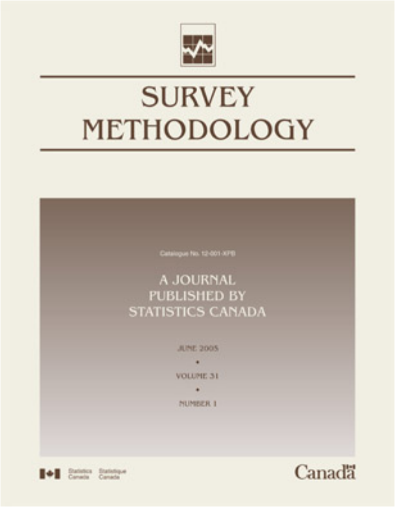
Catalogue no. 12-001-XIE

Survey Methodology

June 2004



SURVEY
METHODOLOGY

Catalogue No. 12-001-XPB

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2005
•
VOLUME 31
•
NUMBER 1

Statistics Canada Statistique Canada

Canada

Statistics Canada Statistique Canada

Canada

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: 1 800 263-1136).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our website.

National inquiries line	1 800 263-1136
National telecommunications device for the hearing impaired	1 800 363-7629
Depository Services Program inquiries	1 800 700-1033
Fax line for Depository Services Program	1 800 889-9734
E-mail inquiries	infostats@statcan.ca
Website	www.statcan.ca

Information to access the product

This product, catalogue no. 12-001-XIE, is available for free. To obtain a single issue, visit our website at www.statcan.ca and select Our Products and Services.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the Agency has developed standards of service that its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1 800 263-1136. The service standards are also published on www.statcan.ca under About Statistics Canada > Providing services to Canadians.

Statistics Canada

Business Survey Methods Division

Survey Methodology

June 2004

Published by authority of the Minister responsible for Statistics Canada

April 2006

Catalogue no. 12-001-XIE
ISSN 1492-0921

Frequency: semi-annual

Ottawa

Cette publication est disponible en français sur demande (n^{o} 12-001-XIF au catalogue).

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

Vol. 30, No. 1, pp. 35-44
Statistics Canada, Catalogue No. 12-001

Weighting Sample Data Subject to Independent Controls

Cary T. Isaki, Julie H. Tsay and Wayne A. Fuller [1]

Abstract

In the U.S. Census of Population and Housing, a sample of about one-in-six of the households receives a longer version of the census questionnaire called the long form. All others receive a version called the short form. Raking, using selected control totals from the short form, has been used to create two sets of weights for long form estimation; one for individuals and one for households. We describe a weight construction method based on quadratic programming that produces household weights such that the weighted sum for individual characteristics and for household characteristics agree closely with selected short form totals. The method is broadly applicable to situations where weights are to be constructed to meet both size bounds and sum-to-control restrictions. Application to the situation where the controls are estimates with an estimated covariance matrix is described.

Key Words: Raking; Regression; Quadratic programming; Coverage adjustment; Integer weights; Weighting area.

1. Introduction

Given the availability of known characteristic totals, it is common among survey practitioners to use such information in estimators of the post stratified, ratio and regression type. The known characteristic totals are sometimes called independent controls because they are derived outside of the survey situation. Use of independent controls tends to reduce the variance of most estimates. Independent controls also often compensate for coverage problems in surveys. See Deville and Särndal (1992) and Fuller (2002).

The U.S. decennial census utilizes a sample for the measurement of selected characteristics. The questionnaire for these characteristics is called the long form and the sample for the long form consists of a random sample of addresses. The long form questionnaire requests information that is asked of all individuals (called short form information) plus information on a set of additional characteristics. In previous Censuses, raking to controls based on short form information was used to construct weights for the long form sample. Two sets of sample weights were created, one for person characteristics and one for housing unit characteristics.

The set of categories used for person weighting was a classification of individuals by race, Hispanic origin, age and sex, family type, and household size. For households, the categories were the cross classification of race by Hispanic-origin-of-householder by tenure by household type and size. In the 1990 Census long form weighting process, persons and housing units were each classified by four sets of classifications for raking in four dimensions. When raking was completed, the long form sample weights were converted to integers. Integer weights are desirable because, unlike real weights, integer weights provide arithmetically consistent totals of integral characteristics. For details, see Schindler, Griffin and Swan (1992).

Long form weighting using short form census information is a part of the Canadian Census of population and housing. Unlike the procedure used by the U.S. Census Bureau (USCB), the procedure used at Statistics Canada constructs a single set of household weights using regression estimation. See Bankier, Houle and Luc (1997). Should the initial weights generated by the regression procedure exceed prescribed bounds, collapsing of cells defining explanatory variables is carried out. Linear dependencies and near linear dependencies among the explanatory variables are also removed by eliminating variables. See Bankier, Rathwell and Majkowski (1992).

Lemaître and Dufour (1987) used a generalized least squares estimator (GLS) to construct weights meeting person and household constraints. Alexander (1987) considers a procedure for constructing household weights in the census setting. One of his distance functions is similar to the one used in this paper.

The use of quadratic programming to compute regression weights in the survey context was suggested by Husain (1969). An application of quadratic programming (QP) in a Census environment is that in Isaki, Ikeda, Tsay and Fuller (2000) where household weights for Census households were obtained using person totals as controls. Motivation for the use of various distance functions can be found in these two papers and in Deville and Särndal (1992) who discuss a general class of estimators called calibration estimators. Fuller, Laughin and Baker (1994) consider a regression weight generation procedure that is modified so that all weights are positive and very large weights are made

[1] Cary T. Isaki and Julie H. Tsay, U.S. Bureau of the Census, Statistical Research Division, Washington, D.C. 20233, U.S.A. E-mail: Julie.Hsu.Ling.L.Tsay@census.gov; Wayne A. Fuller, Iowa State University, Department of Statistics, 221 Snedecor Hall, Ames, Iowa 50011, U.S.A.

smaller than the corresponding least squares weight. Jayasuriya and Valliant (1996) also consider a restricted regression. Fuller (2002) is a review of regression estimation.

Our proposed long form weighting method is a type of regression estimation and, like the Statistics Canada approach, provides a single set of household weights that maintain given independent controls. We generate household weights using quadratic programming with the restrictions that the weights fall within a specified range and that the weights maintain control totals. In the following, we refer to the suggested method as the quadratic programming method or QP.

2. The Quadratic Programming Method

The purpose of quadratic programming is to produce sample weights that i) are close to initial weights, ii) are within reasonable bounds, iii) maintain specified control totals and iv) produce a design consistent estimator. Apart from the bounds on the weight, the weights from quadratic programming are those of a simple regression estimator. We first describe the mathematical form of the QP and then discuss the implementation. Let

i) $\{W_i\,;\, i = 1, 2, ..., n\}$ denote the set of final housing unit weights, where i denotes the i^{th} long form sample household and n is the size of the long form sample,

ii) $\{W_i^{(2)};\ i = 1, 2, \ldots, n\}$ denote the set of initial housing unit weights,

iii) X_{ji}, $j = 1, 2, \ldots, m_p$, $i = 1, 2, \ldots, n$; denote the observation on the j^{th} person control variable for the i^{th} sample household,

iv) Z_{ji}, $j = 1, 2, \ldots, m_h$, $i = 1, 2, \ldots, n$; denote the observation on the j^{th} household control variable for the i^{th} sample household,

v) X_j, $j = 1, 2, ..., m_p$, denote the j^{th} person control,

vi) Z_j, $j = 1, 2, ..., m_h$, denote the j^{th} household control.

The quadratic programming method seeks W_i, $i = 1, 2, ..., n$, that minimize a quadratic objective function subject to linear constraints. In our application we minimize

$$g(W) = \sum_{i=1}^{n} \left(W_i - W_i^{(2)}\right)^2 \left[W_i^{(2)}\right]^{-1}, \qquad (1)$$

subject to

$$\sum_{i=1}^{n} W_i X_{ji} = X_j, \qquad \text{for } j = 1, 2, ..., m_p, \qquad (2)$$

$$\sum_{i=1}^{n} W_i Z_{ji} = Z_j, \qquad \text{for } j = 1, 2, ..., m_h, \qquad (3)$$

$$1 \le W_i \le K \qquad (4)$$

where the summations are over housing units in the long form sample. Observe that the long form household weights are bounded below by one. This is on the basis that an element in the sample should at least "represent" itself. In our program, K was set equal to 48 but the bound was never attained. The lower bound of one was attained. The FORTRAN subroutine from IMSL was used to solve the QP. Other programs, such as LCP of SAS®/IML, are available.

The USCB's current long form weighting procedure rakes the initially weighted long form sample counts to the census counts for the control categories. The weighting is done by subdivisions of the country called weighting areas and is done separately for person and household characteristics. The nominal sample rates for the long form are one-in-two, one-in-six, and one-in-eight. The nominal sampling weights are the inverses of the nominal sampling rates and are denoted by $W_i^{(1)}$. A second set of weights, denoted by $W_i^{(2)}$, are the realized sampling rates calculated for cells, where the cells are required to contain at least five sample households. For details on the USCB's procedures see Schindler *et al.* (1992).

Since we intend to compare the raking and QP methods, we use most of the USCB's person and household categories as the X_j and Z_j control totals in the quadratic program, but some changes were instituted. For example, while we maintained all of the age-race-sex person categories, we did not use a category based on the nominal sampling rates.

We used the USCB's specifications for determining whether a cell category would be retained as a separate control or would be combined with another cell and we used the USCB's procedure for determining the cells to be combined. This capitalized on the USCB's experience and minimized differences between the USCB's set of long form control totals and the set used by the QP method. The procedure used to define $W_i^{(2)}$ is given in the appendix.

Two possibilities exist for the control totals to be used in the construction of weights for the long form of the U.S. 2000 Census. One possibility is to use controls from the 2000 Census short form. That is, the independent controls to be maintained in long form weighting are those that are tabulated from the Census short form. When the Census is used as the control, the person control (X_j) categories include a cross classification of age and sex-race/ethnicity. Other characteristics, such as tenure, were used as additional controls. The majority of the household control categories (Z_j) are defined by a cross classification of household type (*e.g.*, family with children under 18) and household size (*e.g.*, number of persons in the family). The Z_j also include race/ethnicity of the householder cross-classified by tenure.

The other possible set of controls for the 2000 Census is the set of estimates from the post enumeration survey, called the Accuracy and Coverage Evaluation (A.C.E.) survey. The A.C.E. survey is designed to estimate person characteristics only. The X_j for the A.C.E. include age-sex-race/ethnicity-tenure controls.

The last step in long form weighting is to round the W_i to integers. Integer weights prevent discrepancies between sets of estimates caused by rounding of real valued estimates. Sample housing units were grouped by race/ethnicity of the householder and by tenure. Then within each group, the sample was sorted by family type by household size. The weights were then rounded to integers using the cumulate-and-round procedure. Table 1 illustrates the method. The partial sums of the weights are formed (cumulated) as shown in the column CW. The partial sums are then rounded as shown in the column RCW. The integer weight for element i is the difference between successive entries $i-1$ and i in the RCW column.

Table 1
Illustration of Cumulate and Round

Sample Unit	Initial Weight	CW	RCW	Integer Weight
1	3.333	3.333	3	3
2	2.500	5.833	6	3
3	1.428	7.261	7	1
4	1.250	8.511	9	2
5	1.111	9.622	10	1
6	5.021	14.643	15	5

3. Variance Estimation

Variances of long form estimates were estimated using the jackknife method. In the numerical results using census controls, sixteen replicates were formed. Sixteen was chosen for convenience and a larger number could have been used. The long form sample was ordered by the census identification number within blocks and sixteen replicates were formed as the sixteen one-in-sixteen systematic samples. Sixty seven replicates were formed for the estimates using ACE controls.

3.1 Replicates for Census Controls

The jackknife replicate is created by deleting the i^{th} group of elements, computing the quadratic programming weights and rounding the weights to integers. Because of the rounding, the usual jackknife variance estimation procedure required modification. To isolate the effect of rounding, we consider the replicate estimate constructed with real-valued weights. Let

$\hat{\theta}_w$ = the sample estimator with weights rounded to integers,

$\hat{\theta}_R$ = the sample estimator with real-valued weights,

$\hat{\theta}_{R(i)}$ = jackknife replicate estimate with i^{th} group deleted and real-valued weights,

$\hat{\theta}_{w(i)}$ = jackknife replicate estimate with i^{th} group deleted weights rounded to integers,

and let

$$\bar{\theta}_w = r^{-1}\sum_{i=1}^{r} \hat{\theta}_{w(i)}, \tag{5}$$

where r is the total number of replicates. Then the jackknife deviation for the estimator with integer weights can be decomposed as

$$\hat{\theta}_{w(i)} - \bar{\theta}_w = \hat{\theta}_{R(i)} - \hat{\theta}_R + \left[\hat{\theta}_{w(i)} - \bar{\theta}_w - (\hat{\theta}_{R(i)} - \hat{\theta}_R)\right]. \tag{6}$$

We assume that the error in the rounding operation is independent of the group chosen for deletion, a reasonable assumption, given that the deletion produces an entire new set of weights to be rounded. Then

$$E\left\{(\hat{\theta}_{w(i)} - \bar{\theta}_w)^2\right\} \doteq E\left\{(\hat{\theta}_{R(i)} - \hat{\theta}_R)^2\right\} + E\left\{\left[(\hat{\theta}_{w(i)} - \hat{\theta}_{R(i)}) - (\bar{\theta}_w - \hat{\theta}_R)\right]^2\right\}. \tag{7}$$

Assume that the average of the $\hat{\theta}_{R(i)}$ is equal to $\hat{\theta}_R$. Then the last term of (7) is a replicate deviation for the difference between the real and rounded estimates. Then

$$E\left\{\left[\left(\hat{\theta}_{w(i)} - \hat{\theta}_{R(i)}\right) - \left(\bar{\theta}_w - \hat{\theta}_R\right)\right]^2\right\} = r^{-1}(r-1)V\left\{\hat{\theta}_{w(i)} - \hat{\theta}_{R(i)}\right\} = V\left\{\hat{\theta}_w - \hat{\theta}_R\right\} \tag{8}$$

where $V\{\hat{\theta}_{w(i)} - \hat{\theta}_{R(i)}\}$ is the variance due to rounding for a sample of $r-1$ groups and $V\{\hat{\theta}_w - \hat{\theta}_R\}$ is the variance due to rounding for a sample of r groups. In obtaining (8) we assumed the variance due to rounding for a sample of r groups is the variance for $r-1$ groups multiplied by $r^{-1}(r-1)$. Thus

$$E\left\{(r-1)^{-1}\sum_{i=1}^{r}(\hat{\theta}_{w(i)}-\bar{\theta}_{w})^{2}\right\} \doteq E\left\{(r-1)^{-2} r \hat{V}_{R}\{\hat{\theta}_{R}\}\right\} + V\{\hat{\theta}_{w}-\hat{\theta}_{R}\}, \qquad (9)$$

where

$$\hat{V}_{R}\{\hat{\theta}_{R}\} = r^{-1}(r-1)\sum_{i=1}^{r}(\hat{\theta}_{R(i)}-\hat{\theta}_{R})^{2}$$

is the jackknife variance estimator for the estimator with real weights. Then an estimator of the variance due to rounding is

$$\hat{V}\{\hat{\theta}_{w}-\hat{\theta}_{R}\} = r^{-1}(r-1)\left[\begin{array}{l}(r-1)^{-1}\sum_{i=1}^{r}(\hat{\theta}_{w(i)}-\bar{\theta}_{w})^{2}\\ -r(r-1)^{-2}\hat{V}_{R}\{\hat{\theta}_{R}\}\end{array}\right] = r^{-1}\left[\begin{array}{l}\sum_{i=1}^{r}(\hat{\theta}_{w(i)}-\bar{\theta}_{w})^{2}\\ -r(r-1)^{-1}\hat{V}_{R}\{\hat{\theta}_{R}\}\end{array}\right]. \qquad (10)$$

Based on these results, the estimated variance for the rounded estimator is

$$\hat{V}\{\hat{\theta}_{w}\} = (r-1)^{-1}(r-2)\hat{V}_{R}\{\hat{\theta}_{R}\} + r^{-1}\sum_{i=1}^{r}(\hat{\theta}_{w(i)}-\bar{\theta}_{w})^{2}. \qquad (11)$$

3.2 Replicates for A.C.E. Controls

The replicates for estimates constructed with A.C.E. controls were modified so that the estimated variances contained a component for the error in the A.C.E. estimates. The data in a weighting area were assigned to 67 replicates where 67 is the number of controls. The procedure requires the number of replicates to equal or exceed the number of controls if the covariance matrix of the estimated control totals is to be reproduced. More replicates than controls can be used. See Fuller (1998).

The estimator of the total of a characteristic for the long form is a type of regression estimator using the A.C.E. numbers as controls. We write the estimator for the total based on real valued weights as

$$\hat{\theta}_{R} = \hat{\mathbf{X}}_{A}\hat{\boldsymbol{\beta}}, \qquad (12)$$

where $\hat{\mathbf{X}}_{A}$ is the vector of A.C.E. estimates and $\hat{\boldsymbol{\beta}}$ is the regression coefficient computed with the long form data.

Let $\hat{\mathbf{V}}_{AA}$ be the $r \times r$ covariance matrix of the vector of A.C.E. controls, where $\hat{\mathbf{V}}_{AA}$ is estimated as part of the A.C.E. process, and $r = 67$. Let $\lambda_1, \lambda_2, ..., \lambda_r$ be the roots of $\hat{\mathbf{V}}_{AA}$ and let

$$\mathbf{Q}'\mathbf{V}_{AA}\mathbf{Q} = \boldsymbol{\Lambda}, \qquad (13)$$

where $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_r)$, $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_r$, and $\mathbf{Q}$ is the matrix composed of the characteristic vectors of $\hat{\mathbf{V}}_{AA}$. Recall that

$$\hat{\mathbf{V}}_{AA} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}'$$

and

$$\hat{\mathbf{V}}_{AA} = \sum_{j=1}^{r}\mathbf{q}_{\bullet j}\lambda_{j}\mathbf{q}'_{\bullet j} =: \sum_{j=1}^{r}\mathbf{z}_{\bullet j}\mathbf{z}'_{\bullet j} \qquad (14)$$

where $\mathbf{q}_{\bullet j}$ is the j^{th} column of $\mathbf{Q}$ and $\mathbf{z}_{\bullet j} = \lambda_{j}^{1/2}\mathbf{q}_{\bullet \mathbf{j}}$.

Using result (14), controls for the r replicates were constructed as

$$\ddot{\mathbf{X}}_{A(i)} = \hat{\mathbf{X}}_{A} + c\mathbf{z}'_{\bullet i}, \; i = 1, 2, ..., r, \qquad (15)$$

where $\hat{\mathbf{X}}_{A}$ is the row vector of the original controls and c is a constant. The constant c is determined so that the expectation of the sum of the jackknife squared deviations for the elements of the vector $\mathbf{X}$ are the diagonal elements of $\hat{\mathbf{V}}_{AA}$. In our application, the constant c is $(r-1)^{-1/2}r^{1/2}$ and

$$(r-1)r^{-1}\sum_{j=1}^{r}c^{2}\mathbf{z}_{\bullet j}\mathbf{z}'_{\bullet j} = \sum_{j=1}^{r}\mathbf{z}_{\bullet j}\mathbf{z}'_{\bullet j} = \hat{\mathbf{V}}_{AA}. \qquad (16)$$

Thus, if the characteristic being "estimated" is one of the controls used in the QP, the jackknife procedure returns the A.C.E. estimated variance for that characteristic. The $\mathbf{z}_{\bullet j}$ are assigned at random to the r replicates.

Using the regression representation, we write the estimator for the i^{th} replicate as

$$\ddot{\theta}_{R(i)} = \ddot{\mathbf{X}}_{A(i)}\hat{\boldsymbol{\beta}}_{(i)} = \hat{\mathbf{X}}_{A}\hat{\boldsymbol{\beta}}_{(i)} + \left(\ddot{\mathbf{X}}_{A(i)} - \hat{\mathbf{X}}_{A}\right)\hat{\boldsymbol{\beta}}_{(i)} =: \hat{\theta}_{R(i)} + c\mathbf{z}'_{\bullet i}\hat{\boldsymbol{\beta}}_{(i)}, \qquad (17)$$

where $\ddot{\theta}_{R(i)}$ is the real-valued estimator computed with the i^{th} group deleted using $\ddot{\mathbf{X}}_{A(i)}$ as the control vector, $\hat{\boldsymbol{\beta}}_{(i)}$ is the regression coefficient computed with the i^{th} group deleted, and $\hat{\theta}_{R(i)}$ is the real-valued estimator computed with the i^{th} group deleted using $\hat{\mathbf{X}}_{A}$ as the control vector. Then

$$\ddot{\theta}_{R(i)} - \hat{\theta}_{R} = \hat{\theta}_{R(i)} - \hat{\theta}_{R} + c\mathbf{z}'_{\bullet i}\hat{\boldsymbol{\beta}}_{(i)}.$$

Because $\mathbf{q}_{\bullet j}$ are assigned to replicates at random, the expectation of the replicate variance estimator for the real-valued estimator based on A.C.E. controls is

$$E\left\{\hat{V}_R\left(\hat{\theta}_R\right)\right\} = E\left\{ r^{-1}\left(r-1\right)\sum_{i=1}^{r} \left(\ddot{\theta}_{R(i)} - \hat{\theta}_R\right)^2 \right\}$$

$$= E\left\{ r^{-1}\left(r-1\right)\sum_{i=1}^{r} \left(\hat{\theta}_{R(i)} - \hat{\theta}_R\right)^2 \right\} + E\left\{\hat{\boldsymbol{\beta}}'_{(i)}\hat{\mathbf{V}}_{AA}\,\hat{\boldsymbol{\beta}}_{(i)}\right\}. \quad (18)$$

Now, assuming $E\{\hat{\mathbf{V}}_{AA}\} = \mathbf{V}_{AA}, E\{\hat{\boldsymbol{\beta}}_{(i)}\} = \boldsymbol{\beta},$ and that $\hat{\mathbf{V}}_{AA}$ is independent of $\hat{\boldsymbol{\beta}}_{(i)}$,

$$E\left\{\hat{\boldsymbol{\beta}}'_{(i)}\hat{\mathbf{V}}_{AA}\,\hat{\boldsymbol{\beta}}_{(i)}\right\} = \boldsymbol{\beta}'\mathbf{V}_{AA}\,\boldsymbol{\beta} + tr\left\{V\{\hat{\boldsymbol{\beta}}_{(i)}\}\,\mathbf{V}_{AA}\right\},$$

where $\operatorname{tr}\{\mathbf{V}_{AA}\}$ is the trace of the matrix. It follows that

$$E\left\{ r^{-1}\left(r-1\right) \sum_{i=1}^{r} \left(\ddot{\theta}_{R(i)} - \hat{\theta}_R\right)^2 \right\}$$

$$= E\left\{ r^{-1}\left(r-1\right) \sum_{i=1}^{r} \left(\hat{\theta}_{R(i)} - \hat{\theta}_R\right)^2 \right\} + \boldsymbol{\beta}'\mathbf{V}_{AA}\,\boldsymbol{\beta} + O(n^{-2}), \quad (19)$$

where we assume $\operatorname{tr}\{\mathbf{V}_{AA}\} = O(n^{-1})$ and $\operatorname{tr}[V\{\hat{\boldsymbol{\beta}}_{(i)}\}] = O(n^{-1})$, where n is the sample size. The first term on the right of the equality in (19) is the expectation of the variance estimator for the variance due to the sampling of long forms from the census. The second term is the contribution of the variance of the error in the A.C.E. estimates to the total variance. Thus, the variance estimator based on $\ddot{\theta}_{R(i)}$ estimates both components of variation. Observe that the estimated covariance matrix for the controls is $\hat{\mathbf{V}}_{AA}$, as it should be.

4. Numerical Results

We used the USCB's 1990 Census data file to illustrate the application of the QP method to actual data. The file provides data for households and for persons in households, together with long form weights as developed for the 1990 U.S. Census. Hence, the file provides data appropriate for comparing the performance of the USCB's 1990 long form weighting method with the QP method.

The USCB long form sample weighting is done by weighting area, where the weighting areas usually contain two to three thousand housing units. There were about 56,000 weighting areas in the U.S. in 1990. For our numerical work we chose weighting area (WA) 1788 that contains 8,034 occupied housing units and 25,145 persons.

In Table 2 we provide estimates of some person and housing unit characteristics for weighting area 1788. The characteristics in the table, except the number of rented units, were suggested by subject matter personnel at the USCB. In Table 2, Est.(H) is the long form sample weighted estimate computed with housing unit weights, Est.(P) is the long form sample weighted estimate computed with person weights. The quadratic programming estimator constructed with Census controls is called QP in the table, while QPG is used to denote the generalization of the quadratic programming estimator with objective function (20). The QPG estimator is discussed subsequently. The USCB housing unit estimates in Table 2 that are based on person weights were created by using the householder weight as the housing unit weight. Every occupied unit contains a single householder. The householder procedure is called the *principal person method* by Alexander (1987). All estimates in the table are given as a percent of the census count.

Estimates constructed by the two USCB methods can differ by several percentage points with the differences between Est.(P) and Est.(H) for rented units, persons aged 0 to 4 years, persons aged 65 and over, Hispanic, Asian, and persons in rented units being noticeable. The Est.(H) estimate for persons in rented units is closer to 100 than the Est.(P) estimate.

The cell collapsing rules produced 45 person and 22 housing unit controls for WA 1788. An example of a person control is the total number of Non-Hispanic Black males aged 65 and over, while an example of a housing unit control is the total number of Non-Hispanic White owned housing units. Total Black persons is an implicit control in WA 1788. Controls for total persons 18-44, total persons 45-64, total males, total renters and total number of rented housing units were added to the QP. Apart from the controls mentioned above, none of the remaining characteristics in Table 2 is also used as a control in the QP procedure.

The QP estimates and standard errors of the QP estimates are given, as a percent of the census counts, in the fourth and fifth columns of Table 2. The agreement between count and QP estimates for household characteristics are comparable to the USCB household based estimates and superior to USCB person based estimates. For person counts, the QP estimates are generally closer to the census counts than either of the USCB raking estimates.

The largest difference between a QP estimate and the census count relative to the standard error is for the estimate of the number of households with own children present, where the difference is about 1.6 standard errors. The majority of the QP estimates differ from the census count by less than one standard error. A number of the USCB person estimates deviate from the census count by more than one QP standard error.

Table 2
Estimated Occupied Housing Unit and Person Characteristics for WA 1788

	Census Count	$\frac{\text{Est.(H)}^{*}}{\text{Count}}$ (%)	$\frac{\text{Est.(P)}^{**}}{\text{Count}}$ (%)	$\frac{\text{QP}^{\dagger}}{\text{Count}}$ (%)	$\frac{\text{se (QP)}}{\text{Count}}$ (%)	$\frac{\text{QPG}^{\dagger\dagger}}{\text{Count}}$ (%)	$\frac{\text{se (QPG)}}{\text{Count}}$ (%)
Housing unit characteristics							
With Own Children	4,349	100.18	100.45	100.21	0.13	100.18	0.14
Not With Own Children	3,685	99.78	99.67	99.76	0.15	99.78	0.16
With 1 to 4 Persons	6,785	100.00	100.57	100.04	0.05	100.07	0.05
With 5^{+} Persons	1,249	100.00	97.51	99.76	0.30	99.60	0.30
Rented Unit	2,559	100.00	95.97	100.00	0.19	99.92	0.16
Owned Unit	5,475	100.00	102.02	100.00	0.09	100.04	0.08
Person characteristics							
Age 0–4 years	2,493	101.92	97.95	98.84	1.68	99.96	0.29
Age 5–17 years	6,339	103.91	101.07	100.63	0.71	99.98	0.18
Age 18–44 years	12,711	99.50	99.69	100.01	0.05	100.00	0.06
Age 45–64 years	3,028	101.65	101.95	99.90	0.09	99.97	0.09
Age 65^{+} years	574	81.18	93.73	100.17	0.85	100.00	027
Males	12,473	99.95	99.64	100.06	0.08	99.98	0.09
Females	12,672	101.43	100.36	99.95	0.10	100.01	0.09
Hispanic	2,385	95.38	103.40	99.96	0.38	99.87	0.38
Not Hispanic	22,760	101.25	99.64	100.03	0.07	100.00	0.10
Black	1,285	101.08	101.79	100.86	1.22	99.77	0.54
White	22,372	100.69	99.91	100.03	0.07	100.00	0.10
Asian	257	92.60	80.05	96.83	2.32	99.76	0.50
Remainder	1,231	101.94	103.89	105.84	9.54	100.78	1.75
In Rented Unit	7,978	102.04	95.41	100.01	0.24	99.92	0.19
In Owned Unit	17,167	100.06	102.13	100.00	0.09	100.02	0.13

* USCB weights for households
** USCB weights for persons
† QP weights with 82 constraints
†† Generalized QP with 13 constraints and objective function (20)

Because the number of rented units, persons aged 18-44, persons aged 45-64, males, and persons in rented units were used as controls in the QP procedure, differences between QP estimates and census totals for those categories are due to rounding. The standard errors demonstrate that the rounding can lead to sizeable deviations from the controls.

The 45 person and 22 housing unit control totals obtained by the collapsing rules are such that a margin estimate, such as total males, may not be constrained to agree with the count. In addition, for different weighting areas, USCB's collapsing procedure gives different person and housing unit constraints. Thus we considered adding some margin totals to the set of control totals. To reduce the impact of the added controls on the weights, we replaced the original constraints with additional terms in the objective function. The terms are deviations between the final estimates and the control totals. The objective function becomes

$$G(W) = g(W) + \sum_{j=1}^{67} \alpha_j \left(\sum_i W_i X_{ji} - X_j \right)^2, \quad (20)$$

where $g(W)$ is defined in expression (1), the $\{X_{ji}, j = 1, 2, \ldots, 67\}$ is the set of auxiliary variables defining the 45 person and 22 housing unit controls, and α_j are constants to be specified. The X_{ji} for category j of household i for a person characteristic is the number of individuals in category j in the housing unit. The X_{ji} for a housing unit characteristic is one if the housing unit has the characteristic and zero otherwise. In our application, the function is minimized subject to two household controls and eleven person controls. The housing unit controls are rented housing units and owned housing units. The person controls are persons 0 to 4 years, persons 5 to 17 years, persons 18 to 44 years, persons 45-64 years, persons 65 years and over, males, black, white, Asian, Hispanic, and renters. The α_j are $10[\overline{W}^{(2)}]^{-1}[\sigma_j^2]^{-1}$, where $\overline{W}^{(2)} = 8.95$ is the mean of the $W_i^{(2)}$, $\sigma_j^2 = P_j(1-P_j)$, and P_j is the proportion of the population in cell j. The α_j would minimize the mean square error of an estimated total if there was a single control variable and the squared correlation between the control variable and the dependent variable was about 0.9. Thus, the function exerts considerable pressure for the final estimate to be close to the control total.

The QP solution to (20) gives a type of regression estimator. See Fuller (2002) and Fuller and Isaki (2001). Rao and Singh (1997) and Bardsley and Chambers (1984) consider related estimators.

Using $G(W)$ of (20) and the 13 linear constraints, the results in the final two columns of Table 2, under the heading "QPG", were obtained. As expected, the estimates are close to Census totals because the Census marginals were used as constraints. The relative percent differences between the QP estimate and the census count for the 67 characteristics in $G(W)$ of (20) ranged from –3.50% to 3.75% with about 50 of the differences being less than one percent.

The sample weights obtained by the two programming approaches are compared to those of the USCB's household raking method in Table 3. The number and type of controls used under the USCB raking was not determined exactly because the number depends on the execution of the USCB collapsing procedure and on some preliminary files that are not readily available. However, we believe the number to be about 67 because the collapsing procedure used to form the 67 cells is basically that used by the USCB. The QP procedure used 82 controls and the QPG procedure used 90 controls. The range of weights for the two QP methods are similar with a smaller range for raking. There are modest differences among the three sums of squares of the weights. The $g(W)$ values are also similar, with the value for (20) being the largest. The $g(W)$ value is the quantity being minimized by the weights of the first line of the table. The sum of squares of the weights for the QP of (20) could be reduced by reducing the α_j in the objective function.

We also used data from the 1990 Census to simulate the situation in which the controls come from adjusted census counts. For 1990, person estimates from the 1990 Post Enumeration Survey are available, but there are no housing unit estimates based on that survey. We call these estimates A.C.E. estimates. See Hogan (1993) and Isaki, Tsay and Fuller (2000). Estimates for WA 1788 were created by the QP method, using the A.C.E. estimates as controls. We used $G(W)$ of (20) as the objective function with 63 age-race-sex-tenure person characteristics in the second term of the objective function and 11 person constraints. The person constraints are persons 0 to 4 years, 5 to 17 years, 18 to 44 years, 45 to 64 years, 65 and over, total males, total Hispanic, total Black, total White, total Asian and total persons in rented units.

Table 3
Properties of Long Form Housing Unit Sample Weights in WA 1788

Method	Minimum Weight	Maximum Weight	$\sum_i W_i^2$	$g(W)$
QP with $g(W)$ of (1) 72 constraints	1	26.5	78,028	326
QP with $G(W)$ of (20) 13 exact constraints	1	29.9	78,672	383
Raking	4	22	77,000	369

Table 4 contains the estimates for WA 1788 identified as QPG and given as a percent of the census counts. The QPG estimates for these eleven person characteristics agree with the A.C.E estimates, except for rounding error. The standard errors reflect the error in the A.C.E estimates and, hence, are much larger than the standard deviation of rounding error. For example, the rounding error standard deviation for persons 18–44 is 0.06 in Table 2, while the standard error for the ACE estimate of persons 18–44 is 0.63. The QP estimates for household characteristics seem very reasonable. The estimated total number of households is 1.8% larger than the census count while the A.C.E. estimated number of persons is 2.0% larger than the census count. The quadratic programming total number of persons differs slightly from the A.C.E. estimate because of rounding of the weights. The difference is about 7% of the standard error.

Table 4
The Census Count, A.C.E. Estimates and QP Estimates with A.C.E. Controls – WA 1788

	Census Count	$\frac{\text{A.C.E}}{\text{Count}}$	$\frac{\text{QPG}}{\text{Count}}(\%)$	$\frac{\text{s.e.}(\text{QPG})}{\text{Count}}(\%)$
Housing unit characteristics				
With Own Children	4,349	–	101.89	2.09
Not With Own Children	3,685	–	101.66	3.07
With 1 to 4 Persons	6,785	–	102.03	2.03
With 5^+ Persons	1,249	–	100.40	5.92
Rented Unit	2,559	–	104.57	2.62
Owned Unit	5,475	–	100.47	1.50
Total	8,034	–	101.78	1.22
Person characteristics				
Age 0–4 years	2,493	103.17	102.81	1.00
Age 5–17 years	6,339	103.09	103.08	0.96
Age 18–44 years	12,711	101.67	101.67	0.63
Age 45–64 years	3,028	100.26	100.33	0.59
Age 65^+years	574	99.48	98.95	0.70
Males	12,473	102.18	102.01	0.68
Females	12,672	101.74	101.82	0.62
Hispanic	2,385	104.95	104.91	1.09
Not Hispanic	22,760	101.64	101.60	0.60
Black	1,285	104.59	104.82	1.01
White	22,372	101.69	101.69	0.61
Asian	257	100.00	101.95	1.95
Remainder	1,231	104.47	102.92	1.14
In Rented Unit	7,978	104.25	104.21	0.89
In Owned Unit	17,167	100.89	100.84	0.68
Total	25,145	101.96	101.91	0.57

5. Conclusions

The QP method is shown to work well on actual USCB long form data. The QP single household weight method possesses several advantages over the USCB separate weights method. With one set of weights, there will be no confusion as to which weights to use for estimating a given characteristic. Also, estimates of relationships such as ratios of person characteristics to household characteristics are expected to be less variable when a single set of weights is used for both characteristics.

Given that a single set of weights is easier to compute and easier for analysts to use, one would only construct two sets of weights if the weights designed for one type of characteristic give estimates with smaller variance for that type of characteristic. This did not seem to be the case in our example. The single set of QP weights gave favorable results for both household and person characteristics when compared with the USCB weights for the specific category.

The QP estimation module is computationally feasible and can replace the raking estimation module in the USCB operational setting. The QP method can produce long form sample weights for households in an adjustment situation in which only person controls are available.

Acknowledgements

This article reports the results of research and analysis undertaken by the authors. It has undergone a more limited review than official U.S. Census Bureau publications. Research results and conclusions expressed are those of the authors and have not been endorsed by the U.S. Census Bureau. The report is released to inform interested parties of research and to encourage discussion.

This research was partly supported by Cooperative Agreement 43-3AEU-3-80088 between Iowa State University, the National Agricultural Statistics Service and the U.S. Bureau of the Census.

We gratefully acknowledge the comments of the Associate Editor and two referees which led to a much improved paper.

Appendix

Procedure used to define cells and initial weights $W_i^{(2)}$

We used the USCB's procedure to determine the order in which cells are combined (collapsed). The cell collapsing rules specify that each cell contain at least 5 sample households. The procedure below is our extension of the USCB rules for defining $W_i^{(2)}$.

Let two cells under consideration be identified as Cell 1 and Cell 2.

i) Cell 1 is not to be collapsed and $n_1^{-1} N_1 \le B$, where N_1 is the Census count of households in Cell 1 and n_1 is the long form sample count in Cell 1. The constant B is provided by the sponsor and in our work, 27 is used. For household i in Cell 1, let

$$W_i^{(2)} = \max\{1.2, \ddot{W}_i\}, \qquad \text{(A.1)}$$

where $\ddot{W}_i = \min\{Q_1 W_i^{(1)}, B\}$,

$$Q_1 = \left[\sum_{i \in A_1} W_i^{(1)}\right]^{-1} N_1,$$

and A_1 is the set of indices in Cell 1. The number 1.2 is an arbitrary lower bound chosen greater than one and less than the minimum of $W_i^{(1)}$ which is two. Note that the $W_i^{(2)}$ provides reasonable estimated totals for Cell 1. If $n_1^{-1} N_1 > B$, collapse cell 1 with cell 2 as in ii) below.

ii) Cells 1 and 2 are designated for collapse, $(n_1 + n_2)^{-1}$ $(N_1 + N_2) \le B$, $n_1 + n_2 \ge 5$, and $n_1^{-1} N_1 > n_2^{-1} N_2$. Then for i in Cell 1, $W_i^{(2)}$ is defined by (A.1). For i in Cell 2,

$$W_i^{(2)} = \max\{1.2, \ddot{W}_i\},$$

where

$$\ddot{W}_i = \min\{Q_2 W_i^{(1)}, B\},$$

$$Q_2 = \left[\sum_{i \in A_2} W_i^{(1)}\right]^{-1} (N_1 + N_2 - \hat{N}_1),$$

and

$$\hat{N}_1 = \sum_{i \in A_1} W_i^{(2)}.$$

The $W_i^{(2)}$ in $A_1 \cup A_2$, the union of cells 1 and 2, maintains the total households in $A_1 \cup A_2$ and also provide an estimated total for Cell 1 that is reasonably close to the true total.

iii) Cells 1 and 2 are designated for collapse, $n_1 + n_2 \ge 5$, and $(n_1 + n_2)^{-1}$ $(N_1 + N_2) > B$. Then it is necessary to initiate further collapsing. The combined cell becomes the Cell 1 of case (ii). Continue cell collapsing until $(n_1 + n_2 + ..)^{-1}$ $(N_1 + N_2 + ..) \le B$. Case (iii) was not observed in the study data set.

One could repeat the weight construction procedure in an iterative manner by using the $W_i^{(2)}$ as $W_i^{(1)}$ in a second cycle. We tried a second cycle on the data described in the text. There was no discernable improvement in the estimates from using a second cycle.

References

Alexander, C.H. (1987). A class of methods for using person controls in household weighting. *Survey Methodology*, 13, 183-198.

Bankier, M.D., Rathwell, S. and Majkowski, M. (1992). Two step generalized least squares estimation in the 1991 canadian census. Working Paper-Methodology Branch, Census operations section, Social Survey Methods Division, Statistics Canada. SSMD92-007E.

Bankier, M., Houle, A.M. and Luc, M. (1997). Calibration estimation in the 1991 and 1996 canadian censuses. *Proceedings of the Section on Survey Research Methods,* American Statistical Association, 66-75.

Bardsley, P., and Chambers, R.L. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, 33, 290-299.

Deville, J., and Särndal, C.-E. (1992). Calibration estimators in survey sampling, *Journal of the American Statistical Association*, 87, 376-382.

Fuller, W.A. (1998). Replication variance estimation for two phase samples, *Statistica Sinica*, 8, 1153-1164.

Fuller, W.A. (2002). Regression estimation for survey samples. *Survey Methodology*, 28, 5-23.

Fuller, W.A., and Isaki, C.T. (2001). Estimation using estimated coverage in a census. Presented at the CAESAR conference, June, Rome, Italy.

Fuller, W.A., Laughin, M.M. and Baker, H.D. (1994). Regression weighting in the presence of nonresponse with application to the 1987-1988 Nationwide Food Consumption Survey. *Survey Methodology*, 20, 75-85.

Hogan, H. (1993). The 1990 post-enumeration survey: operations and results. *Journal of the American Statistical Association*, 88, 1047-1060.

Husain, M. (1969). Construction of regression weights for estimation in sample surveys. Unpublished M.S. Thesis, Iowa State University, Ames, Iowa.

Isaki, C.T., Tsay, J.H. and Fuller, W.A. (2000). Estimation of census adjustment factors. *Survey Methodology*, 26, 31-42.

Isaki, C.T., Ikeda, M.M., Tsay, J.H. and Fuller, W.A. (2000). An estimation file that incorporates auxiliary information. *Journal of Official Statistics*, 16, 155-172.

Jayasuriya, B.R., and Valliant, R. (1996). An application of restricted regression estimation in a household survey. *Survey Methodology*, 22, 127-137.

Lemaître, G., and Dufour, J. (1987). An integrated method for weighting persons and families. *Survey Methodology*, 13, 199-207.

Rao, J.N.K., and Singh, A.C. (1997). A ridge-shrinkage method for range-restricted weight calibration in survey sampling. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 57-65.

Schindler, E., Griffin, R. and Swan, C. (1992). Weighting the 1990 census sample. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 664-669.