SSC Annual Meeting, July 1995
Proceedings of the Survey Methods Section

IMPACT OF INCOME STRATIFICATION ON THE EFFICIENCY OF INCOME ESTIMATES: AN EVALUATION

Barbara Chun[1]

ABSTRACT

The Survey of Consumer Finances (SCF) is a annual supplement to the Labour Force Survey (LFS) which generates estimates of average income and income distribution at various levels of aggregation. In order to improve the efficiency of income estimates, an additional level of stratification based on high income was introduced in the 1991 redesign of the LFS. A study was undertaken to evaluate the impact of this stratification by estimating the relative efficiency of the estimator for total income under the two designs, *i.e.*, with the additional level of stratification and without it. In this paper the methodology used for the study and results are presented.

RÉSUMÉ

L'Enquête sur les finances des consommateurs (EFC) est un supplément annuel de l'Enquête sur la population active (EPA), qui produit des estimations du revenu moyen et de la répartition du revenu à divers niveaux d'agrégation. Afin d'améliorer l'efficacité des estimations du revenu, on a ajouté un autre niveau de stratification basé sur les revenus élevés lors du remaniement de l'EPA en 1991. On a entrepris une étude afin d'évaluer l'impact de cette stratification grâce à une estimation de l'efficacité relative de l'estimateur du revenu total dans les deux plans de sondage, c'est-à-dire avec et sans le niveau supplémentaire de stratification. On présente dans cet article la méthodologie qui a servi à l'étude et ses résultats.

1. INTRODUCTION

The Survey of Consumer Finances (SCF) generates estimates of average annual income, annual income distributions, low income cut-offs, and incidence of low income for individuals and families. It is an annual supplement to the Canadian Labour Force Survey (LFS) with data collection occurring at the time of the April LFS interview. The SCF sample consists of approximately two-thirds of the LFS sample. Because of the close relationship between the SCF and LFS samples, the efficiency of the income estimates generated by the SCF relies greatly on the efficiency of the LFS design for income.

The LFS is a monthly household survey which generates information on labour market conditions. It has a stratified, multi-stage design with a sample of approximately 59,000 households. It produces estimates of unemployment, employment and labour force participation levels, and the corresponding rates, for various geographic areas.

The LFS is redesigned after every decennial census of population. One of the reasons for this redesign is to incorporate considerations of other surveys that use the LFS sample or frame. Other reasons may be the evolving needs of data users and changes in administrative boundaries. It has been observed (Chen *et al.*) that under the old design, the upper tail of the income distribution as estimated by the SCF differed considerably from that determined using tax data. It is possible that this difference may be due to conceptual differences between the tax data and SCF concepts, an inefficient SCF design for estimating the tail of the distribution, or both. The difference could also be attributable to high non-response or under-reporting of income for high income households. Moreover, the number of high income households falling in the SCF sample fluctuates widely from year to year.

These observations provided the motivation for adding an additional layer of stratification in the 1991 LFS redesign based on high income. It should be pointed out that estimates of average income and income distribution are unbiased under both post-1981 and 1991 designs. However, it is anticipated that the additional stratification in the current redesign will result in more efficient estimates of average income and income distribution. High income strata were introduced in nine Census Metropolitan Areas (CMAs): Montreal, Ottawa, Toronto, Hamilton, London, Winnipeg, Calgary, Edmonton, and Vancouver.

We want to study the impact of high income stratification on the efficiency of the SCF estimator for total income. To keep the study manageable we will examine only the three largest CMAs where high income strata were introduced. These are Vancouver, Toronto, and Montreal. In section 2 we describe the two designs which are compared. The methodology for the study and expression of variance are described in section 3, and

[1] Barbara Chun, Household Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6.

analysis and discussion of the results are given in section 4.

2. SAMPLE DESIGNS

Let D_1 denote the new LFS area sample design, *i.e.*, which includes high income strata, and D_2 the simulated area sample design without high income strata. The design D_1 was introduced over a six month period in 1994-1995. A brief description of the stratification and sample selection for these designs is described below. For more details on the LFS design see Mohl (1994), Singh *et al.* (1990), and Statistics Canada (1994).

2.1 Design D_1

The LFS has essentially a stratified multi-stage area sample design. Within the largest CMAs two separate sampling frames are established - a list frame of high rise apartment buildings and an area frame of all other dwellings. High income strata are created in the area frame. Clusters are created which are small geographic areas made up of a set of blockfaces and contain between 150 and 300 dwellings. These serve as first stage sampling units. Details of the creation of high and non-high income strata are given below.

2.1.1 Stratification

In the large CMAs, dwellings in the area frame (*i.e.*, excluding the apartment frame) are divided into high income and non-high income strata. The designation of high income strata for a CMA is based on the distribution of Census Enumeration Area (EA) average income. The EAs that fall in the upper 3% of the distribution are assigned to the high income stratum for that CMA. Note that the 3% rule is arbitrary. Using this rule, the average household income in the high income stratum in each CMA is over $100,000 and yields a sample of at least 24 dwellings. Clusters are created from blockfaces in the high income EAs. High income strata are formed first, then the rest of the area frame (non-high income) is further stratified into a number of strata based on a combination of geographic and optimization procedures as described below. For further details see Mohl (1994).

For the non-high income areas there are three possible levels of stratification: geo-, super-, and sub-stratification. The first level is geographically based where the CMA is divided into geo-strata which are made up of one or more Census Sub-Divisions (CSDs), usually municipalities. CSDs are made up of contiguous groups of Census Tracts (CTs) which are groups of approximately nine EAs. If the geo-strata are large then super-strata are formed within the geo-strata using an optimization algorithm (Drew *et al.*, 1985) which combines socio-economic and geographic variables resulting in compact and contiguous groups of Census Tracts (CTs) with similar socio-economic characteristics. Final sub-strata are non-compact and non-contiguous groups of CTs within super-strata formed by optimally grouping CTs with similar socio-economic characteristics. Thus the CTs are the basic units of stratification for the non-high income part of the area frame.

2.2 Design D_2

Under D_1, for a given CMA, the high income strata are comprised of all the high income clusters in the CMA. These clusters belong to various Census Tracts (CTs) in the CMA. One can think of the CTs in the non-high income strata as having "holes" in them which are the clusters identified as part of the high income stratum. We want to fill in these clusters and then restratify the area frame without high income strata. Recall that the CT is the basic unit for stratification. As described in section 2.1, three possible levels of stratification can be applied to the CMA. The first level is geographic, hence the geo-stratification of the CT given under D_1 would not be affected by the inclusion of the high income clusters. The next level of stratification is the super-stratum. This is a combination of geographic and optimal stratification with greater weight on the geographic aspect. For the purpose of this study, we will assume that at this level also the super-stratification of the CT given under D_1 would not change with the inclusion of the high income clusters. The third level of stratification is the formation of sub-strata and is based on an optimization algorithm which groups together CTs with similar socio-economic characteristics. These include characteristics such as size based on total dwelling count and CT income. This level of stratification may change from that given under D_1 due to the inclusion of high income clusters.

2.3. Sample Size

In order to ensure that the comparison of the two designs reflects only the impact of high income stratification, it is necessary that the sample sizes under both designs D_1 and D_2 be the same. Sampling ratios in the non-high income, non-apartment frame strata are the same within a CMA. However the sampling ratio in the high income strata may be different from that in the non-high income strata. We use a single sampling ratio for D_2 and set it equal to the sample size for D_1 divided by the population of the CMA. The sample sizes are then equal under the two designs.

2.4. Sample Selection

Sample selection procedures are the same for both designs. The sample is selected in two stages. At the first stage, the sampling unit is the cluster and is selected with probability proportional to size (PPS) using the Rao-Hartley-Cochran (RHC) random group method (Rao *et al.*, 1962). At the second stage of sampling, the sampling unit is the dwelling and is selected systematically.

3. METHODOLOGY

3.1. Approach

For the CMA of interest, let

Y = total income,

$\hat{Y}$ = estimator of Y

$V = V(\hat{Y})$ = variance of $\hat{Y}$, and

$\tilde{V}$ = estimator for $V(\hat{Y})$.

We will attach the subscript i to $\hat{Y}$, V, or $\tilde{V}$ whenever it refers to design D_i, i=1,2.

To compare the efficiencies of the two designs we compute the relative efficiency of D_1 with respect to D_2 denoted by R, and given by

$$R = \frac{V_2}{V_1}.$$

We replace V_1 and V_2 by their respective estimates

$\tilde{V}_1$ and $\tilde{V}_2$ and estimate R by

$$\tilde{R} = \frac{\tilde{V}_2}{\tilde{V}_1}.$$

If $\tilde{R} > 1$ then the design D_1 is more efficient than design D_2. If $\tilde{R} < 1$ then D_2 is more efficient than D_1 and if $\tilde{R} = 1$ then there is no difference between the designs in terms of efficiency.

We estimate the unknown population parameters in the expression of variance, $V(\hat{Y})$, using the 1991 Census 2B sample which is a 20% sample and which collects, among other variables, income data.

3.2. Variance Expression

Let Y be the total CMA income and $\hat{Y}$ its estimate. The estimator used in the SCF design is the regression estimator (Lemaître and Dufour, 1987). For estimation at the CMA level, *i.e.*, with only one auxiliary variable, X, the regression estimator is equivalent to the combined ratio estimator and is given by

$$\hat{Y}_c = (\hat{Y}/\hat{X})\ X = \hat{R}_c\ X$$

where $\hat{R}_c = \hat{Y} / \hat{X}$ is the ratio at the CMA level.

It is well known (Cochran, 1977) that the variance of $\hat{Y}_c$ can be approximated by

$$V(\hat{Y}_c) \doteq V(\hat{Y}-R_c\hat{X}) = V(\hat{U}) = \sum_h V(\hat{U}_h)$$

where $\hat{U}_h = \hat{Y}_h - R_c\ \hat{X}_h$ and $\hat{U} = \sum_h \hat{U}_h$.

The following notation will be used in this section:

y_{hjk} = income of household k in cluster j of stratum h,

x_{hjk} = target population size (15+, no military) of household k in cluster j of stratum h,,

$u_{hjk} = y_{hjk} - R_c\ x_{hjk}$ = value of variable u for household k in cluster j of stratum h,

z_{hj} = size measure of cluster j in stratum h,

Z_h = sum of the size measures over clusters in stratum h,

M_{hj} = dwelling count in cluster j in stratum h,

N_h = total number of clusters in stratum h,

N_{hg} = total number of clusters in group g of stratum h, and

$$A_h = \frac{\sum_{g=1}^{n_h} N_{hg}^2 - N_h}{N_h(N_h-1)}$$

Note that the size measure, z, is the design count, *i.e.*, based on the 1991 Census, and M is the actual count for the survey period. In the case of this study, z and M are the same, as both refer to the 1991 Census day.

Using a tilde (~) to denote a Census sample based estimate, it is shown (Laniel and Mohl, 1994) that an unbiased estimator of the variance of $\hat{U}_h$ is given by

$$\tilde{V}(\hat{U}_h) = A_h\ [\sum_{j=1}^{N_h} \frac{\tilde{U}_{hj}^2 - \tilde{V}(\tilde{U}_{hj})}{z_{hj}/Z_h} - (\tilde{U}_h^2 - \tilde{V}(\tilde{U}_h))] + \sum_{j=1}^{N_h} [W_h - 1 - A_h(Z_h/z_{hj} - 1)] M_{hj} \tilde{S}_{hj}^2$$

when assuming that systematic sampling at the 2nd stage of selection is approximately equal to SRS.

The variance estimate over all strata is then given by

$$\tilde{V} = \sum_h \tilde{V}(\hat{U}_h).$$

Population cluster sizes, dwelling counts, and the number of clusters are known from the 1991 Census. Estimates of the cluster total income, $\tilde{U}_{hj}$, and cluster population variance for income, $\tilde{S}_{hj}^2$, are obtained from the Census 2B sample. The estimates are given by

$$\tilde{U}_{hj} = \frac{M_{hj}}{C_{hj}} \sum_{k=1}^{C_{hj}} u_{hjk}$$

and

$$\tilde{S}_{hj}^2 = \frac{1}{C_{hj} - 1} \sum_{k=1}^{C_{hj}} (u_{hjk} - \bar{u}_{hj})^2$$

where $\bar{u}_{hj} = \frac{1}{C_{hj}} \sum_{k=1}^{C_{hj}} u_{hjk}$ and C_{hj} is the Census 2B sample count of households in cluster j of stratum h.

The stratification of the clusters under D_1 is given by the new LFS design. Under D_2, it is simulated as described in section 2.2. Although the expression for the variance estimator, $\tilde{V}$, is the same under the two designs, the stratification of the clusters will be different. The value of $\tilde{V}$ calculated under the two designs D_1 and D_2 gives the estimate of variance, *i.e.,* $\tilde{V}_1$ and $\tilde{V}_2$ respectively.

4. ANALYSIS AND DISCUSSION

Table 1 presents the relative efficiency of the design D_1 with respect to D_2 for the three CMAs studied.

Table 1: Efficiency of Design D_1 Relative to D_2 for Total Income

CMA	$\tilde{R} = \tilde{V}_2/\tilde{V}_1$
Vancouver	1.18
Toronto	1.18
Montreal	1.17

We note that $\tilde{R} \approx 1.18$ for the three CMAs, *i.e.,* the design D_1 is approximately 18% more efficient than D_2. As 1991 income data were used for both the stratification variable and the study variable, these gains may be slightly higher than one would get in actual practice. It is well known that the efficiency of a design tends to decrease over time. The efficiencies of both designs will be affected. How the changes in the design efficiencies will affect the ratio $\tilde{R}$ depends on the relative rate at which they change. However, over a short period of time the income distribution is relatively stable and hence it is reasonable to expect similar efficiency gains.

Other areas of investigation that are of interest are: (1) to study the efficiency gains due to income stratification when 1996 Census data becomes available, (2) to investigate the use of auxiliary information at the estimation stage, and (3) to study the impact of high income stratification on the efficiency of estimates of income distribution.

ACKNOWLEDGEMENTS

I would like to thank Edward Chen, S. Kumar, Normand Laniel, John Lindeyer, and Chris Mohl for their guidance and assistance during the progress of the study and G.H. Choudhry for his valuable comments during the preparation of this paper.

REFERENCES

Chen, E.J., Gambino, J., Laniel, N., and Lindeyer, J. (1994). "Design and Estimation Issues for Income in the Canadian Labour Force Survey", HSMD internal report, Statistics Canada, Ottawa.

Cochran, W.G. (1977). *Sampling Techniques* (3rd. ed.), New York: John Wiley & Sons, Inc.

Drew, D., Bélanger, Y., and Foy, P. (1985). "Stratification in the Canadian Labour Force Survey", *Survey Methodology,* 11, 95-110.

Laniel, N., and Mohl, C. (1994). "Analysis of Urban Cluster Size in the Canadian Labour Force Survey", *American Statistical Association, Proceedings of the Section on Survey Research Methods,* Vol. II, 931-936.

Lemaître, G.E., and Dufour, J. (1987). "An Integrated Method for Weighting Persons and Families", *Survey Methodology,* 13, 199-207.

Mohl, C. (1994). "Stratification of the Area Frame in SR and NSR Areas for the Post-Censal LFS Design". HSMD, internal report, Statistics Canada, Ottawa.

Rao, J.N.K., Hartley, H.O., and Cochran, W.G. (1962). "A Simple Procedure for Unequal Probability Sampling without Replacement", *Journal of the Royal Statistical Society,* B, 24, 482-491.

Singh, M.P., Drew, J.D., Gambino, J.G., and Mayda, F. (1990). *Methodology of the Canadian Labour Force Survey,* Statistics Canada Catalogue 71-526.

Statistics Canada (1994). "Upcoming Changes to the Labour Force Survey (1994)", Labour Force Survey Sub-Division, Household Surveys Division, internal report, Statistics Canada, Ottawa.