Journal of Official Statistics, Vol. 16, No. 2, 2000, pp. 155–172

An Estimation File that Incorporates Auxiliary Information

C.T. Isaki[1], M.M. Ikeda[1], J.H. Tsay[1], and W.A. Fuller[2]

In the context of a population census employing estimation for undercoverage, we propose a method for constructing an estimation file that can be used for tabulations as if the file were a complete enumeration of the population. Such a data file is said to be transparent. Using data from the 1995 United States Test Census, we construct a transparent file and compare the person estimates for small areas to the corresponding synthetic estimates.

Key words: Post enumeration survey; regression; raking; controlled rounding; transparent file.

1. Introduction

In many situations, the survey statistician is asked to combine survey data and auxiliary information. Typically, the survey data are from a survey based on a probability sample. The auxiliary information may be in the form of known population totals, an example being the surface area of a study region. In other cases, the auxiliary information is in the form of estimates, for example estimates from the first phase of a two-phase sample. One of the methods of creating a tabulation data set in such situations is to use regression estimation. Typically, regression weights are constructed for the survey data such that the weights applied to the sample observations on the auxiliary variable will produce the external estimate.

The estimation problem we discuss is in the general category of sample data plus auxiliary information but has several unique characteristics. In our case, the sample is the data set obtained in a census operation. The number of persons recorded in the Census of Population and Housing conducted by the United States Bureau of the Census is generally agreed to be an undercount of the total number of persons in the country. Studies have

[1] U.S. Bureau of the Census, Statistical Research Division, Washington, D.C. 20233, U.S.A. E-mail: Cary.T.Isaki@ccmail.census.gov
[2] Iowa State University, Department of Statistics, 221 Snedecor Hall, Ames, IA 50011, U.S.A.
Acknowledgments: This article reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. The report is released to inform interested parties of research and to encourage discussion. Statistical Research Division, U.S. Bureau of the Census, Washington, D.C. 20233, U.S.A.
This article reports the results of research and analysis undertaken by W.A. Fuller. Research results and conclusions expressed are those of the author and have not been endorsed by the Census Bureau. The report is released to inform interested parties of research and to encourage discussion. Department of Statistics, Iowa State University, Ames, IA 50011, U.S.A.
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 thank the referees for comments that led to improvements in the manuscript.

been conducted to estimate the undercount by categories of individuals and a statistical method of estimating the total population is being planned as part of the 2000 U.S. Census. Our interest is in combining those estimates of the total population with the enumerated individuals to produce an estimation file that is indistinguishable from estimation files constructed for past censuses.

2. Previous U.S. Census Activities

The reporting unit in the U.S. Census for individuals living together in a household is called a *housing unit* and is what might be loosely called a ''family residence.'' Another reporting unit is called *group quarters* and includes units such as nursing homes and prisons. In what follows, we are primarily concerned with occupied housing units.

The basic data collection instrument for the U.S. Census is called the *short form*. Data elements include tenure class of the housing unit; the age, race, and sex of persons in the housing unit; and the relationships among members of the household. Tenure identifies whether or not the housing unit is rented or owned by the occupant. In the 1990 U.S. Census, short form data were collected from all persons who could be contacted. Census questionnaires were first mailed to all units. Then enumerators were sent to all nonresponding units to obtain short form information in an operation called nonresponse follow-up. We call the data collection operation the *enumeration phase*.

For purposes of conducting the Census, the U.S. Census Bureau subdivides the surface area of the country into sub-units. The smallest subdivision is called a *block*. While the name is in analogy to city blocks, Census blocks cover the entire country. The size of blocks is quite variable with some blocks containing no occupied housing units and some containing hundreds of housing units. Collections of contiguous blocks are called *tracts*. There are about six million blocks and about 56,000 tracts in the United States.

In 1990, a post enumeration survey (an independent enumeration of housing units in a sample of areas taken after the enumeration) was conducted and the results used to produce estimates of the total population. The Post Enumeration Survey is described in Hogan (1992, 1993). Thus, in 1990, there were two sets of population figures – one from the enumeration phase and one produced using the Post Enumeration Survey. The Secretary of Commerce (the Bureau's parent agency) chose the enumeration phase data as the official Census results.

The Post Enumeration Survey was used to create a file of estimated persons by block, but no estimates of housing units at the block level were attempted. The Post Enumeration Survey estimate of number of persons by category for the block was rounded to an integer and the difference between the Post Enumeration estimate and the enumeration count was assigned to the tabulation category of nonhousehold persons. For example, if the Post Enumeration Survey estimate of Black males aged 0 to 17 in a rented housing unit exceeded the enumeration count by two in a given block, then the data record for each of two Black males aged 0 to 17 in a rented housing unit in the block was duplicated (the household relationship variable was removed), and the two duplicate records assigned to a nonhousehold person category. The two records were assigned a weight of plus one. Conversely, if the enumeration exceeded the Post Enumeration Survey figure, data would be obtained in the same fashion except that the assigned weight would be a negative one. The weights were used to tabulate person counts. Data users were uncomfortable with the artificial category because the person estimates were really for persons in housing units. Reports from user groups indicate that block level person estimates based on the Post Enumeration Survey would be accepted by data users. However, use of the nonhousehold person category for assignment of such persons on the data file was judged undesirable because the estimates are for persons in households. Users expressed preference for a procedure that would give estimates for households with individuals in the households.

We describe a methodology that could be used to produce a data file of households for a census such as the 2000 U.S. Census that addresses the deficiencies described above. We present the results for such a file constructed for one of the 1995 U.S. Test Census sites.

3. Transparent File

There are many users of U.S. Census data. They represent the full range of statistical sophistication and are accustomed to tabular data for persons in households in blocks. Given that present plans call for the construction of estimated population numbers as part of the 2000 U.S. Census operation, methods of presenting these data must be developed. Clearly, a file that has the appearance of the traditional Census data file (households with person data and individuals in true group quarters), but yields the adjusted estimates, would be highly desirable for most users. A data file that can be tabulated as if it were a listing of the entire population is said to be a *transparent* Census file. For the U.S. Census, the transparent file would contain a listing of housing units, persons within the unit, their short form data, and the block identification of the housing unit. We briefly discuss the issue of disclosure relative to the transparent file in Section 5.1.

4. Methodology

4.1. Introduction

Present plans call for the 2000 U.S. Census field operation to begin mailing Census questionnaires to the Census list of addresses on or before April, 2000. A small portion of questionnaires will be delivered by enumerators. At a certain point, all addresses for which forms have not been returned or reported vacant by the postal system will be enumerated in the field. At the end of this operation, a file of completed forms, called the *Census file* or *Census Enumeration File*, will be created. This file is in the form of the traditional Census data file produced in past censuses. The U.S. Supreme Court has determined that the *Census file* must be used to determine representation of states in the national government.

As part of the 2000 Census field operation, a sample procedure similar to the PES methodology of the 1990 Post Enumeration Survey will be implemented. The sample and associated methodology is designed to provide estimates of total population by categories of persons for geographic areas. The geographic areas may each be a portion of a state or an entire state. We refer to the estimation operation as Integrated Coverage Measurement (ICM). Under ICM, no estimation of the number of households is planned.

It is planned to construct estimates of persons for finer levels of geography, such as city blocks, by synthetic estimation. An ICM estimate of total number of persons will be constructed for every block. The synthetic estimate for block k is the sum of the products of the number of Census persons in each category in the block multiplied by the estimated adjustment rate for that category. The adjustment rate for a category is the ICM estimate of number of persons in the category in the larger area, divided by the Census number of persons in the category in the larger area.

The ICM estimates and tabulations from the *Census file* will result in two sets of population numbers. The U.S. Census Bureau plans to designate the ICM estimates as the source of all official government figures. We describe a methodology for constructing a transparent Census data file under the ICM situation. Construction details are given in the remainder of this section and for a specific application in Section 5.

The inputs for transparent file construction are the Census Enumeration File, the ICM estimates of total persons by age-race-sex-tenure categories for large geographic areas, and the set of ICM estimates of total persons by block. Given the ICM estimates of persons, the task is to construct housing unit estimates for all Census blocks and to place the estimates in an easy-to-use format. This is done by constructing factors to be applied to the *Census file* housing unit counts to obtain estimates of housing units by blocks and converting the estimates to integer estimates of housing units for each block. Finally, households on the *Census file* are duplicated or deleted at the block level to obtain the desired transparent file. The duplication and deletion operations are described in Sections 4.3 and 5.1.

The Census operation classifies housing units as occupied or vacant. Occupied units are of most interest, but estimates of vacant units are also produced. We concentrate on procedures for occupied housing units.

4.2. Regression estimation

Regression estimation is our basic estimation technology. The regression model postulates that household characteristics are related to the person characteristics of household members. Because of the large number of households in the Census, we reduce the size of the regression problem by placing the observations into housing unit categories. The categories of occupied housing units were defined using characteristics associated with coverage. The factors include race of householder, age of householder, sex of householder, tenure, presence of spouse, and number of persons in the household. In addition, distinctions between certain minority households according to the presence or absence of certain types of persons were made. For example, a separate category for a three-person 30 to 49 aged Black female headed renter household with no spouse present but with young children less than ten years of age was used, separate from households with no children less than ten years old. It is important to separately categorize households with young children because young children were undercounted in 1990. Households that in the past were not subject to large coverage error, such as nonminority households, were placed in less detailed categories. The household categories are defined by variables that are closely related to the explanatory variables in the regression, but the explanatory variables do not completely define the categories.

In the Census, tabulation and estimation will be conducted for certain specified geographic units. The unit might be a state or a portion of a state. We call the units *large geographic areas* or *tabulation units.*

The objective of the first step of the estimation operation is to obtain weights for Census housing units. Two attributes of the weights are desired. First, the sum of the weights applied to Census individuals in a particular person category should give the ICM estimate of persons in that category for the large geographic area. Second, it is also desired that the sum of the weights applied to persons in a block be close to the ICM synthetic estimate of total persons for that block. The weight for an individual in a housing unit is the weight for the housing unit.

To describe the estimation procedure, let

$$\mathbf{x}_{ikt} = (x_{1ikt}, x_{2ikt}, \ldots, x_{m-1,ikt}, x_{m,ikt}, \ldots, x_{m-1+B,ikt})$$

be an $m-1+B$ dimensional vector, where B is the total number of blocks in the large geographic area, x_{jikt}, $j = 1, 2, \ldots, m-1$, is the number of persons in the t-th household of household category i of the k-th block that are in ICM category j, and

$$\begin{aligned} x_{m-1+k,ikt} &= \text{total number of persons in housing unit } ikt \text{ if the unit is in block } \\ &\quad k, k = 1, 2, \ldots, B \\ &= 0 \text{ otherwise.} \end{aligned}$$

Let h_{ik} be the number of housing units in category i in block k, let

$$\mathbf{x}_{ik} = \sum_{t=1}^{h_{ik}} \mathbf{x}_{ikt}$$

be the vector containing the number of Census persons in housing unit category i for the k-th block and let

$$\mathbf{x}_{i} = \sum_{k=1}^{B} \mathbf{x}_{ik}$$

be the vector of number of Census persons in the i-th housing unit category for the large geographic area. The synthetic estimator of the total number of persons for block k is

$$\hat{X}_{Tk} = \sum_{i=1}^{n} \sum_{j=1}^{m} x_{jik} R_{uj} \tag{4.2.1}$$

where x_{jik} is the number of individuals in ICM person category j in housing unit category i in block k in the Census, and R_{uj} is the ICM estimator of the ratio of total persons to census counted persons in person category j for the large geographic area. For example, x_{jik} may denote the number of Hispanic male renters less than 30 years of age in block k of Paterson, New Jersey, that reside in renter households containing five persons headed by a Hispanic male aged 30 to 49 and containing two or more young adult Hispanic males.

Let $\hat{X}_j$, $j = 1, 2, \ldots, m-1$, denote the ICM estimator of the total number of persons in category j for the large geographic area, and let X_{cj}, $j = 1, 2, \ldots, m-1$, denote the Census number of persons in category j. One regression estimator can be written as

$$\hat{Y} = Y_C + (\hat{\mathbf{X}} - \mathbf{X}_C)\hat{\boldsymbol{\beta}} \tag{4.2.2}$$

where $\hat{\mathbf{X}} = (\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_{m-1}, \hat{X}_{T1}, \hat{X}_{T2}, \ldots, \hat{X}_{TB})$ is the vector of ICM estimates, $\hat{\boldsymbol{\beta}}$ is the regression coefficient for the weighted regression of y_{ik} on $\mathbf{x}_{ik}$

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \sum_{k=1}^{B} h_{ik}^{-1} \mathbf{x}'_{ik} \mathbf{x}_{ik} \right)^{-1} \sum_{i=1}^{n} \sum_{k=1}^{B} h_{ik}^{-1} \mathbf{x}'_{ik} y_{ik}$$

Y_C is the census total for the characteristic, $\mathbf{X}_C$ is the census total for $\mathbf{X}$, and y_{ik} is the sum of the y-characteristic for housing units in block k in housing unit category i. The regression estimator can be written as

$$\sum_{i=1}^{n}\sum_{k=1}^{B} w_{ik} y_{ik} \tag{4.2.3}$$

where

$$w_{ik} = 1 + (\hat{\mathbf{X}} - \mathbf{X}_C)\left(\sum_{i=1}^{n}\sum_{k=1}^{B} h_{ik}^{-1}\mathbf{x}_{ik}'\mathbf{x}_{ik}\right)^{-1} h_{ik}^{-1}\mathbf{x}_{ik}'$$

One way to compute the regression weights of (4.2.3) is to minimize

$$\sum_{i=1}^{n}\sum_{k=1}^{B}(w_{ik} - 1)^2 h_{ik} \tag{4.2.4}$$

subject to

$$\sum_{i=1}^{n}\sum_{k=1}^{B} w_{ik}\mathbf{x}_{ik} = \hat{\mathbf{X}}$$

Alternative regression estimators can be computed by generalizing (4.2.4) to

$$\sum_{i=1}^{n}\sum_{k=1}^{B}(w_{ik} - w_{(o)ik})^2 w_{(o)ik}^{-1} h_{ik}$$

where $w_{(o)ik}$ is an initial weight. For example $w_{(o)ik}$ might be the ratio of the ICM estimate to the Census estimate for the person category of the householder. If one were dealing with a probability sample, the initial weight would be the reciprocal of the sampling fraction. Husain (1969) described using quadratic programming to obtain the w_{ik}.

We modify the basic regression procedure in several ways. First, we introduce a bound on the weight. The bound is used to guarantee positive weights and, hence, to guarantee that estimates of positive quantities are positive. Second, because B, the number of blocks, is very large, we use an iterative raking operation to approximate the restrictions that the weights applied to the households in a block equal the ICM synthetic estimate of total persons for the block. The weight construction procedure iterates between a regression weight for the large geographic area controls and a ratio weight for the total population controls for each block.

The adjustment to bring the estimated persons into agreement with the synthetic block estimate is

$$a_{rk} = \hat{X}_{Tk}\left(\sum_{i=1}^{n}\sum_{j=1}^{m} c_{r-1,i} x_{jik}\right)^{-1} \tag{4.2.5}$$

where c_{0i} is the initial weight for an observation in the Census Enumeration File in category i and r is the iteration index. We report results for $c_{0i} = 1$. Let h_i, $i = 1, 2, \ldots, n$, denote the number of occupied units in household category i in the Census file for the large geographic area. Given the $c_{r-1,i}$, a new vector of weights $\{c_{r,1}, c_{r,2}, \ldots, c_{r,n}\}]$ is chosen to minimize

$$f(c) = \sum_{i=1}^{n}(c_{ri} - c_{0i})^2 h_i c_{0i}^{-1} \tag{4.2.6}$$

subject to

$$\sum_{i=1}^{n} c_{ri} \left(\sum_{k=1}^{B} a_{rk} x_{jik} \right) = \hat{X}_j, \quad j = 1, 2, \ldots, m$$

$$c_{ri} > K > 0 \qquad i = 1, 2, \ldots, n$$

where K is a chosen lower bound for the weights. The c-weights that minimize (4.2.6) are then used in (4.2.5) to obtain new a_{rk}, etc. The final estimation weight for housing unit type i in block k is

$$w_{ik} = c_{Li} a_{Lk} \tag{4.2.7}$$

where L is the index of the final iteration. In practice, one may fix L in advance, or use a convergence criterion to stop iteration. Given the weights of (4.2.7), the estimator is

$$\sum_{i=1}^{n} \sum_{k=1}^{B} w_{ik} y_{ik}$$

The weight construction combines elements of raking and least squares regression estimation. Related regression estimators have been considered by Huang and Fuller (1978), Bethlehem and Keller (1987), and Deville and Särndal (1992). Zaslavsky (1988) suggested a procedure for constructing weights for households under the assumption that adjusted estimates of households are available by block.

If we ignore the lower bound restrictions on the weights, the variance of the estimator defined by (4.2.5) and (4.2.6) can be approximated by the variance of the corresponding regression estimator. See, for example, Deville and Särndal (1992) and Deville, Särndal and Sautory (1993). Let Y be the unknown finite population total for the characteristic of interest, let $\mathbf{X}$ be the unknown vector of finite population totals for the x-variables, let $\mathbf{x}_{Tik}$ be the unknown population total of the $\mathbf{x}$ vector for housing unit category i in block k, let y_{Tik} be the population total for housing unit category i in block k, and let h_{Tik} be the population number of households in housing unit category i in block k.

For the purpose of variance estimation, assume the superpopulation regression model

$$y_{Tik} = \mathbf{x}_{Tik} \boldsymbol{\beta} + e_{Tik} \tag{4.2.8}$$

where the e_{Tik} are zero mean independent random variables, and $E\{\mathbf{x}_{Tik} e_{Tik}\} = \mathbf{0}$. The error in the regression estimator (4.2.2) of Y can be written

$$\hat{Y} - Y = (\hat{\mathbf{X}} - \mathbf{X})\boldsymbol{\beta} + (\mathbf{X} - \mathbf{X}_C)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + e_C - e_T + O_p(s^{-1}) \tag{4.2.9}$$

where $\mathbf{X}_C$ is the total of the $\mathbf{x}$-vectors for the Census, e_C is the total of e_{Tik} for the Census, e_T is the total of e_{Tik} for the population, and s is an index of sample size. In our case s is the number of housing units in the Census.

Under the assumptions that the error in $\hat{\boldsymbol{\beta}}$ is independent of the error in $\hat{\mathbf{X}}$, that the covariance between $e_C - e_T$ and $\hat{\boldsymbol{\beta}}$ can be ignored, and that the errors in $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{X}}$ decline at the rate $s^{-1/2}$, an estimator of the variance of $\hat{Y} - Y$ is

$$\hat{V}\{\hat{Y} - Y\} = (\hat{\mathbf{X}} - \mathbf{X}_C)\hat{\mathbf{V}}\{\hat{\boldsymbol{\beta}}\}(\hat{\mathbf{X}} - \mathbf{X}_C)' + \hat{\boldsymbol{\beta}}' \, \hat{\mathbf{V}}\{\hat{\mathbf{X}}\}\hat{\boldsymbol{\beta}} + \hat{V}\{e_C - e_T\} \tag{4.2.10}$$

where $\hat{\mathbf{V}}\{\hat{\mathbf{X}}\}$ is the estimated covariance matrix of the vector of ICM estimates, and $\hat{\mathbf{V}}\{\hat{\boldsymbol{\beta}}\}$ is an estimator of the covariance matrix of $\hat{\boldsymbol{\beta}}$.

4.3. Integer estimates

The w_{ik} are weights for occupied units in the Census file in housing unit category i and block k. Given that h_{ik} denotes the number of occupied housing units in category i and block k in the Census Enumeration File, then $w_{ik}h_{ik}$ is the estimator of the number of housing units in category i in block k.

To make the estimates physically realistic, the category estimates for each block are converted to integers. The weights in Equation (4.2.7) for the U.S. Census are weights for housing units, but the controls are in terms of persons. Hence, the usual types of controlled rounding are not directly applicable, and a sequential procedure was adopted. In the first step, the controlled rounding procedure of Cox and Ernst (1982) was applied to obtain integer estimates of households for each category for each block. Let t_{ik} denote the integer-valued estimate of housing unit counts in category i in block k obtained by controlled rounding. Then $U_{ik} = t_{ik} - h_{ik}$ represents an undercount for category i in block k if U_{ik} is positive and represents an overcount for category i in block k if U_{ik} is negative. The data set is created by either adding category i housing units to block k of the data file if U_{ik} is positive, or removing category i housing units from block k of the data file if U_{ik} is negative.

The required housing units are selected at random from the housing units in the block. A sequential selection procedure is used in an attempt to closely approximate the synthetic estimate of number of persons. In the first step, the required number of housing units with more than three persons is selected. As households with more than three persons are selected, a cumulative count of total persons (those in the original Census plus those in selected households) is maintained and compared to the synthetic estimate of persons. Household selection is terminated when the total number of housing units reaches t_{ik} or when the selection of an additional household produces a sum of individuals (the original census plus individuals in the selected households) that exceeds the synthetic estimate of total persons for the block.

For the large geographic area, let the vector of the number of persons in each category in households with four or more persons be $\hat{\mathbf{X}}_{(1)}$. Then a new control vector for one-, two-, and three-person households is

$$\hat{\mathbf{X}}_{(2)} = \hat{\mathbf{X}} - \hat{\mathbf{X}}_{(1)}$$

where $\hat{\mathbf{X}}$ is the original vector of controls. A set of weights for the one-, two-, and three-person households is computed beginning with the block factor

$$a_{4k} = (\hat{X}_{Tk} - \hat{\mathbf{X}}_{(1)k})\left(\sum_{i\in(1,2,3)}\sum_{j=1}^{m} c_{4i}x_{jik}\right)^{-1}$$

where the summation over i is for one-, two-, and three-person households. Using this block factor, a new set of c-factors is computed for the categories of one-, two-, and three-person households as described previously.

A controlled rounding procedure is applied to obtain integer estimates for the one-, two-, and three-person households. Then two- and three-person households are selected for duplication or deletion for each block. As at the first step, household selection proceeds so that the sum of individuals, in the original and duplicated households, is less than or equal to the synthetic estimate of total persons for the block.

Finally, estimates are created for the one-person households using the control total

$$\hat{\mathbf{X}}_{(3)} = \hat{\mathbf{X}} - \hat{\mathbf{X}}_{(1)} - \hat{\mathbf{X}}_{(2)}$$

where $\hat{\mathbf{X}}_{(2)}$ is the vector of totals from the created file of two- and three-person households.

5. Application to the 1995 U.S. Census Pretest

5.1. Numerical results

A Census pretest, called the *95 Census Test*, was conducted in Paterson, New Jersey and Oakland, California. The test was designed as a test of methodology for sampling nonrespondents and as a test of the ICM post enumeration measurement methods. Transparent files were constructed for both the Paterson and Oakland sites with similar results. We present estimates for the Paterson site. The 1995 Test Census reported 127,954 persons in 42,516 occupied housing units and 3,239 vacant units in Paterson. The direct ICM person estimate of number of persons was 145,508. There are 984 blocks in Paterson, ranging in size from one to 792 housing units. Twelve percent of the blocks contained fewer than eleven housing units.

The estimator is of the form

$$\hat{Y} = \sum_{k=1}^{B} \sum_{i=1}^{n} w_{ik} y_{ik} \tag{5.1.1}$$

where y_{ik} is the census total for household category i in block k. The number of housing unit categories was about 350 and the number of person categories was 42. The weights were chosen to minimize (4.2.4), with three iterations of the step (4.2.5). The weights had largely stabilized at the end of three iterations. The constant K in (4.2.6) was set equal to 0.5 in the minimization. It would be desirable to have a lower limit of one so that any enumerated household would be included in the file. This is not possible because misreporting of household composition results in misclassification. In addition, the enumeration phase produces duplicate housing units. The value of 0.5 was a relatively large value for which the quadratic program converged.

The rounding and imputation steps were carried out as described in Section 4.3. In addition, controls based on the synthetic estimator of race groups were imposed at the tract level. The race groups were Black, Hispanic, and Other. The one-person household counts summarized on a tract by race basis were allocated to one-person housing unit categories in proportion to their count in the transparent file. The race estimates served as one set of controls in a two-way raking procedure applied to one-person housing unit counts. The other controls were synthetic estimates of persons in one-person households by block, where the one-person numbers are totals for the blocks less persons in households of two or more persons in the created data file. The raked one-person households were then control rounded and households duplicated or deleted to complete the transparent file.

In selecting donor households, we randomly selected as donors housing units from among the h_{ik} housing units in the Census Enumeration File beginning with Census housing units in block k. In a few cases, some of the h_{ik} housing units were used as donors three times.

Table 1. Household Census count, regression estimate (5.1.1), and standard error of the regression estimate for Paterson, NJ; 1995 U.S. Census Pretest

Type of household	Census	Estimate	s.e.
Black householder	15,387	18,041	860
Hispanic householder	14,764	16,906	442
Other householder	12,365	12,852	341
Owner	13,407	13,812	443
Renter	29,109	33,987	947
Total	42,516	47,799	996

The source of the housing units was used in determining which units were to be deleted. The deletions were first selected from imputed housing units, where imputed units are units on the Census Enumeration File that were either identified as not containing sufficient information or were nonrespondents to the mailing and not selected for nonresponse follow-up. If more housing units than the number of imputed units in the block were required for deletion, then other units were deleted. Real respondents that were deleted from a block were used to replace housing units in the same housing unit category in another block. This minimized the number of actual responding housing units deleted in the file creation operation.

To prevent disclosure, the U.S. Bureau of the Census exchanges similar households between blocks on the data file. A positive feature of the transparent file is that less exchange of households is required for disclosure limitation than is required for the enumeration file because households are shifted to adjacent blocks in the construction of the transparent file.

Table 1 contains estimates for some types of households. Details on the construction of the standard errors are given in the Appendix. The standard errors are approximations for the regression estimator. They do not contain components for imputation or rounding error. It is estimated that there is about a 12.4 percent undercount in the number of households, compared to the 13.7 percent undercount in persons. The percentage undercount is largest for black households and renter households. The estimated undercount is several times the standard error except for the two categories ‘‘Other householder’’ and ‘‘Owner.’’

In constructing the transparent file, we used only the ICM person estimates. The procedure could easily accommodate controls on estimated households. Isaki and Ikeda (1996) employed the ICM data to construct direct estimates of housing units. They estimated 18,496 Black; 16,724 Hispanic; 12,813 ‘‘other’’; 13,694 owner; and 33,432 renter households. These estimates differ from those in Table 1 by less than one standard error.

We summarize some properties of the ICM estimates and the constructed transparent file for Paterson. Table 2 contains enumerated persons as obtained in the Census operation, as estimated in the ICM, and as they appear in the transparent file for several person categories. The last column gives the differences between the ICM estimates and the transparent file. The differences are due entirely to rounding. Rounding includes rounding to integers and the effect of donor selection of households of different sizes in constructing the transparent file. There would be zero difference if real valued household weights were used. The differences are negligible, relative to the standard errors of the ICM estimators.

Table 2. Census Person count, ICM person estimate, standard error of ICM estimate, and Transparent file person count for Paterson, NJ: 1995 U.S. Census pretest

Category	Census File	ICM Estimate	s.e. of ICM	Transparent File	Differences ICM-T.File
Black Persons	46,673	56,260	2,159	56,266	−6
Black owned	13,767	14,487	767	14,494	−7
Black owned aged 0 to 17	3,528	3,858	248	3,859	−1
Black rented	32,906	41,773	2,251	41,772	1
Hispanic persons	52,268	59,476	1,288	59,474	2
Hispanic owned	14,625	15,929	538	15,915	14
Hispanic rented	37,643	43,547	1,086	43,559	−12
Other persons	29,012	29,772	753	29,767	5
Other owned	14,863	14,741	445	14,713	28
Other rented	14,149	15,031	590	15,054	−23
TOTAL PERSONS	127,954	145,508	2,746	145,507	1

Clearly the difference in total persons could be reduced to zero by adding a one-person household to the transparent file.

The coefficient of variation of the ICM estimate of total persons is 1.9%, while the coefficient of variation for the regression estimator of total households in Table 1 is 2.1%. As one might expect, the coefficient of variation is slightly larger for households because of the additional estimation involved.

The entries in Table 2 are, in general, marginal categories of the 42 race-tenure-age-sex categories in the ICM. Only the category Black owned aged 0 to 17 is one of the original categories. Of the 42 categories, the absolute difference between the ICM and transparent file estimates ranged from 0 to 28 persons. The largest ratio of the absolute difference to the standard error of the ICM estimate was less than one-sixth.

Our research began under the assumption that the transparent file would provide the official estimates of both persons and households. Under this assumption, rounding error that is small relative to estimation error is not important. In our original formulation, we constructed the regression estimator without controlling to the block synthetic estimator. Results of this estimation were presented to the National Academy of Sciences Panel on Census 2000 Methodology. It was the panel's opinion that household estimates should give block estimates of persons ''close'' to the block synthetic estimates (National Research Council 1999, p. 65–66). Furthermore, it was the opinion of the panel that the block estimates of number of persons constructed by the regression estimator using controls for the large geographic area were ''not close enough'' to the synthetic block estimates. In response to the panel's position, we developed the procedure described in the previous section.

Table 3 contains measures of closeness between the ICM synthetic person estimates and those from the transparent file for blocks and tracts in Paterson. Let TF_k, SYN_k and CEN_k denote the transparent file, synthetic estimate, and Census count for a characteristic in the k-th area, where an area can be a block or a tract, and let N denote the number of areas. Define

i) Mean Squared Difference $= MSD = N^{-1}\Sigma_{k=1}^{N}(TF_k - SYN_k)^2$

Table 3. Summary statistics of transparent file and synthetic person estimates at the block and tract levels for Paterson, NJ: 1995 U.S. Census pretest

A. Paterson-Blocks	Number of nonzero blocks	Mean squared difference	Mean absolute difference	5% Point distribution of RD_k	95% Point distribution of RD_k	Mean of TF_k	Mean of CEN_k
Characteristic							
Total Persons	984	0.19	0.31	−0.02	0.01	147.87	130.03
Owners	941	19.00	3.05	−0.18	0.19	47.95	46.00
Renters	970	18.54	2.99	−0.15	0.14	103.49	87.28
Blacks	909	14.30	2.68	−0.29	0.32	61.90	51.09
Hispanics	924	16.94	2.97	−0.17	0.20	64.37	56.72
Others	905	7.78	1.72	−0.23	0.33	32.89	32.15

B. Paterson-Tracts	Number of tracts	Mean squared difference	Mean absolute difference	5% Point distribution of RD_k	95% Point distribution of RD_k	Mean of TF_k	Mean of CEN_k
Characteristic							
Total Persons	33	0.09	0.26	−0.001	0.0002	4409.30	3877.15
Owners	33	450.39	16.39	−0.029	0.044	1368.39	1311.58
Renters	33	451.94	16.43	−0.015	0.015	3041.97	2565.58
Blacks	33	1.55	0.99	−0.002	0.004	1705.03	1407.42
Hispanics	33	1.63	1.00	−0.002	0.002	1802.24	1588.09
Others	33	2.05	1.15	−0.014	0.012	902.03	881.64

ii) Mean Absolute Difference $= MAD = N^{-1}\Sigma_{k=1}^{N}|TF_k - SYN_k|$
iii) Relative Difference $= RD_k = (TF_k - SYN_k)[CEN_k]^{-1}$

The transparent file procedure imposed control for race at the tract level. Hence, the race differences at the tract level are due to rounding, where the original tract synthetic estimates were not rounded. Because there was no direct control for tenure at the tract level, the absolute differences are larger for the tenure categories than for the race categories.

The differences at the block level in total persons are due to rounding. The synthetic estimators were not rounded, and the mean squared difference is less than twice the squared difference due to rounding to integers. The differences in other categories are larger than those for total persons, because no direct restrictions were imposed on those categories at the block level.

The five percent and 95 percent points of the empirical distribution of the relative differences are also given in Table 3. The relative differences of large absolute value are associated with blocks with small numbers of persons. For example, of the 50 blocks with relative differences exceeding the 95 percent point, the average numbers of persons were 16 Black, 19 Hispanic and 6 Other, respectively. In comparison, the average numbers per block in the site were 51 Black, 56 Hispanic, and 32 Other.

5.2. Computing

Throughout the development process, we imposed the requirement that all procedures be such that they could be implemented in the large-scale Census operation. The quadratic program was the least computer intensive of the operations required to create a transparent file. The quadratic programming phases of the operation took about two minutes on a Sparc 40, using a SUN OS4.1.3 operating system. The controlled rounding for the 984 blocks in Paterson took about 12 minutes and the donor selection took about ten minutes at each phase. To fully implement the procedure in the Census would require larger machines or a bank of machines. Improvements in the time required for transparent file construction are possible within the current paradigm by reducing the requirements on the degree of agreement between the block synthetic estimate and the block transparent file. Also, some programming efficiencies could be implemented in a final production program.

6. Comments

In a recent action, the U.S. Supreme Court ruled that estimates of the ICM type cannot be used for the allocation of congressional seats. However, the Court also ruled that ICM estimates can be used for other purposes. Thus, the suggested transparent file could be used to produce tabulations based on short form data similar to those provided in previous censuses, especially tabulations for households. Such tabulations could be used for purposes such as fund allocation and would be constructed after the population counts required for apportionment are released. Also, the term ICM is no longer being used by the Bureau. Because of changes in design, the successor to ICM is called *Accuracy and Coverage Evaluation* (A.C.E.).

A second possible use of the transparent file procedures is for long form data. Long form Census data on items such as utility use and cost, are collected from a sample of households receiving Census forms. In the past, the U.S. Census provided two sets of weights for the long form sample data. One set of weights were to be used for estimating person characteristics and a second set of weights were to be used for estimating household characteristics. The two sets of weights were obtained in two separate raking operations. Many characteristics, such as total rented Hispanic housing units, can be estimated using either set of weights and the two estimates are almost always different. The quadratic programming procedure could be used to provide a single set of housing unit weights for estimating both person and housing unit characteristics. A similar procedure based on the regression approach has been used by Statistics Canada (Bankier 1997) to produce a single set of weights.

Appendix

Standard error calculations

The standard errors in Table 1 of the text were computed using an approximation to equation (4.2.10). For variance calculation purposes, we worked with the 42,516 housing units in the census. For the i-th housing unit category in the k-th block, we observe

$$\mathbf{z}_{ik} = (x_{1ik}, x_{2ik}, \ldots, x_{m-1,ik})$$

and x_{Tik}, where $\mathbf{z}_{ik}$ is the vector of person counts in $m-1$ of the original m person categories and x_{Tik} is the total number of persons in the i-th housing unit category in block k. The vector of regression variables is the vector $\mathbf{z}_{ik}$ plus 984 block-variables, where the value for the k-th block is equal to x_{Tik} for the i-th household category in the k-th block if the observation is in the k-th block, and is zero otherwise.

The regression equation (4.2.2) contains $(42 - 1 + 984)$ regression variables, where 42 is the number of person categories and 984 is the number of blocks. To calculate the coefficients for the elements of $\mathbf{z}_{ik}$, we created the vectors

$$(y_{ik}, \mathbf{z}_{ik}) - \mathbf{b}'_k x_{Tik} = (u_{ik}, \tilde{\mathbf{z}}_{ik}) \quad k = 1, 2, \ldots, 984 \tag{A.1}$$

where

$$\mathbf{b}'_k = (b_{xk}), \mathbf{b}'_{xk} = \left[\sum_{i=1}^{n} h_{ik}^{-1} x_{Tik}(y_{ik}, \mathbf{z}_{ik})\right]\left[\sum_{i=1}^{n} h_{ik}^{-1} x_{Tik}^2\right]^{-1}$$

is the vector of coefficients for the regression of $(y_{ik}, \mathbf{z}_{ik})$ on x_{Tik} in the k-th block. Then the vector of coefficients for $\mathbf{z}_{ik}$ is the coefficient vector obtained in the regression of u_{ik} on $\tilde{\mathbf{z}}_{ik}$. Let this vector of coefficients be denoted by $\hat{\boldsymbol{\beta}}_{m-1}$, where

$$\hat{\boldsymbol{\beta}}_{m-1} = \left(\sum_{k=1}^{B}\sum_{i=1}^{n} h_{ik}^{-1} \tilde{\mathbf{z}}'_{ik} \tilde{\mathbf{z}}_{ik}\right)^{-1} \sum_{k=1}^{B}\sum_{i=1}^{n} h_{ik}^{-1} \tilde{\mathbf{z}}'_{ik} u_{ik} \tag{A.2}$$

An estimator of the variance of $\hat{\boldsymbol{\beta}}_{m-1}$ is

$$\hat{\mathbf{V}}\{\hat{\boldsymbol{\beta}}_{m-1}\} = \hat{\mathbf{A}}_{m-1}^{-1} \mathbf{G}_{m-1} \hat{\mathbf{A}}_{m-1}^{-1} \tag{A.3}$$

where

$$\hat{\mathbf{A}}_{m-1} = \sum_{k=1}^{B}\sum_{i=1}^{n} h_{ik}^{-1} \tilde{\mathbf{z}}'_{ik} \tilde{\mathbf{z}}_{ik}$$

$$\mathbf{G}_{m-1} = n_c(n_c - m - B + 1)^{-1} \sum_{k=1}^{B} \sum_{i=1}^{n} \hat{\mathbf{d}}_{m-1,ik} \hat{\mathbf{d}}'_{m-1,ik}$$

$$\hat{\mathbf{d}}_{m-1,ik} = h_{ik}^{-1} \tilde{\mathbf{z}}'_{ik} \hat{e}_{ik}$$

$\hat{e}_{ik} = u_{ik} - \tilde{\mathbf{z}}_{ik}\hat{\boldsymbol{\beta}}_{m-1}$ and n_c is the number of household-by-block categories with $h_{ik} > 0$.

Let $Y = Y_T =$ true population total of Y_{ik}, $\mathrm{X} = \mathrm{X}_T =$ true population total of X_{ik}, and $(Y_C, \mathrm{X}_C) =$ total of $(Y_{ik}, \mathrm{X}_{ik})$ in the Census. The model (4.2.8) is

$$y_{Tik} = \mathbf{x}_{Tik}\boldsymbol{\beta} + e_{Tik} \tag{A.4}$$

Under the model, the population total is

$$Y_T = \mathrm{X}_T\boldsymbol{\beta} + e_T \tag{A.5}$$

where e_T is the population total of the e_{ik}, and the Census total is

$$Y_C = \mathrm{X}_C\boldsymbol{\beta} + e_C \tag{A.6}$$

where e_C is the Census total of the e_{ik}. Then the error in the regression estimator (4.2.2) can be written

$$\begin{aligned}\hat{Y} - Y_T &= Y_C + (\hat{\mathrm{X}} - \mathrm{X}_C)\hat{\boldsymbol{\beta}} - (\mathrm{X}_T\boldsymbol{\beta} + e_T)\\ &= (\hat{\mathrm{X}} - \mathrm{X}_T)\boldsymbol{\beta} + e_C - e_T + (\mathrm{X}_T - \mathrm{X}_C)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + (\hat{\mathrm{X}} - \mathrm{X}_T)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\end{aligned} \tag{A.7}$$

To calculate the variance of an estimator of the area total in Table 1, we adopt two approximations. First, we approximate the 984 block coefficients with an average coefficient. Second, we use an average of the b-vectors of (A.1) to approximate the covariance matrix of the regression coefficients.

An average coefficient for the 984 block coefficients is obtained by regressing $y_{ik} - \mathbf{z}_{ik}\hat{\boldsymbol{\beta}}_{m-1}$ on x_{Tik} with weights h_{ik}^{-1}. This coefficient is denoted by $\hat{\gamma}$, where

$$\hat{\gamma} = \left(\sum_{k=1}^{B}\sum_{i=1}^{n} h_{ik}^{-1} x_{Tik}^2\right)^{-1} \sum_{k=1}^{B}\sum_{i=1}^{n} h_{ik}^{-1}(y_{ik} - \mathbf{z}_{ik}\hat{\boldsymbol{\beta}}_{m-1})x_{Tik} \tag{A.8}$$

With the coefficient $\hat{\gamma}$ substituted for all of the block coefficients,

$$\hat{Y} - Y = (\hat{\mathrm{X}}_m - \mathrm{X}_{Tm})\boldsymbol{\beta}_m + e_C - e_T + (\mathrm{X}_{Tm} - \mathrm{X}_C)(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_m) + \text{Remainder} \tag{A.9}$$

where $\hat{\boldsymbol{\beta}}_m = (\hat{\boldsymbol{\beta}}'_{m-1}, \hat{\gamma})'$, $\boldsymbol{\beta}_m$ is the corresponding parameter, $\mathrm{X}_{Tm} = (\mathbf{Z}_T, X_T)$, $\mathbf{Z}_T$ is the vector of population totals for the first $m-1$ of the m person categories, X_T is the total number of persons, $\hat{\mathrm{X}}_m$ is the vector of ICM estimates corresponding to X_{Tm}, and X_{Cm} is the corresponding vector of census totals.

The difference $e_C - e_T$ in (A.9) is

$$e_C - e_T = \sum\sum_{ik \in C} e_{Tik} - \sum\sum_{ik \in T} e_{Tik}$$

where the sums are over the Census (C) and over the entire population (T). Assume the population contains N households. Assume that the Census contains c households of which

d_c are duplicates. Then

$$e_C - e_T = \sum \sum_{ik \in C\text{dup}} e_{Tik} - \sum \sum_{ik \notin C} e_{Tik}$$

where C dup is the set of duplicates in the census. Observe that there are $N - (c - d_c)$ elements in the population that are not in the census. Assume the e's have common variance σ^2. Then we have

$$\begin{aligned} V\{e_C\} &= (c - 2d_c + 4d_c)\sigma^2 = (c + 2d_c)\sigma^2 \\ V\{e_T\} &= N\sigma^2 \\ V\{e_C - e_T\} &= [N - c + 2d_c]\sigma^2 \end{aligned} \tag{A.10}$$

where the covariance between e_C and e_T is $c\sigma^2$. An estimator of the variance of $e_C - e_T$ is

$$\hat{V}\{e_C - e_T\} = \hat{g}(c - d_p)^{-1} \sum_{k=1}^{B} \sum_{i=1}^{n} \hat{e}_{ik}^2 h_{ik}^{-1} \tag{A.11}$$

where $\hat{e}_{ik} = u_{ik} - \tilde{\mathbf{z}}_{ik}\hat{\boldsymbol{\beta}}_{m-1}, \hat{g} = (N - c + 2\hat{d}_c), \hat{d}_c$ is an estimator of the number of duplicates and d_p is the number of parameters estimated. The $\hat{g}$-multiplier is based on the assumption of a common σ^2, but the form of (A.10) permits the error variance to differ for different types of households.

The error in $\hat{\boldsymbol{\beta}}$ in (A.2) contains a term of the form

$$\sum_{k=1}^{B} \sum_{i=1}^{n} h_{ik}^{-1} \tilde{\mathbf{z}}_{ik}' \tilde{e}_{ik} \tag{A.12}$$

It follows that an estimator of the covariance between the term in (A.12) and $e_C - e_T$ is

$$\hat{C}\left\{e_C - e_T, \sum_{k=1}^{B} \sum_{i=1}^{n} h_{ik}^{-1} \mathbf{z}_{ik}' \tilde{e}_{ik}\right\} = (c - d_p)^{-1} 2d_c \sum_{k=1}^{B} \sum_{i=1}^{n} \tilde{\mathbf{z}}_{ik}' \tilde{e}_{ik}^2 h_{ik}^{-1} \tag{A.13}$$

If the number of duplicates, d_c, is relatively small, the covariance between $\hat{\boldsymbol{\beta}}_m$ and $e_C - e_T$ is small relative to the variances. If we assume $C\{e_C - e_T, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\}$ is small and that $\hat{\mathbf{X}}_m - \mathbf{X}_{Tm}$ is uncorrelated with $(e_C - e_T, \hat{\boldsymbol{\beta}}_m' - \boldsymbol{\beta}_m')$, the variance of $\hat{Y} - Y_T$ can be estimated with the sum of three estimated variances.

The estimated variance of an estimate of Table 1 is

$$\begin{aligned} \hat{V}\{\hat{Y}\} = {} & (\hat{\mathbf{X}}_m - \mathbf{X}_{Cm})\hat{\mathbf{V}}_{\beta\beta}(\hat{\mathbf{X}}_m - \mathbf{X}_{Cm})' + (\hat{\boldsymbol{\beta}}_{m-1}', \hat{\gamma})\hat{\mathbf{V}}\{\hat{\mathbf{X}}_m\}(\hat{\boldsymbol{\beta}}_{m-1}', \hat{\gamma})' \\ & + \hat{g}(n_c - m - B + 1)^{-1} \sum_{k=1}^{B} \sum_{i=1}^{n} \hat{e}_{ik}^2 h_{ik}^{-1} \end{aligned} \tag{A.14}$$

where

$$\hat{\mathbf{V}}_{\beta\beta} = \mathbf{T}\hat{\mathbf{V}}_{\gamma\gamma}\mathbf{T}'$$

$$\hat{\mathbf{V}}_{\gamma\gamma} = \text{Block diag}[\hat{\mathbf{V}}\{\hat{\boldsymbol{\beta}}_{m-1}\}, \hat{V}\{\hat{\gamma}\}]$$

$$\hat{V}\{\hat{\gamma}\} = \left(\sum_{k=1}^{B} \sum_{i=1}^{n} h_{ik}^{-1} x_{Tik}^2\right)^{-1} G_{\gamma\gamma} \left(\sum_{k=1}^{B} \sum_{i=1}^{n} h_{ik}^{-1} x_{Tik}^2\right)^{-1}$$

$$G_{\gamma\gamma} = n_c(n_c - m - B + 1)^{-1} \sum_{k=1}^{B} \sum_{i=1}^{n} h_{ik}^{-2} x_{Tik}^2 a_{ik}^2$$

$$a_{ik} = y_{ik} - \mathbf{z}_{ik}\hat{\boldsymbol{\beta}}_{m-1} - \hat{\gamma} x_{Tik}$$

$$\mathbf{T} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{b} & 1 \end{pmatrix}$$

$$\mathbf{b} = \left[\sum_{k=1}^{B} \sum_{i=1}^{n} h_{ik}^{-1} x_{Tik}(y_{ik}, \mathbf{z}_{ik}) \right] \left[\sum_{k=1}^{B} \sum_{i=1}^{n} h_{ik}^{-1} x_{Tik}^2 \right]^{-1}$$

$\mathbf{b}$ is the average of the b-vectors of (A.1), n_c is the number of household-by-block categories with $h_{ik} > 0$, $\hat{\mathbf{V}}\{\hat{\mathbf{X}}_m\}$ is the estimated covariance matrix of $\hat{\mathbf{X}}_m = (\hat{\mathbf{X}}_{m-1}, \hat{X}_T)$, $\hat{X}_T$ is the estimated total number of persons for the large geographic area, and $\hat{g}$ is an estimated finite population adjustment factor.

In Paterson, 42,516 households were observed in the Census and it is estimated that there are 47,799 households in the population. On the basis of the dual system sample, it is estimated that three percent of the Census households are duplicates. Then $\hat{g}$ is

$$\hat{g} = 47,799 - 42,516 + 2,551 = 7,834$$

The estimated variance of the ICM estimates was computed using the replication procedure of Fay (1990).

7. References

Bankier, M., Houle, A., and Luc, M. (1997). Calibration Estimation in the 1991 and 1996 Canadian Censuses. Statistics Canada, Draft memo.

Bethlehem, J.G. and Keller, W.A. (1987). Linear Weighting of Sample Survey Data. Journal of Official Statistics, 3, 141–153.

Cox, L.H. and Ernst, L.R. (1982). Controlled Rounding. INFOR 20, 423–432.

Deville, J.C. and Särndal, C.E. (1992). Calibration Estimators in Survey Sampling. Journal of the American Statistical Association, 87, 376–382.

Deville, J.C., Särndal, C.E., and Sautory, O. (1993). Generalized Raking Procedures in Survey Sampling. Journal of the American Statistical Association, 88, 1013–1020.

Fay, R.E. (1990). VPLX: Variance Estimates for Complex Samples. Proceedings of the Survey Research Section, American Statistical Association, 266–271.

Hogan, H. (1992). The 1990 Post-Enumeration Survey: An Overview. The American Statistician, 46, 261–269.

Hogan, H. (1993). The 1990 Post-Enumeration Survey: Operations and Results. Journal of the American Statistical Association, 88, 1047–1060.

Huang, E.T. and Fuller, W.A. (1978). Nonnegative Regression Estimation for Survey Data. Proceedings of the Social Statistics Section, American Statistical Association, 300–303.

Husain, M. (1969). Construction of Regression Weights for Estimation in Sample Surveys. M.S. Thesis, Iowa State University, Ames, Iowa. Unpublished.

Isaki, C.T. and Ikeda, M.M. (1996). Some Estimators of Housing Unit Category Totals in the 95 ICM Via Dual System Estimation and Census Plus. Unpublished manuscript, November 13.

National Research Council (1999). Measuring a Changing Nation: Modern Methods for the 2000 Census, eds. M.L. Cohen, A.A. White, and K.F. Rust. Panel on Alternative Census Methodologies, Committee on National Statistics, Commission on Behavioral and Social Sciences and Education, Washington, D.C.: National Academy Press.

Zaslavsky, A.M. (1988). Representing Local Area Adjustments by Reweighting of Households. Survey Methodology 14, 2, 265–288.

Received April 1998
Revised January 2000