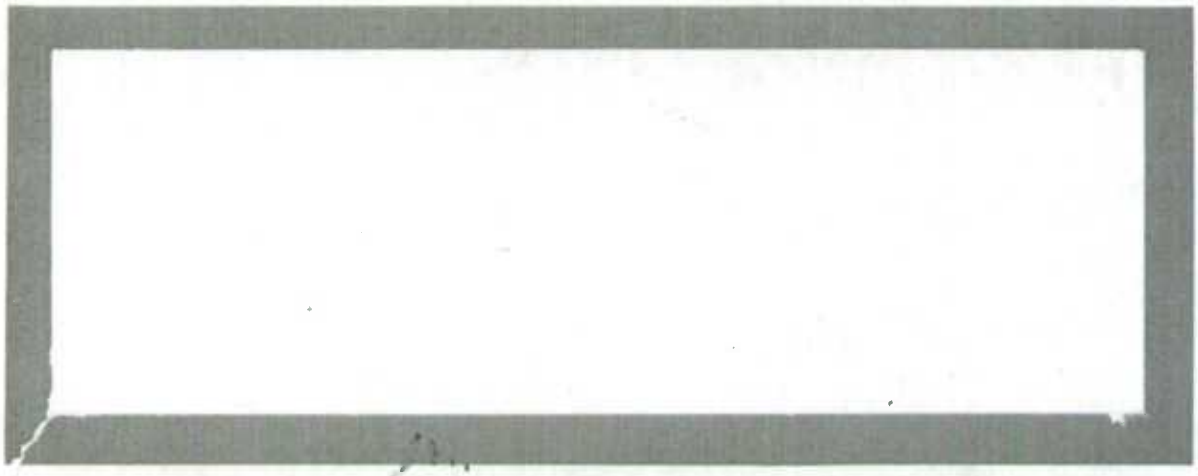
Statistics Canada Statistique Canada



Methodology Branch

Business Survey Methods Division

Direction de la méthodologie

Division des méthodes d'enquêtes-entreprises

Canada

VARIANCE ESTIMATION FOR PUBLIC USE MICRODATA FILES

By

Wesley Yung and Joseph Duggan

BSMD-2000-003E

WORKING PAPER

METHODOLOGY BRANCH

VARIANCE ESTIMATION FOR PUBLIC USE MICRODATA FILES

Wesley Yung and Joseph Duggan

Business Survey Methods Division
Statistics Canada

May, 2000

Variance Estimation For Public Use Microdata Files

Wesley Yung and Joseph Duggan[1]

Abstract

Statistics Canada produces many Public Use Microdata Files (PUMF) which are used by analysts wishing to perform their own analyses of data gathered by Statistics Canada's surveys. However, sample design information necessary to calculate proper design based variances, can not be included on the PUMF due to confidentiality constraints. As part of the PUMF documentation, approximate sampling variability tables are included which allow the user to calculate an approximate coefficient of variation for totals and ratios of categorical variables. Unfortunately, these tables can not be used for continuous data nor for complex statistics such as regression coefficients. In this paper, we investigate two possibilities for replacing the approximate sampling variability tables. The first method, the Generalized Variance Function approach, is easier to use than the existing method but similarly it suffers from a lack of versatility since it can also only be used for totals and ratios. The second method is based on the bootstrap variance estimator and is much more versatile but does have some confidentiality concerns. The two methods are described and empirical results are obtained using two of Statistics Canada's surveys.

Key Words: Bootstrap variance estimator; CV look-up tables; Generalized variance functions; Public use microdata files; Variance estimation.

---

[1]Wesley Yung and Joseph Duggan, Business Survey Methods Division, Statistics Canada.

Estimation de la Variance des Fichiers de Microdonnées à Grand Diffusion

Wesley Yung et Joseph Duggan[2]

Résumé

Statistique Canada produit beaucoup de fichiers de microdonnées à grande diffusion (FMGD) qui sont utilisés par les analystes souhaitant effectuer leurs propres analyses des données recueillies par les enquêtes de Statistique Canada. Cependant, les renseignements sur le plan d'échantillonnage requis pour calculer les variances appropriées, ne peuvent pas être inclus sur le FMGD à cause des contraintes de confidentialité. Les tables de variabilité approximative de l'échantillonnage sont incluses dans la documentation du FMGD afin de permettre à l'utilisateur de calculer un coefficient de variabilité approximative pour les totaux et les ratios des variables nominales. Malheureusement, ces tables ne peuvent pas être utilisées pour les données continues ni pour les statistiques complexes comme les coefficients de régression. L'article que voici, étudie deux méthodes qui pourraient remplacer les tables de variabilité approximative de l'échantillonnage. La première méthode, l'approche de la fonction de variance généralisée, est plus facile à utiliser que la méthode existante mais, elle souffre aussi d'un manque de versatilité puisque nous ne pouvons l'utiliser que pour les totaux et les ratios. La deuxième méthode est fondée sur la méthode bootstrap pour estimer la variance. Cette dernière est plus versatile mais il y a des inquiétudes concernant la confidentialité des données. Nous décrivons les deux méthodes et donnons des résultats empiriques obtenues de deux enquêtes de Statistique Canada.

Mots Clés: Estimation de la variance; Estimation de la variance de boostratp; Fichier de microdonnées à grande diffusion; Fonction de la variance généralisée; Tables de CV à consulter;

---

[2]Wesley Yung et Joseph Duggan, Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada.

TABLE OF CONTENTS

1. INTRODUCTION

Many of Statistics Canada's surveys produce a Public Use Microdata File (PUMF), which is made available to analysts wishing to perform their own analyses of Statistics Canada data. On these microdata files, each record represents a sampled element (business, household, etc...) and includes a weight which usually incorporates adjustments for nonresponse and benchmarking. However, as part of disclosure avoidance procedures, design information such as stratum or cluster identifiers are not normally included on the PUMF. In the absence of this design information, users of Statistics Canada's PUMF's are unable to calculate valid design-based variance estimators. Currently, users are informed of sampling variability by means of *Approximate Sampling Variability Tables*. These tables give an approximate coefficient of variation (CV) for estimates of totals, ratios and proportions for categorical variables. Unfortunately, these tables cannot be used to obtain CV's for continuous variables or for complex statistics such as estimated regression coefficients. As well, this approach, in use since the 1970's, is now felt to be unsatisfactory for practical and statistical reasons.

Users of Statistics Canada's PUMF's have expressed a wide range of views concerning the use of the CV look-up tables. Unsophisticated users do not understand how to use the tables and, as a result, tend not to use them. Others want only an easy procedure to determine the releasability of an estimate: acceptable, marginal, or unacceptable. Many of Statistics Canada's surveys follow the policy that estimates with CV's less than 16.5% are acceptable, while estimates with CV's between 16.5% and 33.3% are marginal and estimates with CV's greater than 33.3% are unacceptable. Sophisticated users find the tables neither detailed enough nor adequate for complex analyses such as linear or logistic regression analyses. Still others find them burdensome because it is a manual procedure and they have many estimates for which they require CV's. In summary, it appears that there are two distinct groups of users:

1. Basic analysts for whom the CV look-up tables are appropriate (if and when they use them).
2. Sophisticated analysts who find the CV look-up tables burdensome and/or inadequate.

For the first group of users, it would be desirable to find a more automated method to obtain the approximate CV's, while for the second group a method for the analyst to calculate a valid design based variance estimator is desired.

In this paper, we investigate two methods for solving the problem of variance estimation for PUMF's while still respecting the confidentiality constraints. The first method uses the Generalized Variance Function (GVF) approach of Wolter (1985) and is intended for the basic analyst. The GVF approach will allow PUMF users to calculate approximate CV's easily and quickly. For the more sophisticated users, we propose a bootstrap variance estimation method for calculating correct design-based variance estimators and respecting the confidentiality of Statistics Canada's respondents. The bootstrap method will allow PUMF users to calculate variance estimates for totals, ratios and proportions for categorical and continuous variables, as well as for more complex statistics such as regression coefficients. In the following section we describe two existing methods at Statistics Canada, the CV look-up tables and a collapsed jackknife method. Section 3 introduces the GVF method and the development of GVF models, while the proposed bootstrap method is presented in section 4. Empirical comparisons of the four methods are given in section 5 using data from two Statistics Canada surveys.

2. EXISTING METHODS

Given that Statistics Canada has a responsibility to provide its data users with measures of data quality, estimated coefficients of variation (CV's) are usually included with estimates published by Statistics Canada. With the increase in computing power and the availability of commercial statistical analysis software, users of Statistics Canada data have demanded that more and more detailed data become available for analysis. Statistics Canada has answered this demand by making PUMF's available. However, along with the responsibility of providing data quality measures, Statistics Canada also has the responsibility to ensure the confidentiality of its respondents. Thus, information necessary to compute proper designed based CV's (survey design information such as stratum or cluster identifiers) is usually suppressed on PUMF's and approximate CV look-up tables are included with the PUMF's. As an alternative to the look-up tables, a method of collapsing and applying the usual jackknife variance estimator to the collapsed strata and clusters has been investigated. These two methods are discussed in more detail in the following sub-sections.

2.1 Approximate Sampling Variability Tables

Approximate sampling variability tables, or CV look-up tables, have been used for many years as a means of informing microdata file users of sampling variability. Typically, these tables are produced at both the national and provincial levels and occasionally at sub-provincial or regional (groups of provinces) levels. To produce a CV look-up table, a set of key categorical variables is identified and design based variances are obtained for each response category cross-classified by the set of variables for which you want to produce a table. For example, since tables are commonly produced for specific age-sex groups, the key variables would be cross-classified by age-sex groups and variances would be calculated for each resulting cell total. In addition to calculating variances under the survey design, variances are calculated under a simple random sampling design and design effects (DEFF's) are obtained for each combination of response category and cross-classification variables. The 75th percentile of these DEFF's is then used as a representative DEFF for use in preparing the CV look-up tables.

Once a representative DEFF is obtained, the following formula is used to calculate CV's for the look-up table,

$$\hat{CV}^2(\hat{X},\hat{P}) = fpc \times DEFF \times \frac{N(1-\hat{P})}{n\hat{X}} \qquad (2.1)$$

where $\hat{P} = \hat{X}/\hat{Y}$ is the estimated proportion, $\hat{X}$ and $\hat{Y}$ are totals of categorical variables estimated from the sample, N is the total population size, n is the sample size and fpc is the finite population correction factor given by

$$fpc = 1 - \frac{n}{N}.$$

Clearly, if $\hat{Y} = N$ then $\hat{P}$ is an estimated proportion whereas if $\hat{Y}$ is an estimated domain total, then $\hat{P}$ is an estimate of a ratio. Equation (2.1) handles both estimated proportions or ratios as long as $\hat{X}$ and $\hat{Y}$ are categorical variables. Unfortunately, the CV look-up tables cannot be used for continuous data or statistics more complex than totals or ratios.

A portion of a typical CV look-up table appears in Table 2.1.

Table 2.1. Portion of a CV Look-up Table

Numerator of Percentage	Estimated Percentage				
	0.1%	1.0%	2.0%	5.0%	10.0%
10000	*****	*****	14.9	14.6	14.3
11000	*****	*****	14.2	14	13.6
12000	*****	*****	*****	13.4	13
13000	*****	*****	*****	12.8	12.5
14000	*****	*****	*****	12.4	12
15000	*****	*****	*****	12	11.6

To use the CV look-up table to obtain an approximate CV for a total, you simply find the estimate in the left most column and then follow the row across to the first numeric value. For example, suppose we estimate 14,000 persons with a particular characteristic. We find 14,000 in the left most column and follow the row over to obtain an approximate CV of 12.4%. To use the table to obtain an approximate CV for a percentage, we find the numerator of the percentage in the left most column and the estimated percentage along the top. The cell where the row and column intersect is the approximate CV. For example, if we estimate that 11,000 persons have a particular characteristic and these 11,000 persons correspond to 2% of the some population, then the estimated CV would be 14.2%. These examples illustrate another problem with the CV look-up tables in the sense that most of the time, the estimated totals or numerators will not be integer values. If we estimated 14,227 persons with a particular characteristic then we would have to interpolate between the approximate CV's for the estimated totals of 14,000 and 15,000. Similarly, if the estimated percentage is not listed in the columns of the table, one would need to interpolate also. The problems of interpolating between table values for estimates and the repetitive manual nature of obtaining CV's from the tables can be avoided by users who were willing to program the table-generating function (2.1).

2.2 Collapsed Jackknife Variance Estimator

For use with Statistics Canada's National Population Health Survey (NPHS), Mayda et al. (1996) proposed using the collapsing method of Rust (1986) to create 'super-strata' and 'super-clusters' and then applying the usual jackknife variance estimator to the super-strata and super-clusters. Following Rust (1986), design strata are collapsed to form super-strata and then the original clusters are collapsed within the super-strata. The clusters are collapsed in such a way that the super-clusters contain original clusters from the same design strata. The super-strata and super-cluster identifiers are then included on the PUMF, thus allowing analysts to use the jackknife variance estimator. This method is illustrated in Mayda et al. using data from the NPHS. Although results from their empirical study are encouraging, one should take care when collapsing strata and clusters within strata, as Valliant (1995) has shown that under certain conditions the balanced repeated replication (BRR) variance estimator can become inconsistent when strata are collapsed. It is unclear at this point whether the inconsistency property of the BRR extends to the jackknife variance estimator, but the asymptotic equivalent of the BRR and the jackknife variance estimators, as shown in Rao and Wu (1988), suggests that the jackknife variance estimator may also suffer due to collapsing. In addition to the possible inconsistency, the collapsing of clusters will affect the stability of the jackknife variance estimator (see Canty

and Davidson, 1999).

3. GENERALIZED VARIANCE FUNCTIONS

3.1 GVF Motivation

The results of an in-house study of survey managers indicated that there is a sizeable group of users of Statistics Canada's PUMF's that find the CV look-up tables appropriate and are unlikely to attempt more-sophisticated methods for the calculation of measures of variability. Although these users appear content, it is desirable to improve on the estimation of sampling variability currently obtained through the supplied CV look-up tables. For this group of users with basic requirements, Generalized Variance Functions (GVF's) are being proposed as a replacement to the CV look-up tables.

The GVF approach has been in use for many years in the United States with some GVF models being theoretically justified using a prediction model approach (Valliant, 1987). In this paper, we will not attempt to theoretically justify the GVF approach, only to illustrate its use and provide a general recipe for obtaining GVF's. The GVF approach and its justification for survey programs are described thoroughly in Wolter (1985, Chapter 5). The reasons for using the GVF approach, as described by Wolter (1985), are briefly mentioned here. The computation of exact variances can be very time consuming and costly given that hundreds or even thousands of estimates may be published for Statistics Canada's surveys. The availability of PUMF's means that many other estimates, that are not included in official publications, can be computed and can even be estimated for arbitrary domains, thus making it impossible to provide data quality measures for all possible estimates. Finally, the generalized nature of GVF's may in fact lead to more stable measures of reliability than approximate variance estimation methods, such as Taylor approximations or the jackknife variance estimator, which are currently used.

The method of generalized variance functions uses a mathematical model to describe the CV of a survey estimator as a function of the expectation of the estimator. Possible model types include, but are not limited to,

$$CV^2 = \alpha + \beta / X, \tag{3.1}$$

$$CV^2 = \alpha + \beta X + \gamma X^2, \tag{3.2}$$

$$\ln(CV) = \alpha + \beta \ln(X), \tag{3.3}$$

where X is an estimator and CV is the associated coefficient of variation. Wolter (1985) notes that there is very little theoretical justification for any of the models given above, although some work has been done to justify some models using a prediction model approach (Valliant, 1987). To utilize the GVF, one would calculate many estimates of survey variables, $\hat{X}_i$, along with their corresponding CV's and then calculate $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$ using ordinary or weighted least squares. These estimated parameters can be included in the PUMF documentation, thus allowing users to calculate approximate CV's on their own. Clearly, by including only the estimated parameters, confidentiality will not be breached.

3.2 Development of GVF Models

The procedure of obtaining GVF's will be illustrated using Statistics Canada's 1995 Survey of Work Arrangements (SWA). The 1995 SWA was conducted nationally as a supplement to the Labour Force Survey and collected information on hours of work, work schedules, reasons for working at home, compensation for overtime, and reasons for self-employment, or for holding more than one temporary or permanent job.

A large number (8,133) of $(\hat{X}_i, C\hat{V}_i)$ pairs were available to develop the GVF models. The 8,133 pairs of estimates and their corresponding CV's represent all response levels of 33 categorical variables within each of the ten provinces, the Atlantic provinces as a whole, the Prairie provinces as a whole and the Canada level (all provinces). The 33 variables (grouped into 16 variable groups based on subject matter knowledge) are presented in Table 3.1. These 8,133 estimates were used to obtain the representative design effects which were used to construct the CV look-up tables included in the SWA PUMF documentation.

Table 3.1. Definition of Variable Groups

Variable Group	Description of variables contained in variable group
1	Education level, Activity of respondent during week, Multiple job holder
2	Actual hours worked during week, Type of job (full/part-time)
3	Class of worker (main job), Wage or salary reporting period
4	Labour force status of spouse, Class of worker (main job) of spouse
5	Usually works Monday to Friday, Usually works Monday to Sunday
6	Usually works Tuesday, Usually works Wednesday
7	Usually works Friday, Usually works Saturday
8	Regular work schedule described, Reason for schedule
9	Can set work hours (within limits), Usually does some work at home
0	Worked paid overtime last week, Rate of pay for overtime
A	Self Employed: works Monday to Friday, Self Employed: works Monday to Sunday
B	Self Employed: usually works Tuesday, Self Employed: usually works Wednesday
C	Self Employed: usually works Friday, Self Employed: usually works Saturday
D	Sex, Age Group
E	Usual hours worked (main job), Is a union member
F	Is job permanent or is there some way not permanent, What way is job not permanent?

We now describe the steps taken to obtain the GVF's for the SWA PUMF. This description is meant to illustrate the model fitting process used to obtain the final GVF's. It is not our intention to represent the procedure which we followed as the only method or the best method that can be used. For example, the final

GVF's were obtained using simple linear regression while a weighted regression approach could well have been utilized. Also note that while all three models (3.1), (3.2) and (3.3) were investigated, model (3.3) was selected as the final model and therefore only the results from this model will be presented.

3.2.1 Single GVF

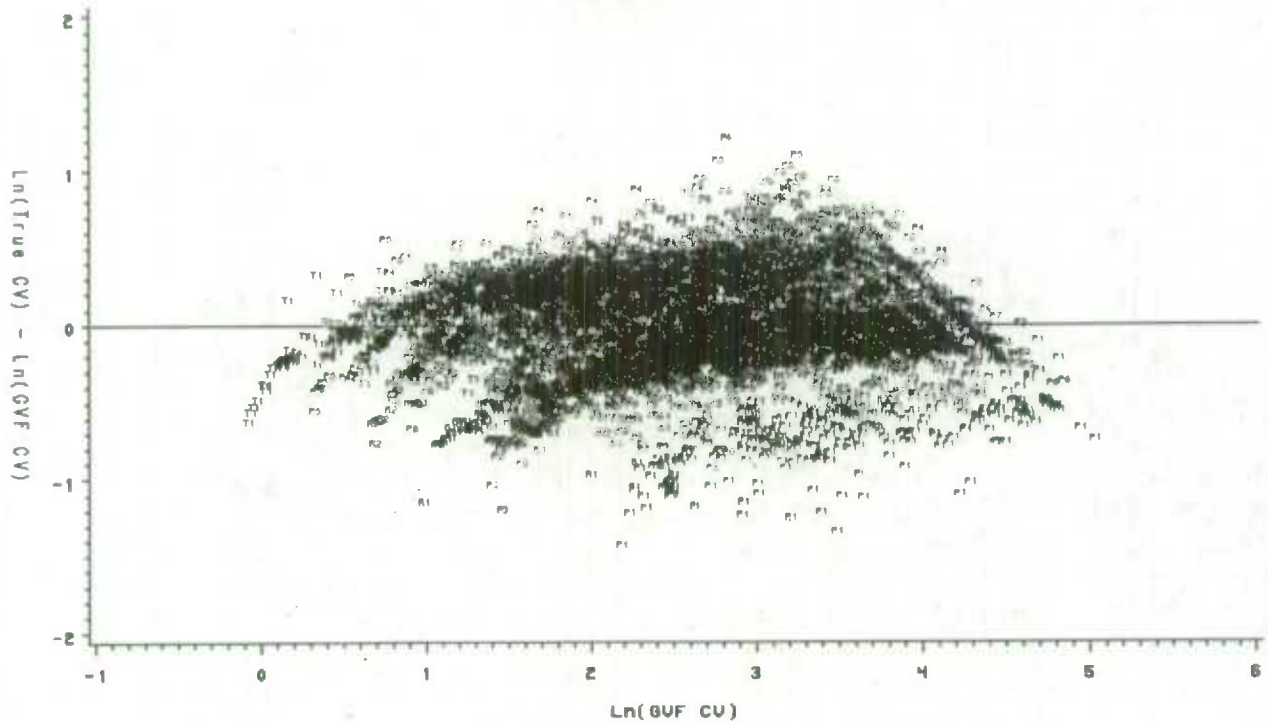
Although is was realized that it would be very unlikely that a single GVF would suffice, the first step taken was to use all 8,133 data points to fit a single regression model with the following results,

$$\ln(CV) = 6.87 - 0.42 \ln(\hat{X}).$$

In order to verify the appropriateness of the model, common diagnostics plots were produced and analyzed. A particularly useful plot was a plot of the residuals against the predicted values where the geographical regions were used as a third dimension. That is, on the plot of the residuals versus the predicted values, the geographical regions were identified by different plotting symbols (see Figure 3.1).

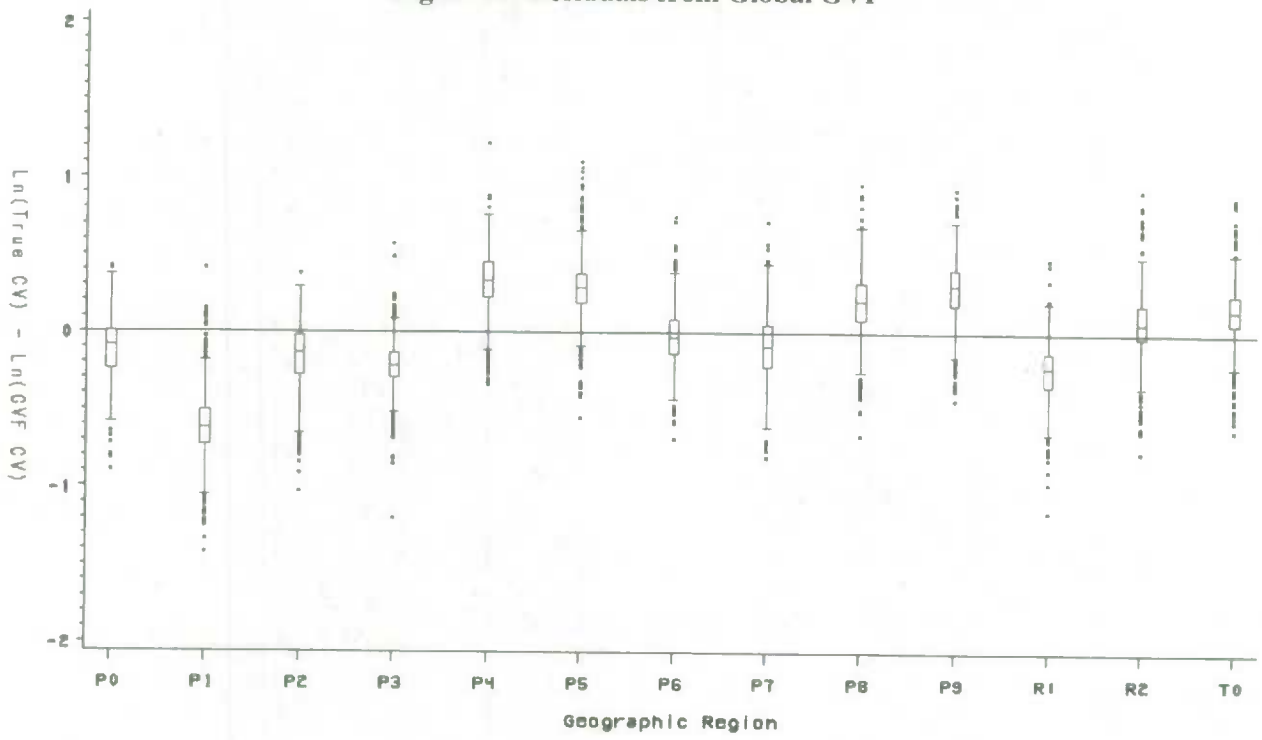
As we can see from Figure 3.1, it appears that the global regression model suffers from serious problems within some geographical regions. For example, the residuals from geographical region 1 appear to be negative which indicates that the global model does not fit well in that particular region. As a result of this, it was felt that separate GVF's should be fit within each of the 13 geographical regions. This is not surprising as CV look-up tables are typically constructed for each province as well as the regions and at the Canada level due to regional differences.

Figure 3.1 Residuals vs Predicted Values
Global GVF



In addition to the typical diagnostic plots, box plots of the residuals for each of the 13 regions were produced (see Figure 3.2). From Figure 3.2, it is clear that the assumption of the residuals having a mean of zero and being homogeneous is severely violated. For example, for province P1 the mean of the residuals is clearly below zero and the residuals appear to be much more variable than the other 12 geographical regions. Based on these observations, it was confirmed that separate GVF's should be fit within each of the 13 regions.

Figure 3.2 Residuals from Global GVF



3.2.2 Separate GVF's within Regions

As a result of the analysis of the residuals, separate GVF's were fit within each of the 13 geographical regions with the estimated regression parameters given in Table 3.2.

Note that while there is no reason to restrict ourselves to the same model within the different regions, for consistency we restrict ourselves to a single model. In fact, all three models presented earlier were investigated and model (3.3) was chosen as the best. Therefore, the results of model (3.3) will be the only results presented here.

Analysis of the residuals was performed separately for each of the 13 GVF's. For brevity, we will discuss the analysis for only one geographical region as the analyses for the other regions were similar. Figure 3.3 presents a plot of the residuals versus the predicted values with the plotting symbols representing the 16 different variable groups given in Table 3.1. From Figure 3.3, it appears that the residuals for variable groups C and D tend to be negative. This indicates that separate GVF's should be considered for these two variable groups. Box plots of the residuals confirm this (see Figure 3.4) and in addition, indicate that separate GVF's should be considered for variable groups A and 5 due to the smaller variability exhibited by the residuals for

variable group A and for the long negative tail of the residuals for variable group 5. Note that although we have included diagnostic plots for only a single province, the diagnostic plots of all regions were analyzed

Table 3.2. Estimated Parameters for Regional GVF's

Region	Alpha	Beta
Newfoundland	7.33	-0.49
Prince Edward Island	6.85	-0.5
Nova Scotia	7.53	-0.51
New Brunswick	7.22	-0.49
Quebec	7.7	-0.47
Ontario	7.48	-0.45
Manitoba	7.28	-0.47
Saskatchewan	7.51	-0.5
Alberta	7.82	-0.49
British Columbia	7.8	-0.48
Altantic Provinces	7.39	-0.5
Priarie Provinces	7.6	-0.48
Canada	7.53	-0.47

before deciding which variable groups warranted a separate GVF. Thus, although it appears that variable group F deserves a separate GVF in Figure 3.4, when all regions were considered it was decided that a separate GVF for variable group F was not necessary.

Figure 3.3 Residuals vs Predicted Values
GVF for Ontario

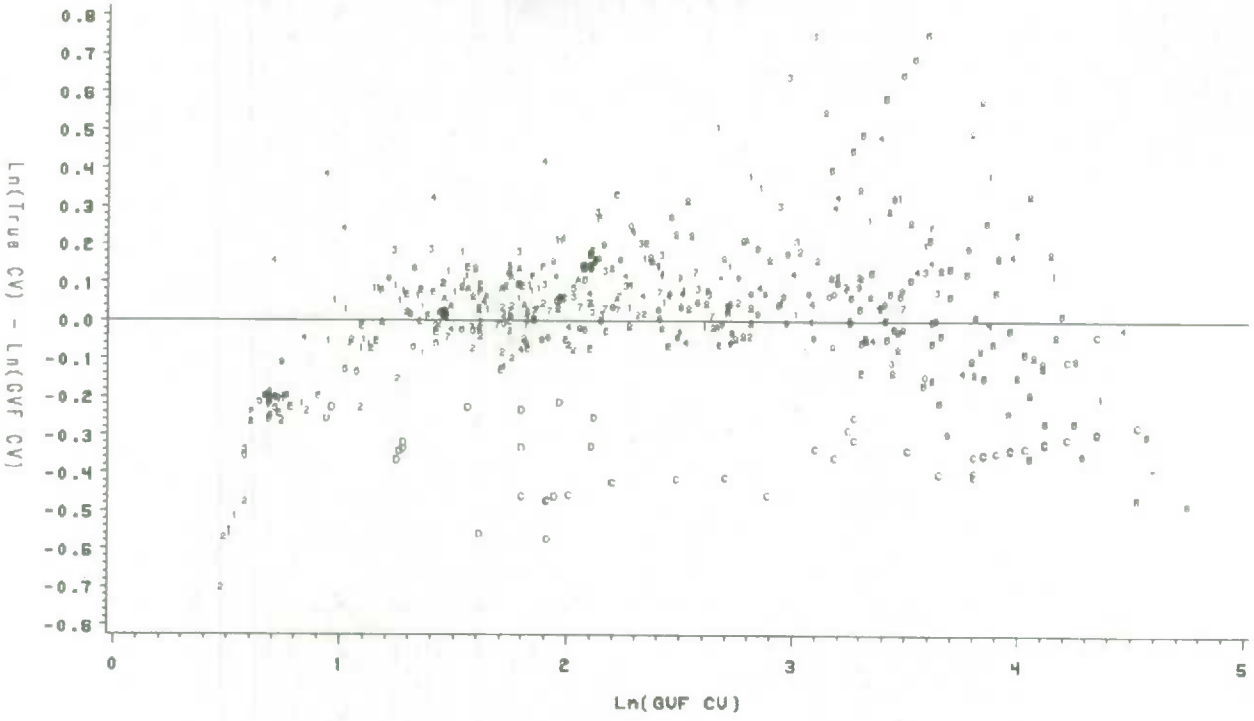
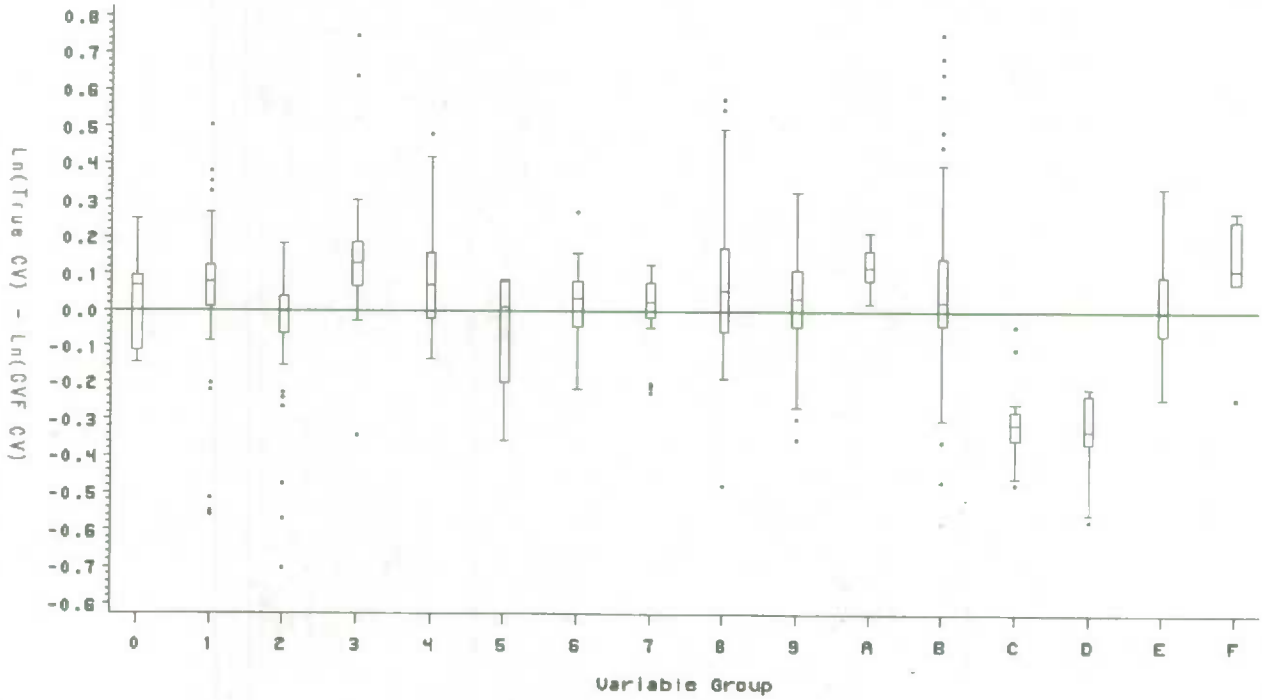


Figure 3.4 Residuals from GVF for Ontario



3.2.3 Separate GVF's for Variables Groups within Geography

Separate GVF's for variable groups 5, A, C and D were investigated within each of the 13 geographical regions in order to be consistent. Before fitting the separate GVF's for the four variable groups mentioned above, logical relationships between the 16 variable groups were examined. Noting that variable groups A, B, and C were asked of self-employed people only, these three groups were further combined to create a new variable group called SELF-EMP. No other combining of variable groups was performed. Thus, separate GVF's were fit for variable groups 5, SELF-EMP, and D within each of the 13 geographical regions.

The resulting 39 GVF's were analyzed using diagnostic plots as previously described. The results of this analysis indicated that variable group 5 did not warrant a separate GVF, while the other two variable groups did. Thus, the final GVF's consisted of three separate GVF's (SELF-EMP, variable group D and all remaining variable groups) in each of the 13 geographical regions. The final GVF's for several provinces appear in Table 3.3.

Table 3.3. Estimated Parameters for Variable Group GVF's in Selected Regions

Region	Variable Group	Alpha	Beta
Newfoundland	Self Employed	6.71	-0.41
Newfoundland	D	5.88	-0.38
Newfoundland	Other	7.36	-0.49
Prince Edward Island	Self Employed	6.53	-0.44
Prince Edward Island	D	6.31	-0.48
Prince Edward Island	Other	6.88	-0.50
Nova Scotia	Self Employed	7.26	-0.48
Nova Scotia	D	6.44	-0.43
Nova Scotia	Other	7.59	-0.52

3.2.4 Modifications to the Final Model

Recall that the CV look-up table method uses the 75th percentile design effect (DEFF) as the representative DEFF. As a result of this, one would expect the CV look-up table method to overestimate the true CV 75% of the time and to underestimate the true CV 25% of the time. The GVF method does not possess this desirable property since if the GVF fits the data well, 50% of the estimated CV's will overestimate the true CV and 50% will underestimate the true CV.

A number of methods for modifying the GVF models to achieve a more conservative estimation of CV's have been investigated for other applications. For the Labour Force Survey (Phillips and Kaushal, 1998) final GVF models were fitted in two passes. The observations that lay below the preliminary regression line were removed and then a second and final regression line was fitted. By this procedure, seventy-five percent of

all observations were below the model regression line and therefore only twenty-five percent of all estimated CV's were underestimated.

For the survey data used in this study, this two-step regression-fitting yielded some problematic results. First of all, the log-log transformations that provided the best linear fits to the data typically produced a plotted line that was slightly concave downward. Linear regressions tended to overestimate CV's for both small and large estimates; as seen in the plots of the residuals. Further attempts to improve the linearity of the data through transformations and other adjustments proved unsuccessful or produced only limited gains for the extra effort invested. Application of the two-step regression to these curves resulted in the truncation of the interval over which the regression model was fitted and overestimated the CV's for the excluded observations.

As an alternative, a conservative upward shift of the regression line was investigated and implemented. The procedure consisted of increasing the estimate for the model intercept term, $\hat{\alpha}$, by a small factor. The new, conservative intercept term $\hat{\alpha}^*$ was specified for the SWA study according to $\hat{\alpha}^* = \hat{\alpha} + z_{0.25} * \hat{\sigma}_{\alpha}$, which is equivalent to the upper bound of a fifty percent confidence interval about the estimated intercept. The factor $z_{0.25}$ is taken from a normal distribution, and $\hat{\sigma}_{\alpha}$ is the estimated standard error of the intercept term.

The conservative shift will still lead to overestimation for the larger and smaller estimates, but the amount of overestimation was observed to be more controlled than for the two-step procedure. A certain component of this overestimation for larger estimates arises from model groups of larger-than-average estimates, as seen at the Canada level. The added distance that these observations lie from the y-axis seems to contribute to the decreased precision with which the intercept of the model can be estimated. As a result, the amount of the shift in the regression line is larger than it might be for smaller regions. Of course, it is possible to further adjust the GVF intercept inflation factors individually to ensure that the regression fits are equally conservative as the CV look-up table method. In mathematical notation, this is equivalent to an inflation factor of $1 + c_k * z_{0.25} * \hat{\sigma}_{\alpha}$, where the c_k are arbitrarily-fixed constants which provide sufficient overestimation when applied to GVF k. The final GVF's, for several provinces, for the SWA PUMF are given in Table 3.4.

3.3 Dissemination of GVF Models

There are many options for dissemination of these estimated GVF's to PUMF users. The traditional CV look-up table format can easily be simulated using the GVF's, the model parameters and transforming equations can be released in the documentation (for example see Table 3.4), or a Visual Basic program (see Brisebois, 1998) can be supplied to compute the appropriate estimated CV. The automated option is an attractive prospect for the more-complicated GVF's; a significant subset of the user population would not have much experience with exponential functions and negative exponential powers.

Table 3.4. Estimated Conservative Parameters for Variable Group GVF's in Selected Regions

Region	Variable Group	Alpha	Beta
Newfoundland	Self Employed	6.75	-0.41
Newfoundland	D	6.22	-0.38
Newfoundland	Other	7.39	-0.49
Prince Edward Island	Self Employed	6.66	-0.44
Prince Edward Island	D	6.63	-0.48
Prince Edward Island	Other	6.91	-0.50
Nova Scotia	Self Employed	7.29	-0.48
Nova Scotia	D	6.81	-0.43
Nova Scotia	Other	7.62	-0.52

4. BOOTSTRAP VARIANCE ESTIMATION

4.1 Rao-Wu Bootstrap

The bootstrap variance estimation method for the *iid* case has been extensively studied, see Efron (1982). Rao and Wu (1988) provided an extension to stratified multi-stage designs but covered only smooth statistics $\hat{\theta} = g(\hat{Y})$. The Rao-Wu bootstrap was extended by Rao, Wu and Yue (1992) to include non-smooth statistics as well as smooth statistics. The design considered by Rao, Wu and Yue, and in this paper, assumes L design strata with N_h clusters in the *h-th* stratum. Within the *h-th* stratum, $n_h \geq 2$ clusters are selected and further subsampling within selected clusters is performed according to some probability sampling design. Although the subsampling is not specified, it is assumed that there is unbiased estimation of cluster totals, Y_{hi}, $h=1, \ldots, L$; $i = 1, \ldots, n_h$. Based on the survey design, design weights, w_{hik}, associated with the *(hik)-th* sampled element are obtained. Also associated with the *(hik)-th* sampled element is the variable of interest, y_{hik}. An estimator of the total Y is then given by

$$\hat{Y} = \sum_{(hik) \in s} w_{hik} y_{hik} \tag{4.1}$$

where s denotes the sampled elements. The design weights are often subjected to adjustments such as poststratification or generalized regression to ensure consistency to known population totals. For example, suppose that each element in the population belongs to a poststratum that can cut across the design strata. Using prescript notation to denote poststrata, the total number of elements in the *c-th* poststratum is ${}_cM$ and is assumed to be known. Letting ${}_cw_{hik}$ represent the poststratified or final weight defined by

$${}_cw_{hik} = \frac{{}_cM}{{}_c\hat{M}} w_{hik},$$

where ${}_c\hat{M} = \sum_{(hik)\in s} w_{hik}\,{}_c\delta_{hik}$ and ${}_c\delta_{hik}$ is the poststratum indicator variable, the poststratified estimator is defined as

$$\hat{Y}_{ps} = \sum_c \sum_{(hik)\in s} {}_c w_{hik}\, y_{hik}\, {}_c\delta_{hik}.$$

To calculate a bootstrap variance estimator for $\hat{\theta} = g(\hat{Y})$, where $\hat{Y}$ is given by equation (4.1) and g is a known function, the Rao-Wu-Yue method proceeds as follows: (note that the poststratified estimator can be expressed in this form)

i) Independently, in each stratum, select a simple random sample of m_h clusters with replacement from the n_h sample clusters.

ii) Let m_{hi}^* be the number of times the *(hi)-th* cluster is selected ($\sum_i m_{hi}^* = m_h$). Define the bootstrap weights as

$$w_{hik}^* = \left[1 - \left(\frac{m_h}{n_h - 1}\right)^{1/2} + \left(\frac{m_h}{n_h - 1}\right)^{1/2} \frac{n_h}{m_h} m_{hi}^*\right] w_{hik}. \tag{4.2}$$

If the size of the simple random sample, m_h, is chosen to be less than or equal to n_h-1, then the bootstrap weights, w_{hik}^*, will all be positive.

iii) To obtain the final bootstrap weight, perform the same weight adjustment with the design weights, w_{hik}, replaced by the bootstrap weights, w_{hik}^*. For example, the final bootstrap weight for the poststratified estimator is

$${}_c\tilde{w}_{hik}^* = \frac{{}_cM}{{}_c\hat{M}^*} w_{hik}^*$$

where ${}_c\hat{M}^* = \sum_{(hik)\in s} w_{hik}^*\,{}_c\delta_{hik}$.

iv) Calculate $\hat{\theta}^*$, the bootstrap estimator of θ, using the final bootstrap weights, $\tilde{w}_{hik}^*$, in the formula for $\hat{\theta}$.

v) Independently replicate steps (I) to (iv) a large number of times, B, and calculate the corresponding estimates, $\hat{\theta}_{(1)}^*, \ldots, \hat{\theta}_{(B)}^*$.

The bootstrap variance estimator for $\hat{\theta}$ is then given by

$$v_B(\hat{\theta}) = \frac{1}{B}\sum_b \left(\hat{\theta}_{(b)}^* - \hat{\theta}_{(\cdot)}^*\right)^2$$

where $\hat{\theta}_{(\cdot)}^* = (1/B)\sum_b \hat{\theta}_{(b)}^*$.

A commonly used value for m_h is n_h-1 in which case equation (4.2) reduces to

$$w_{hik}^* = \frac{n_h}{n_h - 1} m_{hi}^*\, w_{hik}. \tag{4.3}$$

If a sampled element is in a cluster that has not been selected in a particular bootstrap sample, then $m_{hi}^{*} = 0$ and the bootstrap weight is equal to zero. That is, all sampled elements in the cluster have bootstrap weights equal to zero and in the case of multiplicative weight adjustments (e.g. poststratification or regression), will also have bootstrap final weights equal to zero. Now, within each bootstrap sample at least one cluster per stratum will have bootstrap final weights equal to zero, so that members of the same cluster cannot be identified by their zero weights. Unfortunately, when the bootstrap final weights are combined over all B bootstrap samples, cluster membership can be identified. By grouping individuals based on zero and non-zero bootstrap final weights, the members of each cluster can be identified. Although location of the clusters is not given on the PUMF, use of other variables on the PUMF may allow users to deduce the location of a cluster, thus breaching confidentiality.

This problem occurs because under a stratified multi-stage design the bootstrap resamples entire clusters. In the case of stratified simple random sampling, confidentiality is preserved since the cluster consists of a single element. Unfortunately, for stratified multi-stage samples (commonly used in social surveys), the bootstrap method does not meet Statistics Canada's confidentiality guidelines.

As a possible solution to this problem, it was suggested to change the size of the simple random sample, m_h, so that equation (4.2) does not reduce to (4.3). Reducing m_h to be less than $n_h - 1$ may cause problems as it is common to select only 2 clusters per stratum. In this case, some sort of collapsing would be necessary to increase the number of clusters per stratum. Increasing m_h to be greater than $n_h - 1$ would result in negative bootstrap weights, which is not problematic as long as the analysts restrict the use of the negative weights to variance estimation and not for point estimation. Upon closer examination it was noted that the zero weights obtained by using $m_h = n_h - 1$ were replaced by negative weights and the problem with confidentiality still persists.

Two solutions have been investigated to resolve the confidentiality problem: (1) modifying the poststratification adjustment, and (2) replacing the bootstrap weight by an average bootstrap weight. The modified poststratification adjustment is given in Yung (1997). The average bootstrap weight method is described below.

4.2 Mean Bootstrap Weights

The confidentiality problem occurs because m_{hi}^{*} is always equal to zero for one or more clusters. To avoid this problem, produce R bootstrap samples and average the m_{hi}^{*}'s over the R samples. As long as each cluster appears in at least one of the R bootstrap samples, the averages will all be non-zero. The steps to perform the mean bootstrap are as follows:

i) Independently, in each stratum, select a simple random sample of n_h-1 clusters with replacement from the n_h sample clusters.

ii) Repeat step (i) R times.

iii) Let $m_{hi(r)}^{*}$ be the number of times the *(hi)-th* cluster is selected in the *r-th* bootstrap sample. Let $m_{hi(\cdot)}^{*} = (1/R)\sum_r m_{hi(r)}^{*}$ be the average number of times the *(hi)-th* cluster is selected over the R bootstrap samples.

iv) Define the mean bootstrap weight as

$$w^{*}_{hik(\cdot)} = \frac{n_h}{n_h - 1}\, m^{*}_{hi(\cdot)}\, w_{hik}.$$

v) Obtain the bootstrap final weights, $\tilde{w}^{*}_{hik(\cdot)}$, by performing the same weight adjustment substituting the mean bootstrap weight, $w^{*}_{hik(\cdot)}$ for the design weight, w_{hik}.

vi) Calculate $\tilde{\theta}^{*}$ using the bootstrap final weights in the formula for $\hat{\theta}$.

vii) Independently replicate steps (i) to (vi) a large number of times, B, and calculate the corresponding estimates, $\tilde{\theta}^{*}_{(1)},\ldots,\tilde{\theta}^{*}_{(B)}$.

The mean bootstrap variance estimator is then given as

$$v_{MB}\left(\hat{\theta}\right) = \frac{R}{B}\sum_b \left(\tilde{\theta}^{*}_{(b)} - \tilde{\theta}^{*}_{(\cdot)}\right)^2$$

where $\tilde{\theta}^{*}_{(\cdot)} = (1/B)\sum_b \tilde{\theta}^{*}_{(b)}$. We note that the size of R should be large enough so that the chance of $m^{*}_{hi(r)} = 0$ for all $r = 1,\ldots,R$ is very small, but it should not be so large that drawing $R \times B$ bootstrap samples becomes computationally infeasible.

To justify the mean bootstrap variance estimator, we consider the linear case (i.e. $\hat{\theta} = \hat{Y}$ where $\hat{Y}$ is given by equation (4.1)). Letting E_{*} denote expectation with respect to bootstrap sampling, we wish to evaluate

$$E_{*}\left(v_{MB}(\hat{\theta})\right) = \frac{R}{B}\sum_b E_{*}(\tilde{Y}_{(b)} - \tilde{Y}_{(\cdot)})^2$$

where $\tilde{Y}_{(b)} = \sum_{(hik)\in s} w^{*}_{hik(\cdot)}\, y_{hik}$ and $\tilde{Y}_{(\cdot)} = (1/B)\sum_b \tilde{Y}_{(b)}$. Note that $w^{*}_{hik(\cdot)}$ depends on b, but for notational simplicity, we drop the subscript b. Replacing $\tilde{Y}_{(\cdot)}$ with $\hat{Y}$ (the two are asymptotically equivalent, Rao and Wu, 1988), we have,

$$E_{*}(\tilde{Y}_{(b)} - \hat{Y})^2 = E_{*}(\tilde{Y}^2_{(b)}) - 2\hat{Y}E_{*}(\tilde{Y}_{(b)}) + \hat{Y}^2. \tag{4.4}$$

The $m^{*}_{hi(r)}$ follow a multinomial distribution with parameters $(n_h - 1)$ and $p_i = (1/n_h)$ for all i and r. Thus

$$E_{*}(m^{*}_{hi(\cdot)}) = E_{*}(m^{*}_{hi}) \tag{4.5}$$

where $m^{*}_{hi} \sim M((n_h - 1), p_i = 1/n_h,$ for all $i)$. Similarly, we obtain the following bootstrap moments:

$$\begin{aligned}
E_{*}(m^{*2}_{hi(\cdot)}) &= \frac{1}{R}V_{*}(m^{*}_{hi}) + \left[E_{*}(m^{*}_{hi})\right]^2, \\
E_{*}(m^{*}_{hi(\cdot)}\, m^{*}_{hj(\cdot)}) &= \frac{1}{R}C_{*}(m^{*}_{hi}, m^{*}_{hj}) + \left[E_{*}(m^{*}_{hi})\right]^2, \\
E_{*}(m^{*}_{hi(\cdot)}\, m^{*}_{gi(\cdot)}) &= E_{*}(m^{*}_{hi})\, E_{*}(m^{*}_{gi}) \quad \textit{and} \\
E_{*}(m^{*}_{hi(\cdot)}\, m^{*}_{gj(\cdot)}) &= E_{*}(m^{*}_{hi})\, E_{*}(m^{*}_{gj})
\end{aligned} \tag{4.6}$$

where V_{*} and C_{*} denote variance and covariance with respect to bootstrap sampling respectively. Substituting expressions (4.5) and (4.6) in equation (4.4) gives, after some simplification,

$$E_*(\tilde{Y}_{(b)}-\hat{Y})^2 = \sum_h \sum_i \left(\frac{n_h}{n_h-1}\right)^2 \frac{1}{R} V_*(m_{hi}^*)\, y_{hi}^2 + \sum_h \sum_i \sum_{j,j\neq i} \left(\frac{n_h}{n_h-1}\right)^2 \frac{1}{R} C_*(m_{hi}^*, m_{hj}^*)\, y_{hi} y_{hj}.$$

Noting that $V_*(m_{hi}^*) = ((n_h-1)/n_h)^2$ and $C_*(m_{hi}^*, m_{hj}^*) = -((n_h-1)/n_h^2)$, we finally have

$$E_*(\tilde{Y}_{(b)}-\hat{Y})^2 = \frac{1}{R}\sum_h \frac{n_h}{n_h-1} \sum_i (y_{hi} - \frac{1}{n_h}\sum_j y_{hj})^2$$

$$= \frac{1}{R} v(\hat{Y}),$$

where $v(\hat{Y})$ is a commonly used variance estimator (see Yung and Rao, 1996). Hence, in the linear case the mean bootstrap variance estimator reduces to the customary variance estimator, $v(\hat{Y})$.

Upon further investigation of the mean bootstrap weights, Mantel (1999) discovered that the cluster membership may still be disclosed. By calculating the correlations between the B bootstrap weights, it was found that units in the same cluster have highly correlated bootstrap weights, while for units in different clusters, the bootstrap weights were nearly independent. While cluster membership may still be disclosed using the mean bootstrap weights, the work necessary is significantly more than that which is needed to disclose the cluster membership using the Rao-Wu bootstrap weights. This problem of disclosure can be attributed to the selection of clusters by the bootstrap method and the treating of all units within the selected cluster the same. If the bootstrap is to be disclosure safe, collapsing of clusters into super-clusters and selecting bootstrap sample of the super-clusters maybe the only solution, although membership of the super-clusters could still be disclosed. Empirical results of collapsing will be included in section 5.

5. EMPIRICAL COMPARISONS

To compare the proposed methods with the CV look-up table method, empirical results were obtained using PUMF data from two of Statistics Canada's surveys: the 1994 National Population Health Survey (NPHS) and the 1995 Survey of Work Arrangements (SWA). Using the NPHS data, the empirical results of the mean bootstrap and the GVF models were compared with the results of the collapsed jackknife, collapsed bootstrap and the CV look-up methods. A more in depth comparison of the GVF approach and the CV look-up table method was performed using data from SWA.

5.1 National Population Health Survey

The NPHS was designed to collect information related to the health of the Canadian population. The objectives of the NPHS included:

- to aid in the development of public policy by providing measures of the level, trend and distribution of the health status of the population;
- to provide data for analytic studies that will assist in understanding the determinants of health;
- to increase the understanding of the relationship between health status and health care utilization, including alternative as well as traditional services.

The design of the NPHS consisted of a stratified two-stage design. In the first stage, homogeneous strata were formed and independent samples of clusters were drawn from each stratum. In the second stage, dwellings were selected within each sampled cluster. The design weights were obtained based on both

stages of sampling. To obtain the final weight, a series of twelve weighting adjustments were performed with the last adjustment being a poststratification adjustment. More information on the NPHS design and weighting is available in the NPHS PUMF documentation.

Using the NPHS data, empirical comparisons were performed between CV's obtained by the CV look-up table, collapsed jackknife, mean bootstrap, a collapsed bootstrap and the GVF methods. To calculate the look-up table CV's the formula given by equation (2.1) was used. The use of this formula removed the need to interpolate between rows and/or columns of the look-up table and also removed the need to manually look up each estimate. The results for the collapsed jackknife were those obtained by Mayda et al. (1996) and were graciously provided by the authors.

To implement the mean bootstrap, n_h-1 clusters were sampled with replacement within each stratum. One hundred sets of mean bootstrap weights were generated, with each mean weight based on 20 bootstrap samples (i.e. 2000 bootstrap samples in total). For each set of mean bootstrap weights, only the poststratification adjustment was performed.

In addition to the mean bootstrap estimator, a collapsed bootstrap variance estimator was also calculated. The collapsed bootstrap was included for comparison purposes only, so a naive collapsing strategy was performed. Clusters were randomly assigned to one of two super-clusters regardless of the number of original clusters in the stratum. If there were only two clusters within a stratum, no collapsing was performed. As mentioned above, this collapsing strategy is crude and can probably be improved (for example, see Rao and Shao, 1996 or Kovacevic and Yung, 1997). Once collapsed, the super-clusters were treated as original clusters and a simple Rao-Shao bootstrap was performed. As with the mean bootstrap, only the poststratification adjustment was performed.

For comparison purposes, a true CV was computed based on the full jackknife variance estimator (without collapsing), since the jackknife method is currently in use for the NPHS. Table 5.1 gives some results of the comparison between the CV's obtained from the mean bootstrap, collapsed jackknife, collapsed bootstrap, CV look-up tables and the GVF method for means, totals and ratios of categorical and continuous variables.

Table 5.1 - Comparison of CV's for Means, Totals and Ratios for the NPHS

$CV_i - CV_J$	Mean Bootstrap	Collapsed Bootstrap	Collapsed Jackknife	CV Table	GVF
±1 %	71 (78.9%)	53 (58.9%)	41 (45.6%)	45 (50.0%)	23(25.5%)
±2 %	81 (90.0%)	73 (81.1%)	62 (68.9%)	58 (64.4%)	37(41.1%)
±3 %	85 (94.4%)	84 (93.3%)	72 (80.0%)	64 (71.1%)	54(60.0%)
±4 %	87 (96.7%)	87 (96.7%)	77 (85.6%)	71 (78.9%)	58(64.4%)
>4 %	90 (100.0%)	90 (100.0%)	90 (100.0%)	75 (83.3%)	75(83.3%)

In Table 5.1, CV_i denotes the CV based on the mean bootstrap, the collapsed bootstrap, the collapsed jackknife, the CV look-up tables or the GVF method and CV_J represents the true CV. Table 5.1 shows one of the drawbacks of the CV look-up table method. Of the 90 estimates, 15 involved a continuous variable

which could not be handled by the CV look-up tables. Of the remaining 75 estimates, 71 were within ±4% of the true CV. The GVF method suffers this same drawback and in fact performed worse than the CV look-up table with only 58 estimates within ±4% of the true CV. Upon closer examination of the GVF results, it was noted that all of the estimates selected by Mayda et al. (1996) happened to have true CV's that fell above the GVF estimated CV's. While not a justification, it does explain the poor performance of the GVF method. When all of the estimates used in the calculation of the GVF and the CV look-up table are used, we see that the performance of the GVF is similar to the performance of the CV look-up table approach (see Table 5.2).

Table 5.2 Comparison of GVF and CV Look-up Table CV's for NPHS

$CV_i - CV_j$	CV Table	GVF
±1 %	694 (35.1%)	690 (34.9%)
±2 %	1,076 (54.5%)	1,112 (56.3%)
±3 %	1,308 (66.2%)	1,349 (68.3%)
±4 %	1,524 (77.2%)	1,535 (77.7%)
>4 %	1,975 (100.0%)	1,975 (100.0%)

Returning to Table 5.1, we see that the mean bootstrap, collapsed bootstrap and the collapsed jackknife performed well for all estimates with the mean bootstrap performing better than both the collapsed bootstrap and the collapsed jackknife, 78% of the estimates were within ±1% versus 59% for the collapsed bootstrap and 46% for the collapsed jackknife. The effect of the amount of collapsing is apparent when comparing the results of the collapsed bootstrap and the collapsed jackknife. In calculating the collapsed jackknife, Mayda et al. (1996) performed significantly more collapsing than that performed for the collapsed bootstrap. Thus, it is not surprising to see the collapsed bootstrap performing considerably better than the collapsed jackknife.

To illustrate the versatility of the bootstrap method, CV's were calculated for the regression coefficients for the model relating a person's health status to a measure of the restriction of activities, age, type of drinker and household income. The model was fitted separately within five provinces giving a total of 25 parameter estimates. Again, the full jackknife CV was considered to be the true CV. The results of this comparison are given in Table 5.3

From Table 5.3, we see that the collapsed bootstrap performs consistently better than both the mean bootstrap and the collapsed jackknife with 72% of the estimated CV's within ±1% for the collapsed bootstrap compared with 52% for the mean bootstrap and only 24% for the collapsed jackknife. In addition, the collapsed bootstrap has slightly more estimates within ±4% (92% versus 84% for the mean bootstrap and 72% for the collapsed jackknife). This result is somewhat surprising since one would expect the collapsed bootstrap to perform worse than the mean bootstrap. However, as shown in Yeo, Mantel and Liu (1999) the performance of the mean bootstrap will improve as the number of bootstrap samples increases. At this point it is unclear if the collapsed bootstrap will still perform better than the mean bootstrap if the number of bootstrap samples is large, say 500.

Table 5.3 - Comparison of CV's for Regression Coefficients for the NPHS

$CV_i - CV$	Mean Bootstrap	Collapsed Bootstrap	Collapsed Jackknife
±1%	13 (52%)	18 (72%)	6 (24%)
±2%	18 (72%)	19 (76%)	11 (44%)
±3%	20 (80%)	20 (80%)	15 (60%)
±4%	21 (84%)	23 (92%)	18 (72%)
>4%	25 (100%)	25 (100%)	25 (100%)

5.2 Survey of Work Arrangements

The performance of the GVF models was compared with that of the CV look-up tables using data from Statistics Canada's 1995 Survey of Work Arrangements. As indicated earlier, the SWA was conducted nationally and collected information such as hours of work, work schedule, reasons for working at home and reasons for self-employment. To compare the GVF models with the CV look-up tables, approximate CV's were obtained by both methods for the variables which were used to construct the GVF models and the CV look-up tables. While it is noted that using the same variables used in constructing the GVF models and CV look-up tables is not the best method to evaluate the GVF models, this analysis should suffice for comparison purposes. As mentioned earlier, over 8,000 estimates of categorical variables with corresponding true CV's were available for comparison. GVF models and CV look-up tables at the provincial level were used in this analysis. Table 5.4 presents the results of the comparison between the CV's obtained by the GVF method and the CV look-up tables and the true CV's.

Table 5.4 - Comparison of GVF and CV Look-up Table CV's

$CV_i - CV$	CV Table	GVF
±1%	3515 (43.2%)	4297 (52.8%)
±2%	4629 (56.9%)	5520 (67.9%)
±3%	5326 (65.5%)	6112 (75.2%)
±4%	5793 (71.2%)	6517 (80.1%)
>4%	8133 (100.0%)	8133 (100.0%)

From Table 5.4, one can see that the GVF models outperform the CV look-up tables in terms of absolute difference with 52.8% of the GVF CV's within ±1%, compared with only 43.2% for the CV look-up table method. In total, the GVF models produced CV's closer to the true CV than the CV look-up table 64% of the time.

If we compare the GVF CV's and the CV look-up table CV's according to the release category, we see that the release category for most estimates remains unchanged (see Table 5.5).

Table 5.5 - GVF and CV Look-up Table CV's According to Release Category

CV Look-up Table Method	GVF Method			
	Acceptable	Marginal	Unacceptable	Total
Acceptable	4,470	0	0	4,470
Marginal	384	1,177	0	1,561
Unacceptable	0	359	1,743	2,102
Total	4,854	1,536	1,743	8,133

The fact that the CV look-up table method produced fewer acceptable CV estimates is most likely due to the use of the 75% DEFF which produces more conservative estimates.

Now, if we control for the true release category as derived from the true CV, we obtain Table 5.6 which presents the agreement between the true release category and the categories obtained by the GVF and CV look-up table methods.

Table 5.6 - Comparison of True Release Category and GVF and CV Look-up Table Release Categories

	True Release Category		
	Acceptable	Marginal	Unacceptable
GVF method	4,636 (98.2%)	1,290 (78.5%)	1,607 (90.8%)
CV look-up table method	4,405 (93.4%)	1,190 (72.4%)	1,713 (96.8%)

From Table 5.6, we see that the GVF release categories agree with the true categories more often than the CV look-up table release categories for the acceptable (98.2% versus 93.4%) and the marginal (78.5% versus 72.4%) categories, but not for the unacceptable category (90.8% versus 96.8%). That is, by more accurately estimating the CV's of the survey estimates, the GVF method has provided the analyst several hundred more estimates which are acceptable. However, as a trade off of having more acceptable estimates, the GVF method may identify some unacceptable estimates as marginal. The extent of these errors can be controlled using techniques discussed in section 3.2.4.

6. CONCLUSIONS

For those users of Statistics Canada's Public Use Microdata Files who find the CV look-up tables appropriate, the GVF method appears to be a viable replacement for the existing look-up tables. The GVF models are easy to use, are less burdensome than the CV look-up tables and according to the empirical results, are more accurate. In addition, the models can easily be programmed by the analysts or a basic program can be included with the PUMF documentation to calculate the approximate CV's.

For those users who wish to perform more complex analysis of Statistics Canada's data, the addition of bootstrap final weights to PUMF's will allow users to calculate correct design-based variance estimators (and hence CV's) for categorical and continuous variables as well as for complex statistics such as regression coefficients. The added flexibility of the bootstrap method does come at a price as there are still problems with confidentiality. The use of collapsing needs to be further investigated. In spite of this, the empirical comparisons with the currently used CV look-up method demonstrates the superiority of the bootstrap method both in terms of accuracy and the types of estimators for which it can be applied. While comparisons with the collapsed jackknife indicate only a slightly better performance for the bootstrap CV's, at this time the bootstrap methodology has better theoretical justifications.

7. REFERENCES

Brisebois, F. (1998). Détails Techniques Concernant la Construction du Calculateur de CV. Internal Statistics Canada document.

Canty, A.J. and Davidson, A.C. (1999). Resampling-based Variance Estimation for Labour Force Surveys. Unpublished manuscript.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia: Society for Industrial and Applied Mathematics.

Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953). *Sample Survey Methods and Theory* (Vol. 1). New York: John Wiley.

Kovačević, M.S., and Yung, W. (1997). Variance Estimation for Measures of Income Inequality. *Survey Methodology*, **23**, 41-52

Mantel, H. (1999). Private communication.

Mayda, J.E., Mohl, C., and Tambay, J.-L. (1996). Variance Estimation and Confidentiality: They are Related!, in *Proceedings of the Survey Methods Section*, Statistical Society of Canada, pp. 135-141.

National Population Health Survey Public Use Microdata File Documentation, Statistics Canada Publication Number 82F0001XCB

Phillips, O., and Kaushal, R. (1998). Methodology for CV Look-up Tables. Methodology Branch Working Paper, HSMD-98-001E. Statistics Canada.

Rao, J.N.K., and Shao, J. (1996). On Balanced Half-Sample Variance Estimation in Stratified Random Sampling. *Journal of the American Statistical Association*, **91**, 343-348.

Rao, J.N.K., and Wu, C.F.J. (1988). Resampling Inference with Complex Survey Data. *Journal of the American Statistical Association*, **83**, 231-241.

Rao, J.N.K., Wu, C.F.J., and Yue, K. (1992). Some Recent Work on Resampling Methods for Complex Surveys. *Survey Methodology*, **18**, 209-217.

Rust, K. (1986). Efficient Replicated Variance Estimation, in *Proceedings of the Survey Research Methods Section*, Washington, D.C.: American Statistical Association, pp. 81-87.

Valliant, R. (1987). Generalized Variance Functions in Stratified Two-Stage Sampling. *Journal of the American Statistical Association*, **82**, 499-508.

Valliant, R. (1995). Limitations of Balanced Half Sampling when Strata are Grouped, in *Proceedings of the Survey Research Methods Section*, Washington, D.C.: American Statistical Association, pp. 120-125.

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

Yeo, D., Mantel, H., and Liu, T.P. (1999). Bootstrap Variance Estimation for the National Population Health Survey, in *Proceedings of the Survey Research Methods Section*, Washington, D.C.: American Statistical Association, to appear.

Yung, W., and Rao, J.N.K. (1996). Jackknife Linearization Variance Estimators Under Stratified Multi-Stage Sampling. *Survey Methodology*, **22**, 23-31.

Yung, W., (1997). Variance Estimation for Public Use Files Under Confidentiality Constraints, in *Proceedings of the Survey Research Methods Section*, Washington, D.C.: American Statistical Association, pp. 434-439.