do not exhibit monotonic decreasing behaviour, we point out that whenever the difference $(\hat{\gamma}_{j+1} - \hat{\gamma}_j)$ is positive, its magnitude is very small. The positiveness of these differences could be due to the sampling variability rather than a real positiveness of $(\gamma_{j+1} - \gamma_j)$.

In the following discussion, the term relative efficiency of AK composite (or K composite) estimator refers to its efficiency relative to the simple ratio estimator.

Tables 4A and 4B give the results of comparing the estimated variances of three estimators. These are: (i) optimal AK composite estimator, i.e., an estimator having minimum variance among the class of estimators defined by (2.2), (ii) optimal K composite estimator (obtained by taking A = 0 in (2.2) and having minimum variance among all estimators in this subclass), and (iii) the simple ratio estimator. For $0 \leq K < 1$, nearly optimal values of K and (K, A) are also given (K was incremented by 0.1 and the optimal value of A was determined for each fixed K). Here, a value (K, A) is referred to optimal value if the AK composite estimator with this value has the smallest variance among all AK composite estimators defined by (2.2). Similar definition applies to the term "optimal K". Table 4A (computed using $\hat{\gamma}_j$'s given in Table 3) shows that, for all characteristics except "unemployed" there are 18-21% gains in relative efficiency for the K composite estimates and 26-30% gains in the relative efficiency for the AK composite estimates.

To determine the effect of γ_j's on the relative efficiencies, γ_j's were replaced by zero's in the expression for $V(y_m')$ and the optimal K, optimal (K, A), and the relative efficiencies were computed. These results are presented in Table 4B. Note that the optimal K's and optimal (K, A)'s in the Tables 4A and 4B are different. Comparison of the corresponding relative efficiencies in these two tables shows that positive γ's have a negative effect on the reduction in variance, i.e., the gains in relative efficiency are reduced. The greatest reduction in relative efficiency is for the characteristic "employed agriculture". This is the characteristic with relatively high values of $\hat{\gamma}_j$'s. Thus taking γ_j's to be zero, when $\gamma_j > 0$, can result in over-estimation of the relative efficiencies and the degree of over-estimation depends on the magnitude of γ_j's.

As mentioned in the introduction, one of the drawbacks of the composite estimators of level is their bias as compared to the simple ratio estimator. Thus comparing the variances of biased estimators can sometimes result in erroneous conclusions about the relative performance of these estimators. It is appropriate to examine the mean square error in the case of biased estimators. The expression for the bias of y_m' (see Appendix I) involves α_i's (the rotation group biases). The quantity $\hat{\alpha}_i = y_{m,i} - \hat{Y}_m$ is an unbiased estimator of α_i if $\hat{Y}_m$ is an unbiased estimator of Y_m. We assume that the simple ratio estimator $\bar{y}_m$ is an unbiased estimator of Y_m, i.e., $\sum_{i=1}^{6} \alpha_i = 0$. Values of $\hat{\alpha}_i$ $(i = 1,2,...,6)$ for various characteristics are given in Table 5. For each of three characteristics "in labour force", "employed" and "employed non-agriculture", we note that: (i) $\hat{\alpha}_1$ is negative while all other $\hat{\alpha}_i$'s are positive; and (ii) $\hat{\alpha}_1$ is large relative to the other $\hat{\alpha}_i$'s.

Table 6 gives the values of optimal K, the optimal (K, A) and results of comparing mean square errors. The optimal K was determined among 10 values of K = 0(0.1)0.9 in the same manner for Tables 4A and 4B. However, the optimal (K, A) was computed in a different way. It was chosen among all possible combinations of K = 0(0.1)0.9 and A = 0(0.1)1.0 rather than determining optimal A for each fixed K = 0(0.1)0.9 (as used for Tables 4A and 4B). Two criteria of optimality are used. One is based on the concept of minimum variance (as is the case for Tables 4A and 4B), and the other is based on the concept of minimum mean square error.

It is shown in Appendix I that

$$E(y_m') = Y_m + [A\alpha_1 + K(\alpha_6 - \alpha_1)]/[5(1-K)].$$

Bias of each estimate in Table 6 is computed by using $\hat{\alpha}_1$ and $\hat{\alpha}_6$ (given in Table 5) instead of α_1 and α_6 in the above formula. Now we discuss the results of Table 6.

For the K composite estimate (based on minimum mean square error optimality) there is only a moderate gain in relative efficiency for the characteristic "employed agriculture" and a nominal gain for the characteristic "unemployed". Also, the bias of the estimates for these two characteristics is small. For the remaining characteristics, the simple ratio estimate is the optimal K composite estimate.

The K composite estimates (considered in Table 4A and based on minimum variance optimality) for the three characteristics "in labour force", "employed" and "employed non-agriculture" have relative efficiencies less than 10%. In these cases, the poor performance can be attributed to the large bias. For each of the remaining two characteristics, K composite estimate is only marginally better than the simple ratio estimate, i.e., the gain in relative efficiency is insignificant. The difference in the corresponding relative efficiency results in Tables 4A and 6 is due to the different relative efficiency definitions used for the two tables. For Table 4A, relative efficiency is defined as the ratio of appropriate variances whereas for Table 6, mean square errors are used instead of the variances.

The AK composite estimate (based on minimum mean square optimality) shows relative efficiency gains in the range 16-22% for all characteristics except "unemployed". Also, the bias of estimate for each characteristic is small.

However, the AK composite estimate based on minimum variance optimality, like the corresponding K composite estimate, has very low relative efficiency for the characteristics "in labour force", "employed", "employed non-agriculture" because of large bias in these cases. The gain in relative efficiency for the characteristic "employed agriculture" is moderate whereas the corresponding gain the characteristic "unemployed" is nominal.

The results in Table 6 show that, among the four composite estimates discussed above, the optimal AK composite estimates (based on minimum mean square error) have relative efficiencies higher for all characteristics than the corresponding relative efficiencies for other composite estimates. We will discuss later the results in the last column of Table 6.

We note, from the expression for $E(y_m')$ given earlier, the $y_m' - y_{m-1}'$ is an unbiased estimator of $Y_m - Y_{m-1}$, i.e., K or AK composite estimators of change are unbiased. Table 7 gives the optimal K, optimal (K, A), and relative efficiency results for optimal K composite and optimal AK composite estimates of change. The gains in relative efficiency for the characteristics "in labour force", "employed", and "employed non-agriculture" are in the 46-55%

range for K composite and AK composite estimates. For the characteristic "employed agriculture", the optimal AK composite estimate is also optimal K composite and the gain in relative efficiency is about 135%. The gain in relative efficiency for the characteristic "unemployed" is about 6% for both estimates.

It should be pointed out that the optimal value of K or (K, A) is characteristic dependent. Thus the additive property of the estimates is not preserved when different values of K or (K, A) are employed. To preserve additivity, a common value of K = 0.4 and A = 0.4 was selected for estimates of level and change. The following remarks describe the performance of the AK composite estimate with K = 0.4 and A = 0.4. The last column of Table 6 shows that the gains in relative efficiency for AK composite estimates of level are in the 6-10% range for all characteristics except "unemployed". The results of Table 7 show that the gains in relative efficiency for AK composite estimates of change are in the 12-15% range for all characteristics except "unemployed". The gain in relative efficiency for AK composite estimates of level and change is about 2-3% for the characteristic "unemployed".

ACKNOWLEDGEMENT

The authors wish to thank D. Drew, G.B. Gray and M.P. Singh of Statistics Canada for helpful discussions during the progress of this work, and the referee whose suggestions led to many improvements.

TABLE 1

Common and Predecessor Panels Pertaining To Months m and m-j

Panels in Month m	m-1	m-2	m-3	m-4	m-5	m-6	m-7	m-8	m-9	m-10	m-11
1	(6)	(5)	(4)	(3)	(2)	(1)	((6))	((5))	((4))	((3))	((2))
2	1	(6)	(5)	(4)	(3)	(2)	(1)	((6))	((5))	((4))	((3))
3	2	1	(6)	(5)	(4)	(3)	(2)	(1)	((6))	((5))	((4))
4	3	2	1	(6)	(5)	(4)	(3)	(2)	(1)	((6))	((5))
5	4	3	2	1	(6)	(5)	(4)	(3)	(2)	(1)	((6))
6	5	4	3	2	1	(6)	(5)	(4)	(3)	(2)	(1)

The correlation coefficients between common panels of months m and m-j indicated by panels with no parentheses equal ρ_j.

The correlation coefficient between panels of month m and their "single" predecessor of month m-j equals γ_j, the panels indicated by single parentheses.

The correlation coefficient between panels of month m and their double predecessor of month m-j equals $\gamma_j^{(2)}$, the panels indicated by double parentheses. In this report, all $\gamma_j^{(2)}$ are assumed to equal 0.

TABLE 2

Estimated Correlation ρ's (1980-1981 Ontario)

Characteristics $\hat{\rho}$	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$	$\hat{\rho}_4$	$\hat{\rho}_5$
In Labour Force	.843	.782	.717	.674	.631
Employed	.852	.779	.709	.664	.619
Employed Agriculture	.955	.926	.901	.861	.821
Employed Non-Agriculture	.861	.791	.724	.678	.632
Unemployed	.580	.445	.334	.286	.238

TABLE 3

Estimated Correlation γ's (1980-1981 Ontario)

Characteristics $\hat{\gamma}$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_3$	$\hat{\gamma}_4$	$\hat{\gamma}_5$	$\hat{\gamma}_6$	$\hat{\gamma}_7$	$\hat{\gamma}_8$	$\hat{\gamma}_9$	$\hat{\gamma}_{10}$	$\hat{\gamma}_{11}$
In Labour Force	.161	.141	.128	.133	.135	.136	.125	.127	.124	.122	.127
Employed	.164	.136	.142	.142	.146	.149	.148	.150	.153	.141	.148
Employed Agriculture	.477	.483	.474	.486	.480	.474	.459	.429	.394	.323	.252
Employed Non-Agriculture	.184	.150	.147	.157	.162	.167	.166	.169	.174	.156	.166
Unemployed	.141	.074	.076	.063	.057	.051	.045	.060	.077	.136	.074

TABLE 4A & 4B

The Optimal (K, A) and K, and the Relative Efficiencies of K Composite and AK Composite Estimators.

TABLE 4A

$\gamma_i \neq 0$

	K composite		AK composite		
Characteristics	Optimal K	Relative Efficiency	Optimal K	A	Relative Efficiency
In Labour Force	0.7	118.8	0.8	0.48	128.4
Employed	0.7	118.5	0.8	0.49	128.1
Employed Agriculture	0.8	120.6	0.8	0.38	126.9
Employed Non-Agriculture	0.7	119.4	0.8	0.47	129.3
Unemployed	0.3	102.8	0.5	0.38	105.2

TABLE 4B

$\gamma_i = 0$ for all i.

	K composite		AK composite		
Characteristics	Optimal K	Relative Efficiency	Optimal K	A	Relative Efficiency
In Labour Force	0.7	125.5	0.8	0.50	138.4
Employed	0.7	125.3	0.8	0.51	137.9
Employed Agriculture	0.8	167.3	0.9	0.46	187.9
Employed Non-Agriculture	0.7	126.9	0.8	0.49	140.2
Unemployed	0.4	104.4	0.6	0.51	108.4

Relative efficiency is with respect to the simple ratio estimator and is defined as 100 times the ratio V(simple ratio estimator)/V(K or AK composite).

TABLE 5

Estimates (in thousands) of Rotation Group Bias α_i

Characteristics	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$	$\hat{\alpha}_5$	$\hat{\alpha}_6$
In Labour Force	-135.3	39.8	41.1	31.1	15.4	7.9
Employed	-141.7	35.5	34.9	31.3	25.4	14.8
Employed Agriculture	-4.2	-2.6	2.2	-0.1	4.2	0.5
Employed Non-Agriculture	-137.5	38.0	32.7	31.3	21.2	14.3
Unemployed	6.4	4.3	6.2	-0.1	-9.9	-6.9

α_i is defined as $E(y_{m,i}) - Y_m$ and estimated by $y_{m,i} - \sum_{i=1}^{6} y_{m,i}/6$.

TABLE 6

Comparison of the Variances and the Mean Square Errors of Simple, K Composite, and AK Composite Estimators

IN LABOUR FORCE

		Simple Ratio Estimator	K Composite		AK Composite		Common K, A for all Characteristics
			(Min MSE) K=0	(Min Var) K=0.7	(Min MSE) K=0.7 A=0.7	(Min Var) K=0.8 A=0.5	K=0.4 A=0.4
Monthly Level Estimate	10^3	4480.7	4480.7	4547.5	4484.4	4527.6	4481.7
Variance	10^6	432.0	432.0	363.8	358.1	336.5	391.8
Bias	10^3	0	0	66.8	3.7	46.9	1.0
Mean Square Error	10^6	432.0	432.0	4284.5	371.5	2532.4	392.9
Relative Efficiency			100.0	9.0	116.3	17.1	110.0

EMPLOYED

			K=0	K=0.7	K=0.8 A=0.9	K=0.8 A=0.5	K=0.4 A=0.4
Monthly Level Estimate	10^3	4186.0	4186.0	4259.0	4183.6	4240.3	4188.0
Variance	10^6	473.3	473.3	399.6	397.7	369.5	428.9
Bias	10^3	0	0	73.0	-2.4	54.3	2.0
Mean Square Error	10^6	473.3	473.3	5732.2	403.2	3320.9	432.8
Relative Efficiency			100.0	8.3	117.4	14.3	109.4

EMPLOYED AGRICULTURE (TABLE 6 continued)

		Simple Ratio Estimator	K Composite		AK Composite		Common K, A for all Characteristics
			(Min MSE) K=0.6	(Min Var) K=0.8	(Min MSE) K=0.8 A=0.6	(Min Var) K=0.8 A=0.4	K=0.4 A=0.4
Monthly Level Estimate	10^3	142.0	143.4	145.7	143.2	144.1	142.1
Variance	10^6	85.7	75.6	71.1	68.7	67.6	80.8
Bias	10^3	0	1.4	3.7	1.2	2.1	0.1
Mean Square Error	10^6	85.7	77.6	85.1	70.2	71.8	80.8
Relative Efficiency			110.5	110.7	122.2	119.4	106.1

EMPLOYED NON-AGRICULTURE

			K=0	K=0.7	K=0.8 A=0.9	K=0.8 A=0.5	K=0.4 A=0.4
Monthly Level Estimate	10^3	4043.9	4043.9	4114.7	4041.6	4096.6	4045.8
Variance	10^6	498.9	498.9	417.8	418.0	385.9	452.8
Bias	10^3	0	0	70.8	-2.3	52.7	1.9
Mean Square Error	10^6	498.9	498.9	5436.1	423.3	3161.7	456.4
Relative Efficiency			100.0	9.2	117.9	15.8	109.3

UNEMPLOYED (TABLE 6 continued)

	Simple Ratio Estimator	K Composite		AK Composite		Common K, A for all Characteristics
		(Min MSE) K=0.2	(Min Var) K=0.3	(Min MSE) K=0.4 A=0.4	(Min Var) K=0.5 A=0.4	K=0.4 A=0.4
Monthly Level Estimate 10^3	294.8	294.1	293.7	293.9	293.2	293.9
Variance 10^6	117.5	114.9	414.3	112.5	111.7	112.5
Bias 10^3	0	-0.7	-1.1	-0.9	-1.6	-0.9
Mean Square Error 10^6	117.5	115.4	115.7	113.3	114.4	113.3
Relative Efficiency		101.9	101.6	103.7	102.7	103.7

Relative efficiency is relative to the simple ratio estimator and is defined by 100 times MSE(simple ratio estimator)/MSE(K or AK Composite estimator).

TABLE 7

Relative Efficiency of Composite Estimators
for Month-to-Month Change

Labour Force Characteristics	K composite		AK Composite		Common K, A K=0.4 A=0.4
	Optimal K	Relative Efficiency	Optimal K A	Relative Efficiency	Relative Efficiency
In Labour Force	0.9	146.6	0.9 0.1	147.9	113.3
Employed	0.9	151.0	0.9 0.1	152.3	114.1
Employed Agriculture	0.9	234.7	0.9 0.0	234.7	112.3
Employed Non-Agriculture	0.9	154.0	0.9 0.1	155.2	114.1
Unemployed	0.4	106.0	0.6 0.2	106.4	102.9

Relative efficiency is with respect to the simple ratio estimator and is defined as 100 times the ratio of appropriate variances.

APPENDIX I

Derivation of Bias of the Composite Estimator:

As defined in (2.2), the AK composite estimator of Y_m is given by:

$$y'_m = (1 - K + A)y_{m,1}/6 + (1 - K - A/5) \sum_{j=2}^{6} y_{m,j}/6 + K(y'_{m-1} + d_{m,m-1}). \quad (A1.1)$$

It may be noted that the simple ratio estimator now employed in LFS is the average of the six panel estimators and is given by:

$$\bar{y}_m = \sum_{i=1}^{6} y_{m,i}/6. \quad (A1.2)$$

From (2.3), the bias of the i-th panel estimator equals α_i so that:

$E(y_{m,i}) = Y_m + \alpha_i$, recalling that the bias is independent of m. Hence,

$$E(\bar{y}_m) = Y_m + \sum_{i=1}^{6} \alpha_i/6 = Y_m + \bar{\alpha} \text{ (say).}$$

In later developments we assume that $\bar{\alpha} = 0$.

The composite estimator may be rewritten as:

$$y'_m = y_m + K(y'_{m-1} + d_{m,m-1}), \tag{A1.3}$$

where

$$y_m = (1 - K + A)y_{m,1}/6 + (1 - K - A/5) \sum_{j=2}^{6} y_{m,j}/6$$

$$= (1 - K)\bar{y}_m + A(y_{m,1} - \bar{y}_m)/5.$$

Therefore

$$E(y_m) = (1 - K)(Y_m + \bar{\alpha}) + (A/5)(\alpha_1 - \bar{\alpha}). \tag{A1.4}$$

When $\bar{\alpha} = 0$, it simplifies to

$$E(y_m) = (1 - K)Y_m + (A/5)\alpha_1.$$

Using the definition of $d_{m,m-1}$ given in (2.1), we have

$$E(d_{m,m-1}) = E[\sum_{j=2}^{6} (y_{m,j} - y_{m-1,j-1})/5]$$

$$= (Y_m - Y_{m-1}) + (\alpha_6 - \alpha_1)/5. \tag{A1.5}$$

Now y'_m may be expanded by applying (A1.3) recursively and it is found that, up to n months back:

$$y'_m = y_m + Ky_{m-1} + K^2y_{m-2} + \dots + K^{n-1}y_{m-n+1} + K^n y'_{m-n}$$

$$+ Kd_{m,m-1} + K^2d_{m-1,m-2} + \dots. + K^n d_{m-n+1,m-n}. \tag{A1.6}$$

The expected value of y'_m may be readily obtained from (A1.4) and (A1.6) as below:

$$
\begin{aligned}
E(y'_m) &= (1 - K)\ [Y_m + KY_{m-1} + K^2Y_{m-2} + \dots + K^{n-1}Y_{m-n+1}] + K^nE(y'_{m-n}) \\
&\quad + [(1 - K)\ \bar{\alpha} + (A/5)(\alpha_1 - \bar{\alpha})](1 - K^n)/(1 - K) \\
&\quad + K(Y_m - Y_{m-1}) + K^2(Y_{m-1} - Y_{m-2}) + \dots + K^n(Y_{m-n+1} - Y_{m-n}) \\
&\quad + [(\alpha_6 - \alpha_1)/5]K(1 - K^n)/(1 - K) \\
&= Y_m + K^n[E(y'_{m-n}) - Y_{m-n})] \\
&\quad + [(1 - K)\bar{\alpha} + (A/5)(\alpha_1 - \bar{\alpha}) + K(\alpha_6 - \alpha_1)/5](1 - K^n)/(1 - K) \\
&= Y_m + K^n[E(y'_{m-n}) - Y_{m-n}] \\
&\quad + [(1 - K - A/5)\bar{\alpha} + (A/5)\alpha_1 + K(\alpha_6 - \alpha_1)/5](1 - K^n)/(1 - K) \qquad (A1.7)
\end{aligned}
$$

which simplifies for sufficiently large n and for the case $\bar{\alpha} = 0$ to

$$E(y'_m) = Y_m + [A\alpha_1 + K(\alpha_6 - \alpha_1)]/[5(1 - K)]. \qquad (A1.8)$$

Since the bias of y'_m under the model assumed in this paper is independent of m, the difference between composite estimates r months apart is unbiased, i.e.,

$$E(y'_m - y'_{m-r}) = Y_m - Y_{m-r} \text{ for all } r. \qquad (A1.9)$$

APPENDIX II

Derivation of the Variance of the Composite Estimator

We assume that the composite estimators (see (2.2)) have become sufficiently stable over time and hence we shall assume that $V(y'_{m-1}) = V(y'_m)$. Since all correlations 12 or more months apart are assumed to be zero, we shall assume that the LFS composite estimators have become stable after 12 months.

Taking the variance of both sides of (A1.3) and applying the above assumption; one may solve for $V(y'_m)$ to find that:

$$V(y'_m) = [V(y_m) + K^2V(d_{m,m-1}) + 2KCov(y_m, d_{m,m-1})$$

$$+ 2KCov(y_m, y'_{m-1}) + 2K^2Cov(d_{m,m-1}, y'_{m-1})]/(1 - K^2). \qquad (A2.1)$$

To eliminate y'_{m-1} on the right side of (A2.1), we apply (A1.6), replacing m by (m - 1) and n by 12 to obtain:

$$y'_{m-1} = \sum_{g=1}^{12} (K^{g-1}y_{m-g} + K^g d_{m-g,m-g-1}) + K^{12}y'_{m-13}. \qquad (A2.2)$$

Substituting (A2.2) in (A2.1) and dropping zero terms, we have

$$V(y'_m) = \{V(y_m) + K^2V(d_{m,m-1}) + 2KCov(y_m, d_{m,m-1})$$

$$+ 2 \sum_{g=1}^{12} K^g[Cov(y_m, y_{m-g}) + KCov(d_{m,m-1}, y_{m-g})$$

$$+ KCov(y_m, d_{m-g,m-g-1})$$

$$+ K^2Cov(d_{m,m-1}, d_{m-g,m-g-1})]\}/(1 - K^2). \qquad (A2.3)$$

We give the expressions for the variances and covariances on the right side of (A2.3), which may be readily derived by considering (2.1) and (A1.3).

$$V(y_m) = [(1 - K)^2/6 + A^2/30]\sigma^2, \qquad (A2.4)$$

which simplifies to $\sigma^2/6$ when A = K = 0; i.e.,

$$V(\bar{y}_m) = \sigma^2/6, \qquad (A2.4a)$$

the variance of the current LFS estimator.

$$V(d_{m,m-1}) = 2\sigma^2(1 - \rho_1)/5. \qquad (A2.5)$$

$$\text{Cov}(y_m, d_{m,m-1}) = (1 - K)(1 - \rho_1)\sigma^2/6 - A(1 - \rho_1)\sigma^2/30. \qquad (A2.6)$$

To derive the remaining covariances in (A2.3), which involve 'g', an indicator function I(a, b) shall be defined by:

$$\begin{aligned} I(a, b) &= 1 \text{ if } a \leq b \\ &= 0 \text{ otherwise.} \end{aligned}$$

By considering the definitions of y_m in (A1.3), $d_{m,m-1}$ in (2.1) and the corresponding expressions for month (m - g), one would find that the following covariances would be required to derive the remaining covariance of (A2.3).

$$\begin{aligned}
\text{Cov}(\bar{y}_m, \bar{y}_{m-g}) &= (\sigma^2/36)[(6 - g)\rho_g I(g, 5) + (6 - |\gamma - 6|)\gamma_g], \\
\text{Cov}(\bar{y}_m, y_{m-g,1}) &= (\sigma^2/6)[\rho_g I(g, 5) + \gamma_g I(6, g)], \\
\text{Cov}(y_{m,1}, \bar{y}_{m-g}) &= (\sigma^2/6)\gamma_g I(g, 6), \\
\text{Cov}(y_{m,1}, y_{m-g,1}) &= \sigma^2 I(g, 6)I(6, g)\gamma_6 \text{ (= 0 except when } g = 6), \\
\text{Cov}(y_{m-1,6}, \bar{y}_{m-g}) &= (\sigma^2/6)[\rho_{g-1}I(g, 6) + \gamma_{g-1}I(7, g)], \\
\text{Cov}(y_{m-1,6}, y_{m-g,1}) &= \sigma^2[\rho_{g-1}I(g, 6)I(6, g)], \\
\text{Cov}(\bar{y}_m, y_{m-g-1,6}) &= (\sigma^2/6)\gamma_{g+1}I(g, 5), \\
\text{Cov}(y_{m,1}, y_{m-g-1,6}) &= \sigma^2\gamma_1 I(g, 0)I(0, g) \text{ (= 0 for } g \geq 1). \qquad (A2.7)
\end{aligned}$$

The four covariances of (A2.3) that involve g may be readily defined and are found to be as follows:

$$\text{Cov}(y_m, y_{m-g}) = \sigma^2\rho_g I(g, 5)[(1 - K)^2(6 - g)/36$$

$$+ A(1 - K)(g - 3)/90 - gA^2/900]$$

$$+ \sigma^2\gamma_g[(1 - K)^2(6 - |g - 6|)/36$$

$$+ A(1 - K)(|g - 6| - 3)/90 - |g - 6|\ A^2/900]$$

$$+ \sigma^2\gamma_g I(g,\ 6)I(6,\ g)A(1 - K + A)/30, \qquad (A2.8)$$

$$\mathrm{Cov}(d_{m,m-1},\ y_{m-g}) = \sigma^2(\rho_g - \rho_{g-1})I(g,\ 5)[(1 - K)(6 - g)/30 + gA/150]$$

$$+ \sigma^2(\gamma_g - \gamma_{g-1})[(1 - K)(6 - |g - 6|)/30 + |g - 6|A/150$$

$$- (1 - K + A)I(g,\ 6)/30], \qquad (A2.9)$$

$$\mathrm{Cov}(y_m,\ d_{m-g,m-g-1}) = \sigma^2(\rho_g - \rho_{g+1})I(g,\ 5)(1 - K - A/5)(5 - g)/30$$

$$+ \sigma^2(\gamma_g - \gamma_{g+1})[(1 - K - A/5)(6 - I(6,\ g)$$

$$- |g - 6|)/30 + AI(g,\ 5)/25], \qquad (A2.10)$$

$$\mathrm{Cov}(d_{m,m-1},\ d_{m-g,m-g-1}) = \sigma^2[(5 - g)(2\rho_g - \rho_{g-1} - \rho_{g+1})I(g,\ 5)$$

$$+ (5 - |g - 6|)(2\gamma_g - \gamma_{g-1} - \gamma_{g+1})]/25. \qquad (A2.11)$$

Hence, $V(y'_m)$ can be expressed as $aA^2 + bA + c = f(A)$ where a, b and c are functions of K, ρ's and γ's. It can be shown that $a \geq 0$. The values of A that minimize the variance of AK estimator was determined for K = 0(0.1)0.9. Among these (A, K)'s, the optimal value of (A, K) was selected and is presented in Table 4A.

APPENDIX III

Derivation of the variance of $Y'_m - Y'_{m-1}$

From (A1.3)

$$y'_m = y_m + K(y'_{m-1} + d_{m,m-1}), \text{ or}$$

$$y'_m - Ky'_{m-1} = y_m + Kd_{m,m-1},$$

whence

$$(1 + K^2)V(y'_m) - 2K\mathrm{Cov}(y'_m, y'_{m-1}) = V(y_m) + 2K\mathrm{Cov}(y_m, d_{m,m-1}) + K^2V(d_{m,m-1}).$$

When $K \neq 0$, $\mathrm{Cov}(y'_m, y'_{m-1})$ may be obtained from the above and upon substitution of (A2.4), (A2.6) and (A2.5), and from the fact that $V(y'_m - y'_{m-1}) = 2V(y'_m) - 2\mathrm{Cov}(y'_m, y'_{m-1})$, one may find that for $K \neq 0$:

$$V(y'_m - y'_{m-1}) = \sigma^2[A^2/30 - (1 - \rho_1)KA/15 + (1 - K)^2/6 + (1 - \rho_1)K(K + 5)/15]/K - (1 - K)^2V(y'_m)/K. \quad (A3.1)$$

When $K = 0$,

$$y'_m = (1 - A/5)\bar{y}_m + Ay_{m,1}/5,$$

$$\mathrm{Cov}(y'_m, y'_{m-1}) = \mathrm{Cov}[(1 - A/5)\bar{y}_m + Ay_{m,1}/5, (1 - A/5)\bar{y}_{m-1} + Ay_{m-1,1}/5]. \quad (A3.2)$$

Thus for $K = 0$, we have:

$$V(y'_m - y'_{m-1}) = \sigma^2[(1/15 + \rho_1/450 + \gamma_1/90)A^2 + 2(\rho_1 - \gamma_1)A/45 + 1/3 - (5\rho_1 + \gamma_1)/18]. \quad (A3.3)$$

REFERENCES

[1] Bailar, B.A. (1975), "The Effects of Rotation Group Bias on Estimates from Panel Surveys", Journal of the American Statistical Association,

70, 23-29.

[2] Ghangurde, P.D. (1982), "Rotation Group Bias in the LFS Estimates", Survey Methodology, 8, 86-101.

[3] Gurney, M., and Daly, J.F. (1965), "A Multivariate Approach to Estimation in Periodic Sample Surveys", Proceedings of the Social Statistics Section, American Statistical Association, 242-257.

[4] Huang, E., and Ernst, L. (1981), "Comparison of an Alternate Estimator to the Current Composite Estimator in CPS", Proceedings of the Section on Survey Research Methods, American Statistical Association, 303-308.

[5] Lee, H. (1983), "Estimation of Panel Correlations in the Canadian Labour Force Survey", Technical Memorandum (in Preparation), Census and Household Survey Methods Division, Statistics Canada.

[6] Tessier, R. and Tremblay, V. (1976), "Findinqs on Rotation Group Biases", Technical Memorandum, Statistics Canada.

[7] Statistics Canada (1976), Methodology for the Canadian Labour Force Survey, Catalogue 71-526, Occasional.

THE PASSENGER CAR FUEL CONSUMPTION SURVEY

D. Royce[1]

The oil crisis of the mid-1970's triggered a new awareness among Canadians of the importance of energy conservation. The resulting government programs in the transportation sector demanded basic data about on-the-road fuel consumption by motor vehicles operating in Canadian conditions. This paper describes the Passenger Car Fuel Consumption Survey which was developed jointly by Statistics Canada and Transport Canada to meet this need. The methodology of the survey is described and some examples of the results are presented. The paper concludes with some speculation about future directions for the survey and for vehicle-usage statistics in general.

1. INTRODUCTION

1.1 The Need for Fuel Consumption Data

The world oil crisis of the 1970's triggered significant changes in energy policy in Canada. Although it is a net exporter of energy, Canada did not escape the effects of rapidly rising world oil prices and concerns about supply interruptions. By 1980, imports of foreign oil had reached 425,000 barrels a day, or one quarter of total Canadian oil consumption.

The transportation sector is the largest consumer of petroleum products. Transportation accounts for three out of every five barrels of oil consumed, with nearly four-fifths of this consumed by road motor vehicles. Consequently, energy conservation measures for automobiles have been a top priority. Among the actions which the federal government has taken are:

(a) the establishment of new car fuel consumption standards, with the goal of reducing automobile fuel consumption by 40 percent by 1990;

[1] D. Royce, Census and Household Survey Methods Division, Statistics Canada.

(b) the publication of The Fuel Consumption Guide and The Car Economy Book, intended to aid consumers in buying, driving and maintaining their cars to save energy;

(c) the encouragement of lower speed limits, which have now been implemented in all provinces

The evaluation of such programs, and the development of future government policy in the transportation energy field, require basic data about on-the-road fuel consumption by motor vehicles driven in Canadian conditions.

1.2 Development of the Fuel Consumption Survey

In mid-1977, Transport Canada approached Statistics Canada with a proposal to conduct a survey on the use of Canadian motor vehicles. Working closely with officials of Transport Canada, Statistics Canada developed detailed survey objectives and data requirements, identified operational problems that would have to be overcome, and laid out a strategy for the implementation of an ongoing survey by 1979. Because this was the first time such a survey had been attempted, Transport Canada and Statistics Canada agreed to limit it initially to passenger cars operated for personal use.

The three major objectives identified during preliminary discussions were: (a) to monitor both seasonal changes and long-term trends in fuel consumption and vehicle use, (b) to measure improvements in the fuel efficiency of new cars under actual operating conditions, and (c) to characterize the relationship between fuel consumption and vehicle characteristics (e.g., weight, number of cylinders and type of transmission), how the vehicle was maintained (e.g., tune-ups and maintenance of correct tire pressure), and how it was used (e.g., commuting versus long distance travel).

The remainder of the feasibility study was devoted to developing a tentative sample design and data collection methodology. Since no similar survey had been tried before, a period of pilot testing and refinement was required. In the fall of 1978, a series of tests was conducted to evaluate the sampling and data collection activities. The results of these tests were encouraging, and were used to refine the methodology for a full-scale survey carried out in seven provinces during the July-September 1979 quarter. The methodology was further refined during this quarter, and a regular cycle of data collection, processing and publication began for all ten provinces in the fourth quarter of 1979. Personal use light trucks and vans were added to the survey at the beginning of 1981.

Results are published in quarterly bulletins which contain basic tabulations of vehicles operated, distance travelled, fuel consumed, fuel consumption ratios, and average price per litre. More detailed tabulations and analyses are contained in an annual publication on the survey. Microdata tapes are also available, and special tabulations can be run on request.

2. METHODOLOGY OF THE SURVEY

2.1 Sample Design

The population for the survey is personal use passenger cars operated at any time during the survey reference month. Both privately owned and leased vehicles are included. A vehicle is classified as a "passenger car" on the basis of body style. A vehicle is "personal use" if it is operated for personal reasons at any time during the reference month, even though it may also be used for other reasons. Certain exclusions from the population are made for operational reasons. For example, the survey does not include vehicles from the Yukon or Northwest Territories. As well, new model vehicles are not included until they have been on the market for approximately one year. To avoid the difficulties that arise in locating the owners of very old cars, vehicles more than fifteen years old are excluded.

The sampling frame for the survey is constructed from the ten provincial motor vehicle registration files which are supplied specifically for the survey every three months. This approach is much less expensive than the use of a sample of households would be, and there is considerable vehicle data on the files that is very useful in the sample design. As well, all maintenance of the files is done by the provincial governments. The potential of the files as a source of statistics on vehicles not directly included in the survey is another point in their favour.

There are also several problems with using the registration files. For example, the information on the file may be several months out of date by the time fieldwork is underway. Although the files contain the most recent name and address of the registered owner, tracking down a specific vehicle still requires special procedures. The fact that most provinces now use a plate-to-owner system rather than plate-to-vehicle further complicates this tracing operation. Another problem is that some provinces do not supply all records. Some records may contain errors which are awaiting corrections, and even though these vehicles may be in use, the records are not available. Provinces also make changes to their systems and procedures fairly frequently. However, the co-operation of the provincial governments in resolving such problems has been excellent, and they have often made changes to accommodate the needs of the survey. Several provinces are now realizing the usefulness of their files for statistical as well as administrative purposes and are actively considering such needs when they redesign their systems.

Once the files have been reduced to a standard format, a computerized exclusion of vehicles not in the target population is made based on data on the files. The sample is drawn in two phases from the remaining vehicles. A relatively large first phase sample of vehicles is selected and the records are printed out. The owner's name is scanned visually and further exclusions (e.g., those with company names) are made. The vehicles that still remain are then subsampled (the second phase) and split into three random portions for use in the next three months of fieldwork. This two-phase design permits a reduction in sampling variance with very little extra work, and is in fact a special case of double sampling for stratification (Cochran, 1977). More importantly, it reduces the proportion of out-of-scope vehicles that must be handled during field operations. To reduce respondent burden, any vehicle which was in the sample within the previous year is dropped.

The sample is selected in the same manner in both phases. Vehicles are stratified by province, model year class, and either vehicle weight or wheelbase. Because separate data are required for each province and model year class (current year, all previous years), a disproportionate allocation is used. Each province is allocated an equal number of vehicles, and within each province about 40 percent of the sample is allocated to current model year vehicles. Within these two major stratification criteria, the allocation to sub-strata is proportional. Within each stratum, the file is sorted by postal code and a systematic sample of vehicles is selected, thus spreading the sample geographically as well.

A considerable amount of extra sample is selected and sent to the field in anticipation of a certain degree of out-of-scope vehicles being encountered, as well as to compensate for non-response. Whenever a non-response or out-of-scope vehicle is encountered, it is replaced by another vehicle from the same stratum. However, several attempts are made to get a response before a vehicle is replaced in order to avoid introducing more non-response bias than necessary.

2.2 Collection

Data collection for the survey is carried out in two steps. The first involves a telephone interview conducted with the registered owner approximately two weeks before the reference month. Because the sampling frame contains only a name and mailing address, it is necessary to trace the telephone number using telephone books, city directories, and other means. In cases where the vehicle has been sold, the new owner is traced if possible. In cases where the owner absolutely cannot be traced, either due to an unlisted number, the vehicle being sold, or the owner not having a telephone, the vehicle is replaced by another in the same stratum. In the case of leased vehicles, leasing companies are contacted by letter with a list of vehicles in the sample and then are followed up by telephone to request the name and telephone number of the lessees. Once the current owner or lessee is contacted, a screening interview containing questions on the type of vehicle, whether the vehicle will be used for personal reasons, and a few others, is administered. If the vehicle is eligible for the survey, the name and address of the principal driver (if different from the owner) are obtained, and permission to mail out a Fuel Purchase Diary is requested.

The Fuel Purchase Diary, which the principal driver keeps for the one-month reference period, is the second step in the data collection process. The diary itself is large and bright orange to encourage the driver to keep it in a visible location in the vehicle, and a heavy duty vinyl cover is used to prevent it from cracking during the winter. The respondent is also provided with a pen and a table with instructions on how to calculate his fuel economy. For each purchase, the respondent is asked to record the date, odometer reading, amount and type of fuel purchased, price per litre (or gallon), amount paid, and whether the purchase was made in the U.S.A. The respondent is also asked to fill the tank at the first and the last purchase so that the total fuel consumed can be calculated. The diary also contains a few questions asking for basic data on the vehicle, such as the number of cylinders, the type of transmission, and whether it has air-conditioning.

At the end of the reference month, the respondent mails the completed diary pages back to the Regional Office in a postage paid envelope. To improve response, a telephone reminder call is made during the first week of the reference month to answer any questions the respondent might have. After the end of the reference month, non-respondents are followed up several times by telephone to remind them to mail back the diaries or, if the diary was not kept, to determine the reason for non-response.

Table 1 shows an example of the response rates and eligibility rates for the survey, from the first quarter of 1981. During the telephone screening operation, interviews were attempted for 4,968 vehicle owners. Completed interviews were obtained with 3,626 owners for a response rate of 73.0%. The major reason for non-response at this stage was an untraceable telephone number for the registered owner. Of those respondents to the telephone screening, 2,921 were eligible to take part in the survey, a rate of 80.6%. The major reasons for non-eligibility were that the vehicle would not be used, or that it would not be used for personal reasons. Of those eligible, usable diaries were returned for 2,044, a response rate of 70.0% to this part of the study.

TABLE 1. Survey Response Rates and Eligibility Rates (1981 Quarter 1)

Telephone Interview Attempts	4,968	
Telephone Interviews Completed	3,626	(73.0%)
Vehicles Eligible for Survey	2,921	(80.6%)
Usable Diaries Returned	2,044	(70.0%)

2.3 Data Processing

At the present time, all data processing is carried out at Head Office. Prior to data entry, the completed screening questionnaires and diaries are groomed to improve legibility and to catch obvious errors (e.g., diaries with no purchases). In addition, the vehicle's curb weight is coded based on the Vehicle Identification Number, and response codes for both the telephone interview and the diary are added to the screening questionnaire. During data capture, the most important fields are verified on a 100 percent basis. After data entry, the screening questionnaires and diaries are edited separately.

The computer edit of the diaries checks for completeness, validity of codes, internal consistency of data, and reasonableness of derived data such as distance travelled and fuel consumed. The basic strategy in these edits is to use the computer to do mathematical calculations and to identify errors but to make corrections manually. This approach is used because the proper corrective action often cannot be determined without referring to the diary itself. The screening questionnaires are edited in a similar manner, except that the edits are much simpler. The edited diaries and screening questionnaires are then linked together prior to imputation and weighting.

Imputation is used at two points in processing. During the editing of the individual purchase records, if only the amount paid is present, the price is imputed as the average price for that type of fuel in that province in that month. The imputed price is then used to calculate the volume of fuel purchased. This procedure is used in less than one percent of all purchases. Imputation is also used after the screening questionnaire and the diary are linked. If the diary did not contain a pair of fillups sufficiently far apart to calculate the amount of fuel consumed, but the data are otherwise valid, the amount of fuel consumed is imputed from a regression model. The coefficients are calculated from the complete records for vehicles in the same quarter, with the predictor variables being distance travelled and vehicle weight. This procedure is used in 10 to 12 percent of all cases. The imputed records are identified on microdata and users are cautioned to omit these records when doing regression analysis involving fuel consumption.

Weighting of the data is straightforward. A final weight for each vehicle on the clean, edited file is obtained by multiplying together four factors:

1) the inverse of the first-phase probability of selection,
2) the inverse of the second-phase probability of selection,
3) the inverse of the response rate to the telephone interview, and
4) the inverse of the response rate to the diary.

Weights are calculated separately within each stratum. When the number of diaries within a stratum is small, or the response rates to the telephone interview or the diary are low, the stratum is collapsed with a neighbouring stratum before the weights are computed.

Although variance estimation would, ideally, allow for two phases of sampling, this would have required the design and programming of a customized computer system. Instead, it was decided to make the simplifying assumption that the sample had been drawn in one phase rather than two. As well, it was assumed that the sample was drawn by simple random sampling rather than systematically within each stratum. With these assumptions, it was possible to use MINICARP, a system already available in Statistics Canada, to carry out the calculation of sampling error estimates. Some modifications to the MINICARP system were made to allow it to handle large tables, and to improve the system's efficiency by eliminating calculations not needed by the survey.

2.4 Sources of Error

As in any survey, there are numerous sources of error, both sampling and non-sampling, which affect the data. This section discusses a few of the more important sources of non-sampling errors which are known or suspected to exist.

2.4.1 Coverage Errors

Because the sampling frame used for the survey is constructed from provincial files, the coverage of vehicles is highly dependent on the accuracy of these files. Experience has shown that the files can fluctuate considerably from quarter to quarter, as older vehicles are dropped or new vehicles are registered. Such fluctuations often, but not always, are reflected in the estimates of total vehicles, distance travelled and fuel consumption. Such fluctuations indicate that problems with coverage may exist. Some sources of undercoverage are known, such as those provinces which do not forward records with errors in them, but no method of adjusting for those sources has been developed.

When figures from this survey are compared directly to estimates of vehicle registrations (e.g., Statistics Canada 53-219), the survey estimates appear to be low. However, the concepts involved are quite different. The survey counts include only personal use passenger cars of certain model years which happen to be operated during a specific month, while the registration data refer to the number of vehicle registrations over a twelve-month period. As well, the same vehicle may be registered in more than one province when a person moves. Some attempts have been made to adjust for such differences and compare these two data sources, but the results have been inconclusive. In some provinces, the two sets of data appear very similar, while in other provinces the figures are far apart. Further work would be needed to develop methods for more accurate estimation of the level of coverage error. In the meantime, users are cautioned that estimates of level are subject to such errors.

2.4.2 Telephone Interview Non-response

Non-response to the telephone interview is another potential source of bias. A study conducted in early 1982 did show a mild, but not statistically significant, tendency for older cars to have higher non-response rates. Refusals by leasing companies to provide the name of the lessee occur more frequently than refusals by private individuals. A significant cause of non-response to the telephone interview is unlisted numbers, which tend to belong to persons in higher socio-economic classes. Exactly how potential sources of non-response bias affect the data is unknown, however. Ideally one would wish to follow up a sample of non-respondents to study their characteristics further.

2.4.3 Diary Non-response

As in the case of the telephone interview, very little is yet known of non-respondents to the diary. Comparisons of response rates for different sub-groups of the population have occasionally indicated that response does vary between provinces and between urban and rural areas. Again, however, one would have to conduct much more intensive follow up of non-respondents to make more precise statements on the nature of possible biases.

When the non-response rates to the telephone interview and the diary are taken together, it is evident that the overall response rate to the survey (not counting vehicles ineligible for the survey) is of the order of 50%. While this figure is comparable to many other surveys using a diary, it does leave considerable room for doubt about the reliability of the data. The problems with coverage and low response rates are a consequence of the methodology used, but at the same time any other approach would be prohibitively expensive. As often happens, then, the choice of a methodology is subject to considerations of both cost and data quality.

3. SOME RESULTS FROM THE SURVEY

Analysis of results from the survey to date has concentrated on a description of the vehicle fleet in operation, the distances travelled, the fuel consumed, and the fuel consumption ratio (fuel consumed per unit of distance travelled). These variables have been analyzed by vehicle characteristics such as model year, vehicle weight, number of cylinders and type of transmission, as well as by province and quarter.

One of the most interesting results is that the vehicle population in use is much younger than previously believed (Table 2). This table also shows the trend to greater use of newer vehicles, with average kilometres per vehicle declining with increasing age. Over 45% of vehicles operated and 50% of total kilometres driven are accounted for by the four most recent model years. About 90% of both vehicles operated and kilometres driven are accounted for by the nine most recent model years.

TABLE 2. Vehicles Operated, Kilometres per Vehicle, and Total Kilometres Driven, by Model Year (Reference Period October 1980 - September 1981)

Model Year	Vehicles Operated	Kilometres per Vehicle	Total Kilometres Driven (millions)
1980	574,300	20,404	11,718
1979	825,000	18,871	15,568
1978	945,900	15,735	14,884
1977	880,100	16,905	14,878
1976	891,800	17,671	15,759
1975	676,300	15,414	10,425
1974	637,300	12,252	7,808
1973	473,500	12,810	6,066
1972	361,700	12,569	4,547
1971 and previous	789,600	11,224	8,862

Turning to seasonal variations, Table 3 shows a peak in vehicle use during the summer quarter. About 33 percent more distance is travelled than in the winter quarter, and about 17 percent more than in the fall quarter. The fuel consumption ratio also changes with the season. On average, vehicles used 17 percent more litres per kilometre during the October to December period compared to the April to June quarter. Much of this difference is undoubtedly due to differences in climate, but the spring and summer months also contain a higher proportion of more fuel-efficient highway travel. The exact contribution of these two factors is unknown.

TABLE 3. Total Kilometres Driven and Fuel Consumption Ratio, by Quarter (Reference Period, October 1980 to September 1981)

	Total Kilometres (millions)	Fuel Consumption Ratio	
		l/(100 km)	MPG
January-March	23,059	17.4	16.2
April-June	30,468	15.1	18.7
July-September*	30,716	15.5	18.2
October-December	26,273	17.7	16.0

* Estimates for this quarter are based on data for only July and September due to a postal strike which occurred in August 1981. This may account for the Fuel Consumption Ratio being higher in the summer than in the spring.

Table 4 shows the improvements that have taken place in the fuel consumption ratio during the past decade. The ratio rose slightly between 1973 and 1975 with the introduction of stiffer pollution standards in those model years, but since then fuel efficiency has steadily improved.

TABLE 4. Fuel Consumption Ratio, by Model Year (Reference Period, October 1980 to September 1981)

Model Year	Fuel Consumption Ratio	
	l/(100 km)	MPG
1980	12.9	22.0
1979	14.3	19.8
1978	15.2	18.6
1977	16.5	17.1
1976	17.1	16.5
1975	18.7	15.1
1974	18.1	15.6
1973	18.1	15.6
1972	17.5	16.1
1971 and previous	19.6	14.4

Of all the factors examined, however, vehicle weight has the greatest impact on fuel consumption. Table 5 illustrates its effect: the heaviest cars consume more than twice as much fuel per kilometre as the lightest cars. In fact, the recent improvements in fuel efficiency shown in Table 4 have been achieved primarily by reductions in the average vehicle weight.

TABLE 5. Fuel Consumption Ratio, by Vehicle Weight
(Reference Period, October 1980 to September 1981)

Vehicle Weight	Fuel Consumption Ratio	
	1/(100 km)	MPG
Under 1000 kg	9.7	29.2
1000 to 1271 kg	13.3	21.3
1272 to 1544 kg	15.6	18.1
1545 to 1816 kg	18.1	15.6
1817 kg and over	20.4	13.8

Finally, regression analysis was used to look at the effect of vehicle characteristics and distance driven on the fuel consumption rate. Table 6 shows a typical result from the third quarter of 1981. The cumulative R-squared reaffirms the importance of the vehicle weight, but the age of the car and the number of cylinders also enter into the equation. An interesting finding is the negative coefficient for kilometres travelled. One hypothesis is that vehicles travelling longer distances tend to have better fuel efficiency because a higher proportion of their travel is on the highway. The R-squared value for the equation indicates that about 30 percent of the variation in the

TABLE 6. Multiple Regression Analysis of the Fuel Consumption Rate*
(Reference Period, July to September 1981)

Independent Variable	Regression Coefficient	Standard Error	Cumulative R-squared
ln (weight)	0.646	0.038	57.2
ln (distance)	-0.132	0.008	66.4
ln (age)	0.049	0.007	67.5
ln (cylinders)	0.245	0.033	69.0
Constant Terms	-1.578	0.232	

* The dependent variable was ln (Fuel Consumption Rate)

the fuel consumption rate remains unexplained.

4. FUTURE DIRECTIONS FOR THE SURVEY

Development of the survey was hindered for several years by its financial arrangements. With Transport Canada providing funding only on a year-to-year basis, personnel and other resources could not be permanently assigned to the project. Early in 1983, however, Transport Canada was able to make a three year commitment to the survey.

One of the first priorities under the new arrangement is to exploit the potential of the registration files as a source of statistics on motor vehicles. Descriptions of the entire motor vehicle fleet will be useful to planners in both governments and private industry. To assist in the analysis of these files, Statistics Canada has purchased the "VINA" system from R.L. Polk and Company. This is a computer system that both verifies the Vehicle Identification Number (or "serial number") and decodes it to give the make, model, weight, engine displacement, and other data. The system will also be useful in conducting the survey. Positive identification of vehicles during the sampling and the automated coding of vehicle data are two potential applications.

A second priority is to expand the coverage of the survey to other categories of vehicles and vehicle use. The major classes not covered by the present survey are commercially-used passenger cars, commercially-used light trucks and vans, and heavy trucks. A study to develop sampling and data collection methods for these vehicles began in the second half of 1983.

Several spin-off studies are also possible. A survey on vehicle maintenance was conducted in August 1983 and will be repeated in February 1984. Another study involves the oversampling of specific makes and models of vehicles. This would allow comparisons between laboratory measurements of fuel efficiency and measurements made under actual conditions of use.

Finally, there is a need for more information on the relationship between fuel consumption and specific trip characteristics. Very little is known, for example, about how total fuel consumption is split between commuting trips, shopping trips and business/commercial trips. The speed and distance of the trip, the type of roadway used, the weather conditions, and the number of passengers are a few of the other factors affecting fuel use.

Previous surveys, notably the National Driving Survey and the Canadian Travel Survey, have shown the viability of collecting detailed trip data from respondents. Unfortunately, the "fill-refill" method used in the present survey does not permit the measurement of the amount of fuel consumed for an individual trip. In order to do this, a vehicle would have to be equipped with an instrument, similar to an odometer, that accumulates the amount of fuel consumed. Until this happens, surveys of trip making and surveys of fuel consumption will likely continue to develop along separate lines.

The past ten years have seen a rapidly developing awareness of the importance of energy in all sectors of the Canadian economy. Although the world oil situation has changed as a result of the economic times, the need for reliable data on transportation energy use is an established fact. New fuels, new technologies, and new ways in which Canadians view energy use will all make their effects felt. The Fuel Consumption Survey will continue to provide important information for the shaping of future energy policy in Canada.

ACKNOWLEDGEMENTS

The author would like to thank Kevin Murphy, Wilma Shastry, Michel Fluet, Ken Sorrenti and Ray Ryan of Statistics Canada and John Lawson of Transport Canada for valuable assistance in preparing this paper.

REFERENCES

[1] Cochran, W.G., Sampling Techniques, 3rd edition, J. Wiley and Sons, New York, 1977.

[2] Energy, Mines and Resources Canada, "Fuel Consumption and Associated Trends in New Automobiles, 1975-1980", Technical Report TE-82-2, Ottawa, July, 1982.

[3] Energy, Mines and Resources Canada, The National Energy Program, Ottawa, 1980.

[4] Energy, Mines and Resources Canada, "Transportation Energy Data Book", Technical Report TE-82-5, Ottawa, December, 1982.

[5] Environment Canada, "Canadian Automobile Driver Survey", Economic and Technical Review Report EPS 3-AP-73-10.

[6] Lawson, J.J., "Canadian Department of Transport National Driving Survey, 1978-79", Accident Analysis and Prevention, Vol. 14, No. 5, pp. 371-380, 1982.

[7] Lawson, J.J., "Presentation on Fuel Consumption Survey Results", Transport Canada Technical Memorandum TMSE8301, Ottawa, March 1983.

[8] Lawson, J.J., "Recent National Surveys of Motor Vehicle Use in Canada", paper presented at the 2nd International Conference on New Survey Methods in Transport, Hungerford Hill, Australia, September 1983.

[9] Reilly-Roe, P., "Canada's Automobile Fuel Consumption Policies and Programs", Energy, Mines and Resources Canada Technical Report TE-81-41, Ottawa, 1981.

[10] Sorrenti, K. and Petherick, T., "Canadian Passenger Car Fuel Consumption Survey", in Consumers and Energy Conservation, Praeger Publishers, New York, 1981.

[11] Statistics Canada, "Canadian Travel Survey", Catalogue 87-504, Occasional.

[12] Statistics Canada, "Evaluation of the Pre-test, Automobile Fuel Consumption Survey", internal Statistics Canada report, Ottawa, February 1979.

[13] Statistics Canada, "Fuel Consumption Survey", Annual Report, non-catalogued publication, August 1983.

[14] Statistics Canada, "Passenger Car Fuel Consumption Survey", non-catalogued publication, Quarterly.

[15] Statistics Canada, "Quarterly Report on Energy Supply - Demand in Canada", Catalogue 57-003, Quarterly.

[16] Statistics Canada, "Report on Provincial Interviews, Automobile Fuel Consumption Survey", internal Statistics Canada report, Ottawa, August 1978.

[17] Statistics Canada, "Road Motor Vehicles - Fuel Sales", Catalogue 53-218, Annual.

[18] Statistics Canada, "Road Motor Vehicles - Registrations", Catalogue 53-219, Annual.

THE REGRESSION ESTIMATES OF POPULATION FOR SUB-PROVINCIAL AREAS IN CANADA[1]

Ravi B.P. Verma, K.G. Basavarajappa,
and
Rosemary K. Bender[2]

In order to improve the timeliness, accuracy and consistency of population estimates for different geographic areas, Statistics Canada has developed new methods of estimation for sub-provincial areas (census divisions and census metropolitan areas). Beginning with 1982, two sets of population estimates (regression and component based) will be published yearly, appearing 3-4 months and 12-15 months, respectively, from the reference date.

The regression technique uses family allowance recipients as the main symptomatic indicator and where available, additional indicators - reference population from provincial health insurance files and hydro accounts - to derive population change for the current year. The first set is obtained by adding this change to the second set for the previous year produced by the component method, with births and deaths from vital registers, and estimated migration from Revenue Canada taxation files. The two sets were found to be statistically similar with respect to accuracy, though the first set is more timely, and the second provides more details on the components of population change.

1. INTRODUCTION

Annual estimates of population for sub-provincial areas such as Census Divisions and Census Metropolitan Areas are useful for the planning of housing, hospitals, schools, colleges and universities and other social service programmes, studies of labour market areas, allocation of funds, computing vital

[1] The earlier version of this paper was presented at the Joint Statistical Meetings of the American Statistical Association, the Biometric Society, the Institute of Mathematical Statistics and the Statistical Society of Canada in Toronto, August 1983.

[2] Ravi B.P. Verma, K.G. Basavarajappa and Rosemary K. Bender, Demography Division, Statistics Canada.

rates, designing some surveys, computing the index of economic disparities, forecasting the number of tax payers etc. In particular, these estimates are required for weighting the results of Labour Force and Consumer Income and Expenditure Surveys and by the Department of Regional Economic Expansion, Revenue Canada and some provincial governments.

In order to improve the timeliness, accuracy, and to maintain the consistency among population estimates for different geographic areas, Statistics Canada has recently devised new methods for estimating the total population for census divisions and census metropolitan a eas. The objectives of this paper are to describe the post 1981 methodology for estimating the total population for these sub-provincial areas, discuss the accuracy of the methodology, review the work done by the provincial governments, and to discuss some factors which have played a vital role in the selection of some of these methods.

2. POPULATION ESTIMATION PROGRAMME FOR THE POST 1981 PERIOD

Beginning with 1982, for each year, Statistics Canada will publish postcensal population estimates for census divisions at two different periods. This is similar to the established practice for census metropolitan areas.

The first set which is based on a combination of regression model and component procedure and which provides no details on components of population change, will be published near the end of September of each year, i.e., 4 months from the reference date. Those estimates are termed regression-nested estimates (see Table 1).

As can be seen from Table 1, the first set of population estimates as of June 1, 1982 are the regression estimates, but for other years 1983 to 1986 they will be obtained by adding the change in the regression estimates to the second set of population estimates (obtained by the component procedure) for the preceding year. This approach ties together the two sets of the postcensal estimates for a specific reference date.

Specifications of the regression method by province, for census divisions and census metropolitan areas are presented in Table 2. For Census Divisions in each province, regressions will be based on the best available symptomatic indicators of population change namely, the number of family allowance beneficiaries aged 1-14, reference population taken from health insurance files (Saskatchewan and Alberta), and the number of hydro accounts (British Columbia). Regression models are primarily designed to maximize the accuracy of population estimates. For census metropolitan areas, the first set of population estimates for the post 1981 period will take input from the regression (Ratio-correlation) method using family allowance recipients aged 1-14 as the symptomatic indicator of population change. The form of regression as well as the variables selected are based on the criterion of minimum average absolute error (defined in Table 2) for alternate estimation methods for the 1976-81 period. These new methods are expected to result in estimates that are more reliable than those actually produced by alternate procedures for the intercensal years between 1976 and 1981.

The second set of estimates, produced using the component method, will provide details on each of the components of population change, and would be published about 12-15 months following the reference date. Birth and death data will be obtained from the vital statistics records, and the migration data will be derived from Revenue Canada Tax Files (Norris and Standish, 1983). The components of international migration derived from Revenue Canada Files, however, need further adjustment. For census divisions, this adjustment will be based on the immigration data eminating from the department of Employment and Immigration, and the independent estimate of emigration derived by Statistics Canada using the Family Allowance Files (Raby and Parent, 1982). For census metropolitan areas, no adjustment is needed for the immigration data, as they will be compiled from the intended destination of immigrants to CMA, from the department of Employment and Immigration. But the adjustment is applied to the estimates of emigrants which are derived as in the case of CDS.

For the first set of postcensal population estimates for census divisions by province, a further adjustment (prorating) is made to make them consistent with the corresponding provincial population totals. This is not necessary for the second set, as each of the components of population chanqe is already consistent with the corresponding provincial total. Similarly, an adjustment is made only to the first set of postcensal population estimates for census metropolitan areas.

3. EVALUATION OF ESTIMATION METHODS

The relative accuracies of the reqression method, the methods used during the period 1976-81, and the component method using migration data from Revenue Canada Tax Files are examined elsewhere (Norris, Britton and Verma, 1982). The accuracy is measured by comparing the 1981 estimates constructed from the 1976 base population with the enumerated totals as obtained from the census of 1981.

Methods of estimation are evaluated with respect to three criteria: (i) greater accuracy, (ii) timeliness and (iii) consistency. As mentioned above, accuracy of population estimates is measured by comparing the estimates with the census counts by computing the indices of average absolute error and index of misallocation. The error is defined as the difference between the estimate and the census count. Average absolute error is the arithmetic mean of percentage errors disregarding siqn (see Table 2). Index of misallocation is the index of dissimilarity, defined as half of the sum of absolute differences of the two sets of percent distribution of population estimates. Timeliness refers to the availability of estimates within as short a time as possible after the reference date. Consistency refers both to data consistency in the estimation of population being developed at various area levels of disaggregation using the same data source, and to uniformity in the methods of estimation. It must, however, be recognized that in some cases, the use of different data sources and different methods may be unavoidable.

3.1 Census Divisions

Relative accuracies of four sets of estimates as of June 1, 1981 obtained by four different methods for census divisions are presented in Table 3. It appears that each of the alternative methods (reqression, reqression-nested, and component) is superior to the old methods used during the period 1976-81. For Canada as a whole, among the alternative methods, regression-nested, derived by adding the change between two year regression estimates to the previous year component estimate, seems to be the most accurate with a lowest average absolute error of 1.7%. Betweeen regression and component methods, the regression-direct is observed to be less accurate. This is true in all provinces except the province of Saskatchewan in which the regression estimates are based on the reference population from health insurance files as the indicator of the population change. The accuracy of this indicator in estimating population is very high as indicated by a very low average absolute error, 1.43 percent. In 5 out of 10 provinces the regression-nested is slightly more accurate than the component method.

In order to assess the relative accuracy of each of the alternative methods, the paired t-test was also carried out. For Canada as a whole, it was found that the differences were statistically significant between the estimates obtained from the regression-direct and component method. This is also true in Ontario and Quebec. In contrast, the differences were not statistically significant between the regression-nested and the component method indicating that both these methods are statistically similar in terms of the accuracy. Similar results were observed when the t-test was performed on the weighted average of absolute errors which takes into account size of population.

3.1.1 Temporal Stability

In order to illustrate the temporal stability of three sets of postcensal estimates for census divisions (regression, component and regression-nested), the index of dissimilarity was computed for the years 1977 to 1981 and is presented in Table 4. It may be observed that both the disparities between the regression and component estimates (A) and the regression and nested estimates (C) increase over time. However, the disparity between the regression-nested and component estimates fluctuates slightly and is found to be minimum. Thus these two methods, in general, provide similar results during 1976 to 1981.

The component and regression methods are independent and so the results may be expected to diverge, whereas, the regression-nested and the component methods overlap and so the results tend to be similar.

The largest gap between the regression and the component estimates is not expected to fall, because there are some theoretical weaknesses inherent in the regression method. For example, the model may fit well for the previous time period, but may predict poorly during the succeeding period. The assumption in the regression method that the vector of regression coefficients for symptomatic indicators is invariant from the immediately preceding intercensal period to the postcensal period is often questionable. In practice, this invariance may not hold good over time, both because of structural changes in the underlying relationships of the variables, and also because of the improvement in the quality of the symptomatic indicators over time.

3.1.2 The Effects of Structural Changes

In order to examine the effects of structural changes on the differences between the 1976 and 1981 average errors, the 1981 average errors resulting from the equations of the model 1971-76 were compared with those resulting from the regression equations of the model period, 1976-81. It may be seen from Table 5 that the 1981 average errors resulting from the equations for two different time periods are quite comparable in all provinces except Saskatchewan, where the error declined by nearly 50% from 1.3% to 0.7%.

Due to structural changes, the relationship between the variations in symptomatic indicators (vital events and family allowance) and variations in population have undergone changes during the period, 1976-81. This is probably so for the provinces Quebec, Manitoba and Alberta. During the period 1976-81, the characteristics of the people moving from the eastern and maritime provinces to the western provinces may have changed considerably. For example, as the family allowances are limited to the families with children, movement of single persons and families without children were not captured by the changes in the family allowance indicator. Due to this, the family allowance recipients as an important predictor of the population change in the regression model, 1976-81 failed to predict adequately. Thus, it is clear that the average errors in 1981 resulting from models of both time periods, 1971-76 and 1976-81 were high, because of structural changes.

A part of the difference in the average errors between 1976 and 1981 is also due to changes in the quality of family allowance data. The numbers of family allowance recipients are produced at the census division level by converting postal codes to standard geographic codes. In 1976, the conversion file had problems of missing and overlapping postal codes. In particular the percentage of missing codes in 1976 was high in maritime provinces and Ontario. But, by 1981, the magnitude of the problem of missing postal codes in the FA files had declined in all provinces. Thus, the change in the quality of the family allowance indicator between the years 1976 and 1981 may have also affected the quality of the regression coefficients during the period 1976-81.

3.2 CMAs and the Non-CMA Unit

Table 6 presents the average absolute errors for CMAs according to three types of estimates, viz., regression-direct, nested and component. It may be seen that the component method provides estimates with the lowest errors at Canada level. The regression nested procedure comes second best. The same findings as for census divisions hold good when we consider the indices of dissimilarity which are given below:

Nested vs. Component	0.98%
Regression-direct vs. Component	1.15%
Regression-direct vs. Nested	1.09%

3.2.1 Consistency and Timeliness

In terms of the accuracy of population estimates and consistency with respect to sources of input data and methods used for estimating the population of different geographic areas, (provinces and territories, CDs and CMAs), the component method appears to be the most suitable. In addition, the component method provides more detailed and consistent information on components of population change, e.g., consistent set of internal migration figures classified by streams (in- and out-) and by broad age groups and sex for the province and its sub-provincial areas. However, this method does not provide timely estimates. The delay is expected to be about 12-15 months. The proposed regression method using family allowance recipients and/or other symptomatic indicators on the other hand, can provide estimates with a delay of about 3-4 months.

From Table 6, it may also be seen that in terms of the accuracy of the population estimates, the component and the regression-nested are closer to each other than regression-direct and component. But in terms of the timeliness, the regression-nested is superior to the component method.

4. ALTERNATIVE METHOD TESTED

The type of regression method by province shown in Table 2 is the most accurate for a given province among several alternative methods of estimation which were tested over the period 1976-81. These methods are: two types of component methods using migration estimates from school enrolment data and tax files, vital rates method, ratio method using the provincial administative files, proportional allocation method based on family allowance recipients, and six types of regression methods (ratio-correlation, weighted ratio-correlation, ridge weighted ratio-correlation, difference-correlation, weighted difference-correlation, and ridge weighted difference-correlation). Of these methods, the methods used for official estimates during 1976-81 include the component method using migration estimates from school enrolment, ratio method and ratio-correlation method (Dominion Bureau of Statistics, 1967).

Weighted regression method was adopted in order to control for heteroscedasticity. In this procedure, we transform the data set with the calculated weights such that one obtains a random error term (e) with constant variance.

We have used the Goldfield-Quandt procedure for testing the assumption of homoscedasticity (Johnston, 1963). Ridge regression controls for multicollinearity. In this procedure, estimates of β - coefficients are obtained by adding a small value K (.04) to the diagonal of the correlation matrix (X'X).

The accuracy of all these methods of estimation are thoroughly evaluated and the results presented in the three reports by Verma, Basavarajappa and Bender (1982a, 1982b, 1982c).

5. BACKGROUND HISTORY

In adopting the post 1981 methodology for estimating the population for sub-provincial areas, the following points were considered: the accuracy of the methods used during the period 1976-81, theoretical issues in the regression method, review of the work done by the provincial governments, two sets of official estimates for certain census divisions - one produced by Statistics Canada and the other produced by some provincial governments, consideration for a small area data development project and demand by other private users. A brief discussion of some of these points is given in the following paragraphs.

5.1 Review of Methodology Used during 1976-81 for Census Divisions and Census Metropolitan Areas.

Methods used during 1976-81 census divisions were specific to the provinces as presented in Table 3. These methods had many limitations (Verma and Basavarajappa, 1982a). These included the inadequacies of symptomatic indicators in capturing the current population changes (e.g., births and deaths), excessive time lag of about 2 years (due to delay in obtaining data on school enrolments) and some specification problems. The latter arose because in some provinces, particularly in those with large rural areas, school enrolment may not conform precisely to residential patterns due to transportation of children across census division boundaries. As a result of these limitations, the accuracy of the estimates for census divisions became unsatisfactory.

During the period 1976-1981, the component method was used to produce estimates for census metropolitan areas in Canada using births and deaths from Vital Statistitcs registers and Immigrants to CMAs from Employment and Immigration Department. The accuracy of the population estimates for CMAs was unsatisfactory primarily due to the weaknesses in the methodology for estimating emigration and internal migration for which no direct sources were available (Catalogue No. 91-207).

5.1.1 Ratio vs. Difference Correlation Methods.

Schmitt and Crosetti and many others have claimed that the ratio-correlation method is one of the most accurate methods (using as the criterion the Average Absolute Error - AAE) (Balakrishnan, 1960; Goldberg, Rao and Namboodiri, 1964; Swanson, 1978; N.R.C., 1980; Mandell and Tayman, 1982). Later, some researchers including Schmitt and Grier suggested that the difference-correlation method is an improvement over the ratio-correlation method (Schmitt and Grier, 1966; O'Hare, 1976). This was because the difference-correlation method produced constant mean, a lower mean square error (M.S.E.), higher intercorrelation among the variables, and a resulting higher squared value of the coefficient of multiple correlation (R^2). These features are often used to evaluate the fitting of a regression model and are considered desirable.

However, no consistent relationship between the higher R^2 and the average absolute error has been observed. The accuracy of population estimates produced by the regression method is highly dependent on the temporal stability of the regression coefficients. In this respect, a recent study has shown that the ratio-correlation method was more suitable than the difference-correlation method (Mandall and Tayman, 1982). The difference-correlation method produced a higher multi-collinearity than the ratio-correlation. Due to this, the difference-correlation shows higher instabibility in the regression coefficients over time-periods (Spar and Martin, 1979).

A review of both techniques has revealed that neither the ratio-correlation, nor the difference-correlation method uniformly or routinely outperforms the other (O'Hare, 1980). This was also confirmed by Verma, Basavarajappa and Bender (1982a).

In light of the above findings, a multiple-model frame work seems to be the most appropriate course for evaluating the competing estimation techniques. In fact, this is what has been employed in the present estimation programme.

5.1.2 Review of the Work done by Provincial Governments

A survey of provincial/territorial agencies producing population estimates and projections revealed that neither the methods nor the geographic divisions for which the estimates were produced were uniform. Some prepared estimates for census divisions and other areal units, non prepared estimates for census metropolitan areas.

To estimate populations of census divisions, or counties, one popular approach adopted by Ontario, Alberta and Northwest Territories is the component method described earlier.

The Northwest Territories obtain births and deaths from its Bureau of Vital Statistics. It estimates net migration with a time related cohort model for the population subgroup 1-14 years of age using family allowance recipients and school enrolment data.

Ontario also uses birth and death data from their Bureau of Vital Statistics. However, it estimates net migration from the changes of addresses from the drivers licence files of the Ministry of Transport.

Alberta uses a combination of two techniques. A ratio-correlation method estimates population change using births, school enrolment and the provincial health insurance plan as symptomatic indicators. Using the component method approach, the net migration is then obtained as a residual of the regression-based population growth, and births and deaths.

British Columbia, on the other hand, uses a combination of the difference-correlation method, with hydro billings, family allowance and vital statistics as symptomatic indicators, and the proportional allocation method. It is the only province that adjusts its subprovincial estimates so as to correspond to the provincial total published by Statistics Canada.

Quebec uses the best of several techniques to estimate their municipal populations. One method uses the figures provided by the municipalities and if found reasonable, these have priority over all others. Other methods use the rates of growth in hydro billings in combination with estimates/counts of the preceding year.

Newfoundland also estimates their communities using hydro billings as a source of input data. It combines this with preceding census counts, number of households, and average number of people per household.

Manitoba estimates the population of its municipalities in much the same way as Saskatchewan does at the provincial level. The count of elegible persons registered under its provincial medical health insurance plan, along with appropriate adjustment factors, is used to directly estimate the municipal populations in these two provinces.

However, no systematic evaluation of these estimates is available. Newfoundland, Quebec and Ontario are in the process of evaluating their estimates, British Columbia's evaluation of their estimates supports the continuation of their estimation methodologies for the post 1981 period.

The time lag after the reference date for which estimates become available ranges up to six months. Manitoba and Saskatchewan produce data within two months, the Northwest Territories, British Columbia and Ontario within four months and Quebec and Newfoundland within six months.

In conclusion, there is no uniformity of methods across the country. Each provincial/territory uses techniques that suit its particular needs, and which take advantage of provincial administrative data files.

5.1.3 Federal-Provincial Consultations

The new techniques devised for estimating the population of sub-provincial areas were discussed at meetings of the Federal-Provincial Committee on Demography. It is well to remember that the regression method was devised primarily for providing timely preliminary totals and the updating of these is firmly anchored in the component method. The question of the usefulnes of figures for CDs and CMAs is also worth considering. While the provinces need population estimates for municipal and administrative regions more than for CDs and CMAs, the latter are needed for Statistics Canada's internal uses and as building blocks for specified areas. Over the years, it has been observed that there has been a sufficient demand for estimates for CDs and CMAs. The lack of resources is also an important factor in preventing the extension of the estimation procedures for small sub-divisions of the provinces. Because of this, with some technical assistance from Statistics Canada, some provinces are planning to undertake the task of preparing population estimates at the municipal and other smaller divisions.

It may be noted that the above arrangement also avoids the duplication of efforts by the provincial and federal governments relating to the preparation of estimates for provinces and sub-provincial areas.

6. EVALUATIVE DISCUSSION

The research during the past year, carried out in collaboration with several provincial statisticians, resulted in the development of improved methods for estimating the population of census divisions and census metropolitan areas. As of 1982, for each year, Statistics Canada will publish two sets of post-censal population estimates for sub-provincial areas at two different periods.

The first set which is based on a regression model (and which refers to June 1 of each year) will be published near the end of September of each year, i.e., with a delay of utmost 4 months. The second set of estimates referring to the same date, produced by the component method using migration data derived from Revenue Canada Taxation Files, and the numbers of births and deaths from Vital

Registers, will be published about 12-15 months following the reference date.

These new methods are expected to result in estimates that are more reliable than those actually produced for the intercensal years between 1976 and 1981. These, more accurate and timely sub-provincial population estimates will be crucial to the Small Area Data Development Program that has just been launched by Statistics Canada.

It should be realized that the types of regression method that gave rise to a satisfactory pattern of error during 1976-81 for each province may turn out to be unsatisfactory during 1981-86, thereby giving rise to estimates with higher errors than anticipated. For example, on average, the regression model error was 2% in 1976 but when coefficients of the 1971-76 were applied to produce the estimates in 1981, the accuracy of the 1981 population estimates as compared to the 1981 census counts for census divisions was found to be 2.54%. Thus, we anticinate that the error as shown in Table 2 may increase by about 0.50 percentage points. However, the error in 1986 for the regression-nested estimate derived by adding the change in the regression estimates to that obtained by component method is expected to be very close to that of the component method.

One might argue that the practice of changing one set of estimates with another set of estimates for a specific reference date will have a negative impact on the planning for different social programmes. Also, if the two sets do not differ from each other very much, is there any need for producing both sets? The defense is that the first provided timely data and of acceptable quality, and the second, besides providing the relatively more detailed information on the components, provides estimates of acceptable and perhaps better accuracy.

ACKNOWLEDGEMENTS

We are thankful to Mr. Gordon Brackstone, Mr. Garnet Picot and Mr. Doug Drew of Statistics Canada, Dr. N.M. Lalu of University of Alberta, and the referee

as well as the assistant editor of this journal for many helpful comments.

TABLE 1

Methodology for the First Set of Population Estimates (Regression-Nested) for Census Divisions and Census Metropolitan Areas

Time	Regression Estimate	Component Estimate*	Regression-nested Estimate
t	P_t	P_t^c (census)	P_t
t+1	P_{t+1}	P'_{t+1}	P_{t+1}
t+2	P_{t+2}	P'_{t+2}	$P'_{t+1} + [P_{t+2} - P_{t+1}]$
t+3	P_{t+3}	P'_{t+3}	$P'_{t+2} + [P_{t+3} - P_{t+2}]$
t+4	P_{t+4}	P'_{t+4}	$P'_{t+3} + [P_{t+4} - P_{t+3}]$
t+5	P_{t+5}	P'_{t+5}	$P'_{t+4} + [P_{t+5} - P_{t+4}]$

* The method uses births and deaths from Vital Registration Records and migration data from Revenue Canada Taxation Files.

TABLE 2

Specifications of the Regression Method by Province for Estimating the Population Totals for Census Divisions and Census Metropolitan Areas, Post 1981 Period

Area/Province	Type*	Model Period	Symptomatic Indicator	Test 1981 AAE
Census Divisions				
Nfld. - P.E.I.	RC	1976-81	F	1.27
N.S.	RC	1971-76, 1976-81	F	1.50
N.B.	RC	1976-81	F	1.30
Quebec	RC	1976-81	F	1.81
Ontario	RC	1976-81	F	1.99
Manitoba	WDC	1971-76, 1976-81	F	3.13
Saskatchewan	DC	1976-81	CP	0.62
Alberta	WRC	1976-81	F, HC	1.89
B.C.	WDC	1971-76, 1976-81	F, Hydro	2.14
TOTAL				1.84
CMAs	RC	1976-81	F	2.30

Note: F: Family Allowance Recipients aged 1-14 years old.
CP: Covered Population.
HC: Health Care Files.

AAE: Average Absolute Error $= \frac{1}{N} \Sigma \left|\frac{E_i - P_i}{P_i}\right|$.

E_i: Estimated Population for Census Divisions.
P_i: Census Population for Census Divisions.
N: Number of Census Divisions with Province.
RC: Ratio-correlation.
WDC: Weighted-Difference correlation.
WRC: Weighted-Ratio-correlation.
DC: Difference-correlation.
CMAs: Census Metropolitan Areas.

* For a description of the types of regression methods, the readers are referred to the paper by W. O'Hare [10].

TABLE 3

Evaluation of Population Estimates, June 1, 1981
(Average Absolute Error)

Province	No. CDs.	Regression Direct (1)	Regression Nested	Component	Old Method Used (2)
NFLD. - P.E.I.	13	1.36	0.67	1.00	2.6
N.S.	18	1.64	1.27	1.07	6.8
N.B.	15	1.59	1.05	1.06	3.3
Quebec	76	3.10	1.63	2.02	2.5
Ontario	53	2.17	1.26	1.21	1.5
Manitoba	23	3.33	2.57	2.58	4.4
Saskatchewan	18	1.43	1.96	2.10	2.0
Alberta	15	4.45	2.84	2.39	5.1
B.C.	29	2.45	2.50	2.39	9.2
TOTAL	260	2.55	1.72	1.80	2.9

Notes: (1) The method uses as symptomatic variables reference population for Saskatchewan and family allowance recipients for other provinces.

The model period for all provinces in 1971-1976, using weighted ratio correlation for Alberta, weighted difference correlations for British Columbia, and ratio correlation for all other provinces.

(2) Methods used during 1976-81: Component II: Prince Edward Island, Nova Scotia, New Brunswick, Manitoba, Alberta and British Columbia.

Ratio Method: Ontario and Saskatchewan.

Ratio-correlation: Newfoundland and Quebec.

For a description of all these old methods, the readers are referred to the Statistics Canada Catalogue No. 91-206 [15].

TABLE 4

Temporal Stability of Three Sets of Postcensal Estimates for Census Divisions (Regression-direct, Regression-nested, Component) 1977-1981

Provinces		1977	1978	1979	1980	1981
NFLD.	A.	0.17	0.33	0.41	0.34	0.51
	B.	0.17	0.19	0.19	0.13	0.13
	C.	0.00	0.17	0.34	0.35	0.41
P.E.I.	A.	0.17	0.26	0.25	0.51	0.51
	B.	0.17	0.08	0.19	0.02	0.24
	C.	0.00	0.17	0.26	0.52	0.26
N.S.	A.	0.29	0.53	0.60	0.63	0.64
	B.	0.29	0.30	0.18	0.23	0.19
	C.	0.00	0.53	0.38	0.45	0.70
N.B.	A.	0.52	0.38	0.46	0.71	0.48
	B.	0.52	0.48	0.44	0.52	0.37
	C.	0.00	0.53	0.38	0.45	0.70
QUE.	A.	1.02	0.64	0.81	0.99	1.13
	B.	1.02	0.72	0.27	0.57	0.54
	C.	0.00	1.05	0.66	0.80	0.98
ONT.	A.	1.69	0.58	0.70	0.99	0.94
	B.	1.69	1.75	0.31	0.49	0.56
	C.	0.00	1.67	0.55	0.71	0.96
MAN.	A.	0.21	0.39	0.60	0.70	0.80
	B.	0.21	0.26	0.26	0.21	0.19
	C.	0.00	0.20	0.42	0.59	0.70
SASK.	A.	0.37	0.52	0.53	0.70	0.78
	B.	0.37	0.18	0.26	0.25	0.18
	C.	0.00	0.38	0.51	0.55	0.68
ALTA.	A.	0.45	0.45	0.57	0.89	1.18
	B.	0.45	0.21	0.27	0.41	0.36
	C.	0.00	0.44	0.43	0.56	0.86
B.C.	A.	0.39	0.45	0.76	0.95	0.93
	B.	0.39	0.32	0.41	0.23	0.29
	C.	0.00	0.37	0.43	0.76	0.94

Note: Index of dissimilarity between estimates E_1 and E_2 for a province with n census divisions and total population P is given by:

$$\frac{1}{2}\sum_{i=1}^{n}\frac{|E_{1i} - E_{2i}|}{P}$$

A: Index of dissimilarity between regression-direct and component estimates.
B: Index of dissimilarity between regression-nested and component estimates.
C: Index of dissimilarity between regression and regression-nested estimates.

Source: Demography Division, Statistics Canada, February 1983.

TABLE 5

Comparaison of the Accuracy of the Regression Methods for the Model Periods 1971-76 and 1976-81

	Regression		Model 1971-1976		Model 1976-81
	Type	Indicator	Test 1976 AAE	Test 1981 AAE	Test 1981 AAE
Nfld. - P.E.I.	RC	F	1.6	1.4	1.3
N.S.	RC	F	1.8	2.0	1.6
N.B.	RC	V, F	2.0	1.0	0.9
Quebec	RC	V, F	1.4	2.3	1.8
Ontario	RC	V, F	2.0	2.5	2.1
Manitoba	RC	F	1.9	3.3	3.5
Saskatchewan	RC	CP	1.5	1.3	0.7
Alberta	RC	F	3.1	4.6	4.2
B.C.	WDC	F	3.1	4.0	2.3
CANADA			1.96	2.54	2.04

Note: WDC: Weighted difference correlation.
RC: Ratio correlation with ordinary least square.
F: Family allowance recipients.
V: Vital events (Births + deaths).
CP: Covered population in Saskatchewan.
AAE: Average absolute error.

TABLE 6

Evaluation of 1981 Population Estimates
(CMAs and Non-CMA)

Method	Average Absolute Error (%)
Regression (F), (1971-76)	2.25
Regression-Nested	2.21
Component (Tax)	1.47

Note: F: Family Allowance Recipients Aged 1-14 years.

$$\text{Average Absolute Error} = \frac{1}{n} \Sigma \left|\frac{\text{Estimate - Census}}{\text{Census}}\right| \times 100.$$

n = Number of CMAs and non-CMAs.

REFERENCES

[1] Dominion Bureau of Statistics (1967). Population Estimates for Counties and Census Divisions, Catalogue No. 91-206, Ottawa: The Queen's Printer.

[2] Goldberg, D. and Balakrishnan, T.R. (1960). A Partial Evaluation of Four Estimation Techniques. Paper presented at the Annual Meeting of the Social Statistics Section, American Statistical Association.

[3] Goldberg, D., Rao, V.R. and Namboodiri, N.R. (1964). A Test of the Accuracy of the Ratio Correlation Population Estimates. Land Economics 40: 100-102.

[4] Government of British Columbia Central Statistics Bureau (1980). British Columbia School District Population Estimates, and also, British Columbia Municipal Population Estimates. Victoria, Ministry of Industry and Small Business Development, Government of British Columbia; unpublished report.

[5] Johnson, T. (1963). Econometric Methods, New York, McGraw-Hill Book Company, p. 219.

[6] Mandell, M. and Tayman, J. (1982). Measuring Temporal Stability in the Regression Models of Population Estimation. Demography, 19(1): 135-146.

[7] National Research Council (1980). Estimating Population and Income of Small Areas. Washington, D.C., National Academy Press. pp. 1-247.

[8] Norris, Douglas A., Britton, Malcolm and Verma, Ravi (1982). The Use of Administrative Records for Estimating Migration and Population. Presented at the Annual Meeting of the American Statistical Association, Cincinnati, Ohio, August 16-19.

[9] Norris, Douglas A. and Standish, Linda D. (1983). A Technical Report on the Development of Migration Data from Taxation Records. Ottawa, Statistics Canada. Draft.

[10] O'Hare, W. (1976). Report on a Multiple Regression Method for Making Population Estimates. Demography 13: 369-380.

[11] O'Hare, W. (1980). A Note on the Use of Regression Methods in Population Estimates. Demography 7: 87-92.

[12] Raby, R. and Parent, P. (1982). Postcensal Emigration Estimates 1980-1982. Ottawa, Statistics Canada, (Mimeographed). pp 1-29.

[13] Romaniuc, A., Raby, R. and Parent, P. (1982). The Choice of Methods for Estimating Interprovincial Migration for the Post-1981 Period. Ottawa, Statistics Canada, (Mimeographed). pp. 1-43.

[14] Spar, M. and Martin, J. (1979). Refinements to Regression-Based Estimates of Postcensal Population Characteristics. Review of Public Data Use 7 No. 5/6.

[15] Statistics Canada, Annual. Estimates of Population for Census Divisions. Cat. 91-206. Ottawa, Ministry of Supply and Services, Government of Canada.

[16] Statistics Canada, Annual. Estimates of Population for Census Metropolitan Areas of Canada. Cat. 91-207. Ottawa, Ministry of Supply and Services, Government of Canada.

[17] Swanson, D. (1978). An evaluation of Ratio and Difference Regression Methods for Estimating Small, Highly Concentrated Population: The Case of Ethnic Groups. Review of Public Data Use 6: 18-27.

[18] Verma, Ravi B.P. and Basavarajappa, K.G. (1982a). A Sub-provincial Estimation of the Population in Canada: A review of Estimation Methods and Prospects for Development. Presented at the Population and Family Planning Session, American Public Health Association Meeting, Montreal, November 1982. pp. 1-32.

[19] Verma, Ravi B.P., Basavarajappa, K.G. and Bender, Rosemary (1982b). New Approaches to Methods of Estimating the Population of Census Divisions. Ottawa, Statistics Canada, (Mimeographed). pp. 1-77.

[20] Verma, Ravi B.P., Basavarajappa, K.G. and Bender, Rosemary (1982c). New Approaches to Methods of Estimating the Population of Census Metropolitan Areas. Ottawa, Statistics Canada, (Mimeographed). pp. 1-39.

A BIBLIOGRAPHY FOR SMALL AREA ESTIMATION

With the growing demand for small area (small domain) estimates and the establishement of the Small Area Data Program within Statistics Canada, there is an increasing need to develop and evaluate methods for small area estimation. From research conducted within Statistics Canada and elsewhere, it is clear that there is no single best solution to the problem. Rather, for a particular application, the method to be selected from those available will depend on a variety of factors, including the availability of census, survey and administrative data.

The bibliography was developed by the Small Area Estimation Research team[1] primarily to assist persons in Statistics Canada engaged in research activities related to small area estimation techniques.

It represents an attempt to document both the range of techniques that have been used for small area estimation, and experiences with their use. It was built from references compiled by persons associated with this work and is undoubtedly incomplete.

This bibliography also represents a snapshot of an evolving duocument, to which other references will be added as time goes on. Readers are encouraged to bring to the attention of the members of the Small Area Estimation Research team any omissions in this document. Updated copies can be obtained from the Editor or persons involved in its development.

[1] The following persons are currently members of the Small Area Estimation Research team: Jean Dumais and David Paton, Institutions and Agriculture Survey Methods Division; Ravi Verma, Demography Division; Stephen Earwaker and Jean-François Gosselin, Census and Household Survey Methods Division; K.P. Srinath, Business Survey Methods Division.

Battese, G.E., and Fuller, W.A. (1981). Prediction of county crop areas using survey and satellite data. 1981 Proc. Surv. Meth. Res. Sec., American Statistical Association, 500-505.

Baxter, R., and Williams, I. (1978). Population forecasting and uncertainty at the national and local scale. Progress in Planning 4, Pergamon Press, Great Britain, 1978.

Ben-Akiva, M., and Lerman, S.R. (1975). Use of forecasting models in transportation planning. Conference on Population Forecasting for Small Areas, Oak Ridge, Tn., Oak Ridge Associated Universities, Inc., Tn.

Bender, R.K., and Verma, R.B.P. (1983). Translation for demographic data between overlapping subprovincial areas. Presented at the American Statistical Association meeting, Toronto.

Bogue, D.J. (1950). A technique for making extensive population estimates. JASA 45, 149-163.

Bogue, D.J., and Duncan, B.D. (1959). A Composite Method for Estimating Postcensal Population of Small Areas by Age, Sex and Colour. National Office of Vital Statistics. Vital Statistics -- Special Reports 47, No. 6.

Bousfield, M.V. (1977). Intercensal estimation using a current sample and census data. Review of Public Data Use 5, 6-15.

Box, E.P., and Tiao, G.C. (1973). Bayesian Inference in Statistical Analysis. Addison Wesley, Cambridge, Mass.

Bresee, J.C. (1975). Forecasting population for energy plant sites. Conference on Population Forecasting for Small Areas, Oak Ridge, Tn., Oak Ridge Associated Universities, Inc., Tn.

Cardenas, M., Craig, M.E., and Blanchard, M. (1979). Small area estimators: county crop acreage estimates using LANDSAT data. Review of Public Data Use 7, 23-28.

Carrol, S.J., Caggiano, M.N., McCarthy, K.F., Morrison, P.A., and Quint, B. (1980). City Data: a Catalogue of Data Sources for Small Cities. RAND Corp., Santa Monica, Ca.

Chambers, R.L., and Feeney, G.A. (1977). Log-linear models for small area estimation. Presented at the Joint Conference of the CSIRO Division of Mathematics & Statistics & the Australian Region of the Biometrics Society, Newcastle, Australia, Biometrics Abstract No. 2655.

Choudhry, G.H. (1979). Synthetic estimation for special areas at the economic region level. Unpublished memorandum, Census and Household Survey Methods Division, Statistics Canada.

Chu, S.F. (1974). On the use of the regression method in estimating regional population. International Statistical Review 42, 17-28.

Cohen, S.B. (1978). A modified approach to small area estimation. Synthetic Estimates for Small Areas, NIDA Research Monograph 24, 98-134.

Cohen, S.B., and Kalsbeek, W.D. (1977). An alternative strategy for estimating the parameters of local areas. 1977 Proc. Soc. Statist. Sec., American Statistical Association, 781-785.

Cohen, Reuben (1978). Drug Abuse Applications: Some regression explorations with National Survey Data. Synthetic Estimates for Small Areas, NIDA Research Monograph 24, 194-213.

Conopask, J.V. (1978). A Data-Pooling Approach to Estimate Unemployment Multipliers for Small Regional Economies. Economics, Statistics and Co-operatives Service, Washington, D.C.

Crosetti, A.H., and Schnitt, R.C. (1956). A method of estimating the intercensal population of counties. JASA 51, 587-590.

Delbert, J.E. (1981). A Technique for Estimating Age Specific Net Migration for Short-term Projections of Country Population. Population Research Center, University of Texas.

Deming, W.E., and Stephan, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. Annals of Mathematical Statistics 11, 427-444.

Deer, G.E. (1975). Forecasting population of small areas to faciliate state health-care planning and development. Conference on Population Forecasting for Small Areas, Oak Ridge, Tn., Oak Ridge Associated Universities, Inc., Tn.

Dominion Bureau of Statistics (1967). Population Estimates for Counties and Census Divisions, Cat. No. 91-206.

Drew, J.D., and Choudhry, G.H. (1979). Small area estimation. Unpublished memorandum, Census and Household Survey Methods Division, Statistics Canada.

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

Efron, B., and Morris, C. (1973). Stein's estimation rule and its competitors - an empirical Bayes approach. JASA 68, 117-130.

Efron, B., and Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. JASA 70, 311-319.

Elvers, E., Sarndal, C.E., Wretman, J.H., and Örnberg, G. (1983). Regression analysis for domains, a randomization theory approach. Unpublished.

Ericksen, E.P. (1975). Outliers in regression analysis when measurement error is large. 1975 Proc. Soc. Statist. Sec., American Statistical Association, 412-417.

Ericksen, E.P. (1974). A regression method for estimating population changes of local areas. JASA 69, 867-975.

Ericksen, E.P. (1974). Developments in statistical estimation for local areas. Census Tract Papers Series GE-40 No. 10, USBC, U.S. Government Printing Office, Washington, D.C.

Ericksen, E.P. (1973). A method of combining sample survey data and symptomatic indicators to obtain population estimates for local areas. Demography 10, 137-160.

Ericksen, E.P. (1973). Recent developments in estimation for local areas. Proc. Soc. Statist. Sec., American Statistical Association, 37-41.

Ericksen, E.P. (1971). A method of combining sample survey data and symptomatic indicators to obtain estimates for local areas. Unpublished Ph.D. thesis, University of Michigan, Ann Abor, Michigan.

Ervin, D.J. (1982). County population estimation with early release data: A new approach to data quality determination. Presented at the Population Association of America meeting, San Diego.

Espenshade, T., Hobbs, F.B., and Pol, L.G. (1981). An experiment in estimating postcensal age distribution of state population from death registration data. Review of Public Data Use 9, 97-114.

Fay, R.E. (1978). Some recent Census Bureau applications of regression techniques to estimation. Synthetic Estimates for Small Areas, NIDA Research Monograph 24, 155-184.

Fay, R.E., and Herriot, R. (1977). Estimates of income for small places: An application of James-Stein procedures to census data. JASA 74, 269-277.

Fay, R.E. (1974). Statistical considerations in estimating the current population of the U.S. Unpublished Ph.D. thesis, Department of Statistics, University of Chicago.

Feeney, G. (1983). A research proposal for the estimation of unemployment at the regional level. Unpublished, Australian Bureau of Statistics.

Ferguson, A.D. (1975). The necessity of forecasting for education planners. Conference on Population Forecasting for Small Areas, Oak Ridge, Tn., Oak Ridge Associated Universities, Inc., Tn.

Fisher, A., Fisher, W., and Starler, N. (1975). Small area economic analysis: Are Dunn and Bradstreet the answer? Northeast Regional Science Review 5, 270-276.

Ford, B.L. (1981). The Development of County Estimates in North Carolina. USDA Staff Report AGES 811119, Washington, D.C.

Ford, B.L., Bond, D., and Carter, N. (1983). Combining Historical and Current Data to Make District and County Estimates for North Carolina. USDA Staff Report AGES 8301_, Washington, D.C.

Ford, B.L., Bond, D., and Carter, N.J. (?). Research into small area estimation at the U.S. Department of Agriculture. Unpublished.

Freeman, D.H., and Koch, G.G. (1976). An asymptotic covariance structure for testing hypotheses on raked contingency tables for complex sample surveys. 1976 Proc. Soc. Statist. Sec., American Statistical Association, 330-335.

Froland, C. (1978). Synthetic Estimates as an Approach to Needs Assessment: Issues and Experience. Synthetic Estimates for Small Areas, NIDA Research Monograph 24, 246-258.

Fuller, W.A., and Battese, G. (1981). Regression estimation for small areas. Appendix G (pp. 572-586) in Rural America in Passage: Statistics for Policy, National Academy Press, Washington, D.C.

Ghangurde, P.D., and Gray, G.B. (1981). Estimation for small areas in household surveys. Communications in Statistics A10(22), 2327-2338.

Ghangurde, P.D., and Singh, M.P. (1978). Evaluation of efficiency of synthetic estimates. 1978 Proc. Soc. Statist. Sec., American Statistical Association, 52-61.

Ghangurde, P.D. (1978). Evaluation of synthetic estimation in self-representing areas of LFS. Technical Memorandum, Household Surveys Development Division, Statistics Canada.

Ghangurde, P.D., and Singh, M.P. (1977). Synthetic estimation in periodic household surveys. Survey Methodology 3, 152-181.

Ghangurde, P.D., and Singh, M.P. (1977). Evaluation of synthetic estimation in the LFS. Technical Report, Household Surveys Development Division, Statistics Canada.

Ghangurde, P.D., and Singh, M.P. (1976). Synthetic estimation in the LFS. Technical Report, Household Surveys Development Division, Statistics Canada.

Goldberg, D., Rao, V.R., and Namboodiri, N.R. (1964). A test of the accuracy of the ratio-correlation population estimates. Land Economics 40, 100-102.

Goldberg, D., and Balakrishnan, T.R. (1961). A partial evaluation of four estimating techniques. Department of Sociology, University of Michigan, Ann Arbor.

Gonzalez, M.E., and Hoza, C. (1978). Small area estimation with application to unemployment and housing estimates. JASA 73, 7-15.

Gonzalez, M.E. (1978). Case studies on the use and accuracy of synthetic estimates: unemployment and housing applications. Synthetic Estimates for Small Areas, NIDA Research Monograph 24, 142-154.

Gonzalez, M.E., and Hoza, C. (1975). Small area estimation of unemployment. 1974 Proc. Soc. Statist. Sec., American Statistical Association, 437-443.

Gonzalez, M.E. (1973). Use and evaluation of synthetic estimates. 1973 Proc. Soc. Statist. Sec., American Statistical Association, 33-36.

Gonzalez, M.E., and Waksberg, J.E. (1973). Estimation of the error of synthetic estimates. Presented at the first meeting of the International Association of Survey Statisticians, Vienna, Austria.

Government of British Columbia, Central Statistics Bureau (1980). British Columbia School District Population Estimates. Unpublished, Ministry of Industry and Small Business Development.

Government of British Columbia, Central Statistics Bureau (1980). British Columbia Municipal Population Estimates. Unpublished, Ministry of Industry and Small Business Development.

Gray, G.B., and Ghangurde, P.D. (1975). On a ratio estimate with post-stratified weighting. Survey Methodology 1, 134-144.

Greenberg, M.R. (1975). Use of the mathematical extrapolation methods to forecast population of small areas. Conference on Population Forcasting for Small Areas, Oak Ridge, Tn., Oak Ridge Associated Universities, Inc., Tn.

Greenberg, M.R., Krueckeber, D.A., Michaelson, C.O., Mautner, R., and Newman, N. (1975). Local Population and Employment Projection Techniques. The Center for Urban Policy Research.

Harms, L.T., James, R., and Springer, R.C. (1966). Projective Models of Employment by Industry and by Occupation for Small Areas: A Case Study. Temple University, Philadelphia, Pa.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953). Sample Survey Methods and Theory. John Wiley & Sons, Inc., New York.

Hawkins, D.M. (1980). Identification of Outliers. Chapman and Hall, New York.

Henry, L. (1976). Population Analysis and Models. Edward Arnold, London.

Herring, D.E. (1974). Procedures for preparing annual population estimates for counties and census divisions. Unpublished, Demography Division, Statistics Canada.

Holt, D., Smith, T.M.F., and Tomberlin, T.J. (1977). Synthetic estimation for small sub-groups of a population. Unpublished Technical Report, University of Southhampton, England.

Holt, D., Smith, T.M.F., and Tomberlin, T.J. (1979). A model based approach to estimation for small sub-groups of a population. JASA 74, 405-410.

Hughes, P.J., and Choy, C.Y. (1982). Regression Techniques for Local Government Area Population Estimation, Australian Bureau of Statistics Occasional Paper 1982/1.

Irwin, R. (1978). Aggregate medicare enrollment by age, sex and race as a resource in analysing demographic change for local areas. Prepared for presentation at the NBER Workshop on Population Analysis with Social Security Research Files.

Irwin, R. (1975). Methods and data sources for small areas. Conference on Population Forecasting for Small Areas, Oak Ridge, Tn., Oak Ridge Associated Universities, Inc., Tn.

Irwin, R. (1975). Use of the cohort-component method in population projections for small areas. Conference on Population Forecasting for Small Areas, Oak Ridge, Tn., Oak Ridge Associated Universities, Inc., Tn.

James, W., and Stein, C. (1961). Estimation with quadratic loss. Proceedings of the First Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, 361-379, University of California Press, Berkeley.

Jones, D.H., and Coopersmith, L.A. (1976). A ratio estimator of the total of a sub-population. Communications in Statistics A5(3), 251-260.

Jones, H.L. (1974). Jackknife estimation of functions of stratum means. Biometrika 61, 343-348.

Kalsbeek, W.D. (1973). A method for obtaining local postcensal estimates for several types of variables. Unpublished Ph.D. thesis, University of Michigan, Ann Arbor, Michigan.

Kendal, M.C. (1975). Labour market models. Conference on Population Forecasting for Small Areas, Oak Ridge, Tn., Oak Ridge Associated Universities, Inc., Tn.

Kish, L. (1965). Survey Sampling. John Wiley & Sons, Inc., New York.

Kish, L. (1980). Design and estimation for domains. The Statistician 29, 209-222.

Koch, G.G. (1973). An alternative approach to multivariate response error models for sample survey data with applications to estimators involving subclass means. JASA 68, 906-913.

Krebs, H.C., Feeney, D.T., Palit, C., Voss, P., and Kale, B. (1982). On measurement of error in estimating population change: a partial critique of the report of the panel on small area estimates of population and change. Presented at the Population Association of America meeting, San Diego.

Laake, P. (1978). An evaluation of synthetic etimates of employment. Scandinavian Journal of Statistics 5, 57-60.

Laake, P. (1979). A prediction approach to sub-domain estimation in infinite populations. JASA 74, 355-358.

Lalu, N.M. (1977). Monitoring Population Change in Postcensal Years, Edmonton Population Research Lab., University of Alberta.

Lee, E.S., and Goldsmith, H.F. (1978). Population estimates for small area analysis. Presented at the conference of the National Institute for Mental Health, Annapolis.

Levy, P.S. (1978). Small area estimation - Synthetic and other procedures. Synthetic Estimates for Small Areas, NIDA Research Monograph 24, 4-19.

Levy, P.S., and French, D.K. (1977). Synthetic estimation of state health characteristics based on the Health Interview Survey. Vital and Health Statistics Series 2, No. 75, U.S. Department of Health, Education & Welfare.

Levy, P.S. (1971). The use of mortality data in evaluating synthetic estimates. 1971 Proc. Soc. Statist. Sec., American Statistical Association, 328-331.

Lundstrom, S. (1983). Estimation for small domains: two studies using combined data from censuses, surveys and registers. Report to 1983 Meeting on Statistical Methodology, Statistical Commission for Europe, U.N. Economic Council.

Mandell, M., and Tayman, J. (1982). Measuring temporal stability in regression models of population estimation, Demography 19, 135-146.

Mandell, M., and Tayman, J. (1979). A test of the ratio-correlation methods of population estimation in a high growth state. Presented at the Southern Regional Demographic Group meeting, Myrtle Beach.

Martin, J.H., and Spar, M.A. (1981). Increasing the accuracy of postcensal country population estimates: the case for area specific methods. Presented at the Population Association of America meeting, Washington, D.C.

Martin, J.H., and Serow, W.J. (1978). Estimating demographic characteristics using the ratio-correlation method. Demography 15, 223-233.

McCleary, A. (1982). Small area population estimation: A note on the state of the art. Scottish Geographical Magazine 98(3).

Morrison, P.A. (1975). Forecasting population of small areas (an overview). Conference on Population Forecasting for Small Areas, Oak Ridge, Tn., Oak Ridge Associated Universities, Inc., Tn.

Morrison, P.A.(1982). Different Approaches to Monitoring Local Demographic Change, Rand Corporation Paper No. P.6743, Santa Monica.

Morrison, P.A., and Relles, D.A. (1975). A method for monitoring small area population changes in cities. Review of Public Data Use 3(2), 10-15.

Namboodiri, N.K. (1972). On the ratio-correlation and related methods of sub-national population estimation. Demography 9, 443-453.

Namboodiri, N.K., and Lalu, N.M. (1971). The average of several simple regression estimates as an alternative to multiple regression estimates in postcensal and intercensal population estimation: a case study. Rural Sociology 36, 187-194.

Nameketa, T. (1974). Synthetic state estimates of work disability. Unpublished Ph.D. thesis, University of Illinois, Champagne, Illinois.

National Center for Health Statistic (1980). Small Area Estimation: An Empirical Comparison of Convertional and Synthetic Estimators for States. D.H.E.W Pub. No. 80-1356, U.S. Government Printing Office, Wahsington, D.C.

National Center for Health Statistic (1978). State Estimates of Disability and Utilization of Medical Services, 1974-1976. D.H.E.W Publication No. (PHS) 78-1241, U.S. Government Printing Office, Washington, D.C.

National Center for Health Statistic (1977). State Estimates of Disability and Utilization of Medical Services, 1969-1971. D.H.E.W. Publication No. (HRA) 77-1241, U.S. Government Printing Office, Washington, D.C.

National Center for Health Statistics (1977). Synthetic Estimation of State Health Characteristis Based on the Health Interview Survey. D.H.E.W. Publication No. (PHS) 78-1349, U.S. Government Printing Office, Washington, D.C.

National Center for Health Statistics (1968). Synthetic State Estimates of Disability. P.H.S. Publication No. 1759, U.S. Government Printing Office, Washington, D.C.

National Research Council (1980). Estimating Population and Income of Small Areas, National Academy Press, Washington, D.C.

Nicholls, A. (1977). A regression approach to small area estimation. Unpublished, Australian Bureau of Statistics, Canberra, Australia.

Norris, D.A., Britton, M.; and Verma, R. (1982). The Use of Administrative Records for Estimating Migration and Population, Internal Revenue Service, Washington.

Office of Population, Censuses and Surveys (1981). The Revised Mid-1971 Population Estimates for Local Authorities Compared with the Original Estimate. Occasional Paper 22, Population Statistics Division, OPCS.

Office of Population, Censuses and Surveys (1980). Estimating the population of local authorities. Population Trends 20, 12-17.

Office of Population, Censuses and Surveys (1980). Local Authority Population Estimation Methodology, Occasional Paper 18, Population Statistics Division, OPCS.

O'Hare, W.P. (1980). A note on the use of regression methods in population estimation. Demography 17, 341-343.

O'Hare, W. (1976). Report on a multiple regression method for making population estimates. Demography 13, 369-379.

Pittenger, D., et al (1977). Making the housing unit population method work: a progress report. Presented at the Population Association of America meeting, St. Louis.

Platek, R. (1980). Small area estimation. Unpublished, Census and Household Survey Methods Division, Statistics Canada.

Promisel, D.M. (1978). Applications of synthetic estimates to alcoholism and problem drinking. Synthetic Estimates for Small Areas, NIDA Research Monograph 24, 63-87.

Purcell, N.J. (1979). Efficient small domain estimation. A categorical data analysis approach. Unpublished Ph.D. thesis, University of Michigan, Ann Arbor, Michigan.

Purcell, N.J., and Linacre, S. (1976). Techniques for the estimation of small area characteristics. Presented at the 3rd Australian Statistical Conference, Melbourne, Australia.

Purcell, N.J., and Kish, L. (1980). Postcensal estimates for local areas (or domains). International Statistical Review 48, 3-18.

Purcell, N.J., and Kish, L. (1979). Estimation for small domains. Biometrics 35, 365-384.

Pursell, D.E. (1970). Improving population estimates with the use of dummy variables. Demography 7, 87-91.

Raby, R. (1980). Intercensal estimates of the population of Canada and the provinces, census metropolitan areas and census divisions. Unpublished, Demography Division, Statistics Canada.

Rao. C.R., and Shirozaki, M. (1978). Precision of individual estimates in simultaneous estimation of parameters. Biometrika 65, 23-30.

Rao, J.N.K. (1984). Some thoughts on small area estimation. Unpublished.

Rasmussen, N. (1974). The use of driver license address change records for estimating interstate and intercounty migration. Presented at the Small Area Statistic Conference, St. Louis.

Rives, N.W. (1976). A modified housing unit method for small area population estimation. 1976 Proc. Soc. Statist. Sec., American Statistical Association, 717-720.

Rosenberg, H. (1968). Improving current population estimates through stratification. Land Economics 44, 331-338.

Royall, R.M. (1978). Prediction models in small area estimation. Synthetic Estimates for Small Areas, NIDA Research Monograph 24, 63-87.

Royall, R.M. (1977). Statistical theory of small area estimates - use of predictor models. Unpublished technical report presented to National Center for Health Statistics.

Royall, R.M. (1974). Discussion of papers by Gonzalez and Ericksen. Census Tract Papers Series GE-40, No. 10, U.S.B.C., U.S. Government Printing Office, Washington, D.C.

Royall, R.M. (1973). Discussion of two papers on recent developments in estimation of local areas. 1973 Proc. Soc. Statist. Sec., American Statistical Association, 43-44.

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

Sarndal, C.E. (1981). Design consistent versus model dependent estimation for small domains. Unpublished manuscript.

Schaible, W.L., Brock, D.B., Casady, R.J., and Schnack, G.A. (1979). Small area estimation: An empirical comparison of conventional and synthetic estimators for states. Public Health Service Series 2-82 (No. 80-1356), NCHS, D.H.E.W., U.S. Government Printing Office, Washington, D.C.

Schaible, W.L. (1978). Choosing weights for composite estimates for small area statistics. 1978 Proc. Soc. Statist. Sec., American Statistical Association, 741-746.

Schaible, W. (1978). A composite estimator for small area statistics. Synthetic Estimates for Small Areas, NIDA Research Monograph 24, 36-53.

Schaible, W.L., Brock, D.B., and Schnack, G.A. (1977). An empirical comparison of the simple inflation, synthetic and composite estimators for small area statistics. 1977 Proc. Soc. Statist. Sec., American Statistical Association, 1017-1021.

Schaible, Wesley L., Brock, D.B., and Schnack, G.A. (1977). An empirical comparison of two estimators for small areas. Presented at the Second Annual Data Use Conference of the National Center for Health Statistics, Dallas.

Schaible, W.L. (1975). A comparison of the mean square errors of the postratified, synthetic and modified synthetic estimators. Unpublished report, Office of Statistical Research, National Center for Health Statistics.

Schmitt, R.C., and Grier, J.M. (1966). A method of estimating the population of minor civil divisions. Rural Sociology 31, 355-361.

Schmitt, R.C., and Crosetti, A.H. (1954). Accuracy of the ratio-correlation method of estimating postcensal population. Land Economics 30, 279-280.

Serow, W.J., and Martin, J.H. (1977). Estimating demographic characteristics using the ratio-correlation method. Demography 15, 223-233.

Shryock, H.S., Siegel, J.S., and Assoc. (1975). The Methods and Materials of Demography. U.S. Government Printing Office, Washington, D.C.

Simmons, W.P. (1973). Adjustment of data-synthetic estimates. Presented at the first meeting of the International Association of Survey Statisticians, Vienna, Austria.

Singh, M.P., and Tessier, R. (1975). Some estimators for domain totals. Survey Methodology 1, 77-86.

Snow, E.C. (1911). The application of the method of multiple correlation to the estimation of postcensal population. Journal of the Royal Statistical Society 74, 575-620.

Spar, M.A., and Martin J.H. (1979). Refinements to regression based estimates of postcensal population characteristics. Review of Public Data Use 7, 16-22.

Starsinic, D.E. (1974). Development of population estimates for revenue sharing areas. Census Tract Papers Series GE-40, No. 10, U.S.B.C., U.S. Government Printing Office, Washington, D.C.

Starsinic, D.E., and Zitter, M. (1968). Accuracy of the housing unit method in preparing population estimates for cities. Demography 5, 475-484.

Statistics Canada (1979). Revised Annual Estimates of Population for Census Divisions. Catalogue 91-521.

Steahr, T.E., and Heston, J.F. (1983). Adjusted estimates of population by age and sex for towns in Connecticut, 1940-1980. Demographic Technical Paper 75-5, University of Connecticut.

Strohmenger, C. (1980). Estimates of population for census metropolitan areas. Unpublished, Demography Division, Statistics Canada.

Sturdevant, T.R. (1978). Using available resources to generate small area data. Presented at the American Statistical Association meeting, San Diego.

Swanson, D.A., and Tedrow, L.M. (1983). Improving the measurement of temporal change in regression models used for county population estimates. Presented at the American Statistical Association meeting, Toronto.

Swanson, D.A. (1980). Allocation accuracy in population estimates: an overlooked criteria with fiscal implications. Presented at the American Statistical Association meeting, Houston.

Swanson, D.A. (1980). Improving accuracy in multiple regression estimates of population using principles from casual modelling. Demography 17, 413-427.

Swanson, D.A. (1978). An evaluation of "ratio" and "difference" regression methods for estimating small, highly concentrated populations: the case of ethnic groups. Review of Public Data Use 6, 18-27.

Swanson, D.A. (1977). Preliminary results of an evaluation of the utility of ridge regression for making county population estimates. Staff Document 40, Office of Financial Management, State of Washington.

Tam, S.M. (1982). Postcensal estimates for local areas using current sample with census as the source of sampling frame. Int. Statist. Review 50, 125-134.

Tayman, J. (1980). Total and occupied housing units as useful symptomatic indicators in population estimation models. Presented at the Federal Population Estimates Cooperative Meeting, Denver, Colorado.

Terasvitra, T. (1981). Superiority comparisons of homogeneous linear estimators. Research Report No. 30, Department of Statistics, University of Helsinki.

Thibault, N. (1980). Population estimates for census divisions. Unpublished, Statistics Canada.

Théroux, G. (1979). Estimation of under-enumeration of the population in small areas. Unpublished, Census and Household Survey Methods Division, Statistics Canada.

Thompson, J.R. (1978). Some shrinkage techniques for estimating the mean. JASA 63, 113-122.

Treasury, Economic and Intergovernmental Affairs (Ontario) (1977). Demographic information from administrative files: recent developments in Ontario. Demographic Bulletin.

United Nations (1950). The Preparation of Sampling Survey Reports. Statistical Papers, Series C, No. 1.

U.S.B.C. (1981). Small area population estimates methods and their accuracy and new metropolitan area definitions and their impact on the private sector (GE-41, No. 6). U.S. Government Printing Office, Washington, D.C.

U.S.B.C. (1980). Population and per capita money income estimates for local areas: detailed methodology and evaluation. Current Population Reports, Series P-25, No. 699, U.S. Government Printing Office, Washington, D.C.

U.S.B.C. (1980). Small area statistics papers: An evaluation of small area data forecasting models and 1980 census small area statistics program (GE-41, No. 6). U.S. Government Printing Office, Washington, D.C.

U.S.B.C. (1977). Interrelationship among estimates, surveys and forecasts produced by federal agencies. Presented at the American Statistical Association Conference on Small Area Statistics.

U.S.B.C. (1975a). Population estimates and projection. Current Population Reports, Series P-25, No. 580, U.S. Government Printing Office, Washington, D.C.

U.S.B.C. (1974). Intercensal estimates for small areas and public data files for research. Presented at the American Statistical Association Conference on Small Area Statistics.

U.S.B.C. (1973). Federal State co-operative program for local population estimates: Test results April 7, 1970. Current Population Reports, Series P-26, No. 21, U.S. Government Printing Office, Washington, D.C.

U.S.B.C. (1970). Inventory of state and local agencies preparing population estimates, survey of 1969. Current Population Reports, Series P-25, No. 454, U.S. Government Printing Office, Wahsington, D.C.

U.S.B.C. (1969). Estimates of the population of counties and metropolitan areas, July 1, 1966: A summary report. Current Population Reports, Series P-25, No. 427, U.S. Government Printing Office, Washington, D.C.

U.S.B.C. (1966). Methods of population estimation: Part I, Illustrative procedure of the Census Bureau's component method II. Current Population Reports, Series P-25, No. 339, U.S. Government Printing Office, Washington, D.C.

U.S.B.C. (1949). Illustrative examples of two methods of estimating the current population of small areas. Current Population Reports, Series P-25, No. 20, U.S. Government Printing Office, Washington, D.C.

U.S.B.C. (1947). Population special reports. Series P-47, No. 4, U.S. Government Printing Office, Washington, D.C.

Verma, R.B.P., Basavarajappa, K.G., and Bender, R. (1983). The regression estimates of population for subprovincial areas in Canada. Presented at the American Statistical Association meeting, Toronto.

Verma, R.B.P., Basavarajappa, K.G., Bender, R.K., and Stepien, B. (1983). Generalized System for evaluation and production of total population estimates for sub-provincial areas. Unpublished, Statistics Canada.

Verma, R.B.P., and Basavarajappa, K.G., and Bender, R. (1982). New approaches to methods of estimating the population of census metropolitan areas. Presented to the Ad hoc Committee on Demography, Ottawa.

Verma, R.B.P., and Basavarajappa, K.G. (1982). Sub-provincial estimation of population in Canada: a review of estimation methods and prospects for development. Presented at the American Public Health Association meeting, Montreal.

Verma, R.B.P., Basavarajappa, K.G., and Bender, R. (1982). New approaches to methods of estimating the population of census divisions. Unpublished, Statistics Canada.

Woodruff, R.S. (1976). A simple method of approximating the variance of a complicated estimate. JASA 66, 411-414.

Woodruff, R.S. (1966). Use of a regression technique to produce area breakdowns of the monthly national estimates of retail trade. JASA 62, 496-504.

Young, M.E. (1975). Population Projections for Small Areas (A Bibliography with Abstracts). National Technical Information Service, Springfield, Va.

Zetter, M., and Shryock, H.S. (1964). Accuracy of methods of preparing postcensal estimates for states and local areas. Demography 1, 227-241.

ACKNOWLEDGEMENTS

The Editorial Board wish to thank Mrs. P. Pariseau, for her patient typing of the Journal and to Mr. D. Lemire for his service as liaison between the translation services and the production services. Thanks are also due to Mrs. Sheila Ranger, Printing Liaison Officer, Publishing Section, Production and Support Services for her help in the production of this Journal.

Acknowledgement is also due to H. Gough, N. Brien, D. Lemire, Y. Bélanger and A. Cyr for their proofreading of the French versions of papers.

Finally, the Editorial Board wish to thank the following persons who have served as referees during the past year.

M. Bankier
S. Cheung
B.N. Chinnappa
G.H. Choudhry
D. Drew
P. Foy
J.H. Gough
G.B. Gray
M.A. Hidiroglou
G. Hunter
J. Kovar
S. Kumar
M. Lawes
C. Patrick
D. Royce
A. Satin
K.P. Srinath
R. Sugavanam
P.F. Timmons
R.A. Veevers

SURVEY METHODOLOGY

1983 Vol. 9 No. 1

A Journal produced by Statistics Canada

CONTENTS