Survey Methodology, December 1988
Vol. 14, No. 2, pp. 265–288
Statistics Canada

Representing Local Area Adjustments by Reweighting of Households

ALAN M. ZASLAVSKY[1]

ABSTRACT

Suppose that undercount rates in a census have been estimated and that block-level estimates of the undercount have been computed. It may then be desirable to create a new roster of households incorporating the estimated omissions. It is proposed here that such a roster be created by weighting the enumerated households. The household weights are constrained by linear equations representing the desired total counts of persons in each estimation class and the desired total count of households. Weights are then calculated that satisfy the constraints while making the fitted table as close as possible to the raw data. The procedure may be regarded as an extension of the standard "raking" methodology to situations where the constraints do not refer to the margins of a contingency table. Continuous as well as discrete covariates may be used in the adjustment, and it is possible to check directly whether the constraints can be satisfied. Methods are proposed for the use of weighted data for various Census purposes, and for adjustment of covariate information on characteristics of omitted households, such as income, that are not directly considered in undercount estimation.

KEY WORDS: Undercount; Raking; Local-area adjustment; Missing data.

1. HOUSEHOLD-LEVEL ADJUSTMENT BY WEIGHTING

A major research effort has been devoted to methods for estimation of the undercount in the 1990 Census in the United States (National Academy of Sciences 1985). In one of the primary methodologies that has been proposed, a Post Enumeration Survey (PES) would be conducted shortly after the Census in a sample of blocks. The fraction of persons in the PES who were omitted from the Census enumeration yields an estimate of Census undercoverage. Estimates of the undercount would be carried down to some geographical level (possibly the smallest geographical unit used by the Census, the block). These estimates would apply to classes formed on the basis of characteristics of persons, as well as possibly some household or block-level characteristics. The term "class" will be used henceforth to refer to estimation or adjustment classes or cells; the term "block" will refer to the smallest geographical unit for which undercount estimates are calculated. The 1980 Census found approximately one hundred million households in two to four million blocks, depending on the definitions used.

For each block, the outcome of the processes described above would be a vector of estimated undercounts, with S components corresponding to the adjustment, or estimated number of persons omitted from the census in that block, from each of S adjustment classes. The methods by which these estimates are arrived upon are beyond the scope of this paper. However, in our examples we shall assume that for each class within each block there is an undercount rate, expressing estimated omissions as a fraction of enumerated persons in that class and block. In this paper, the term "adjustment" refers to any process which incorporates the estimated undercount into the enumeration. The adjustment classes might be, but would

[1] Alan M. Zaslavsky, Statistics Center, Massachusetts Institute of Technology, Room E40-111, Cambridge, MA 02139, U.S.A. and Harvard University Department of Statistics, Cambridge, MA 02138, U.S.A.

not necessarily be, the same as the post-strata formed in analysis of a Post-Enumeration Program. For forming simple marginal tabulations of persons by characteristics, this information might well be adequate. In particular, small-area counts used for various official and commercial purposes could be calculated from block totals.

However, for some purposes it would be desirable to place the added persons in households. We assume for these purposes that there is also an estimate of the number of omissions of whole households in each block. There might also be information distinguishing omissions of persons within enumerated households from those in omitted households.

If the resulting adjusted records are to be meaningful, the composition of the added households and the relationships of its individual members must be logically consistent and typical of the types of households found in that area. The term "composition" will be used to refer to the number of household members from each adjustment class. Thus, for example, a household consisting of a 20-year old white female head of household, a 75-year-old Chinese male, and a 10-year-old black daughter would not be a very plausible household, even if all of its members were from classes that are well represented in the block. Yet abstractly to describe these patterns and create new households that fit them is a daunting task.

Example 1: *Forming a roster of households.*

Table 1 illustrates part of a census enumeration as it might appear on a microdata tape.

Table 2 represents the same roster, showing how the composition of the households might be summarized if there were only three estimation classes: (1) men over 20 years of age, (2) women over 20 years of age, and (3) children up to 20 years of age.

Table 1

A piece of a sample microdata file

Name	Address	Sex	Age
John Smith	328 Main Street	M	34
Mary Smith	328 Main Street	F	32
Louise Smith	328 Main Street	F	7
Nancy Chen	330 Main Street	F	62
Jorge Ramirez	332 Main Street	M	21
Juan Ramirez	332 Main Street	M	24

Table 2

Microdata file recoded by household, showing composition of households

Address	Count of persons by class		
	Class 1	Class 2	Class 3
328 Main Street	1	1	1
330 Main Street	0	1	0
332 Main Street	2	0	0

Essentially the same problem arises in many situations in which a household survey must be reweighted to match known marginal totals for various classes of individuals.

The essence of the method proposed in this paper is to assign weights to the households enumerated in the census lists for the block, so that the weighted totals of persons in each adjustment class and the weighted total number of households are precisely equal to the corresponding adjusted totals. Thus, although the weighting changes the proportionate composition of the block, all of the households are real and possess characteristics and relationships that are logically consistent and reasonable for that block. This weighting methodology is similar to the standard raking adjustment, in which the weight applied to counts in a cell of a contingency table is the adjusted count divided by the original count. The household weights are calculated *after* the block totals have been adjusted and will be consistent with those totals. For most Census purposes, the weighted records would be an adequate basis for forming published tables and sampled lists.

This proposal might be contrasted with imputation methods, in which undercounted units are represented by whole units added to the roster. The imputed units may be either persons or households. Although individual persons may be imputed into the block, the problem of fitting these persons into plausible households remains unsolved. Placing them in fictitious "group quarters," as was done in some tests of adjustment procedures, sidesteps this problem at the cost of creating a skewed picture of relationships in the block. Reweighting or imputation of individuals would be appropriate for residents of institutions or group homes, for whom the particular configuration of persons in the dwelling unit has no particular significance.

Another approach to imputation starts with probability models for omissions of households and of persons within households, and draws imputed households from the posterior distribution of the omissions given the enumerated households. This methodology is suited to the multiple imputation approach (Rubin 1987), in which the entire imputation process is repeated several times to represent the variability introduced by the underenumeration. However, in each block roster that is created, totals based on enumerated and imputed households would not necessarily be precisely equal to the desired adjusted totals. In this paper, our concern is with methods that give an *exact* fit to population estimates derived at a preceding stage.

The remaining sections of this paper develop methods for the proposed weighting adjustment. Section 2 gives a mathematical formulation of the objectives of the weighting scheme, while Section 3 explains how to fit the weights. Section 4 explains how to incorporate the distinction between omissions in enumerated and omitted households into the scheme. Section 5 introduces some refinements that improve the robustness of the procedure against the variability of small blocks. Section 6 describes simulation results. Section 7 discusses the use of weighted data for various Census purposes, while Section 8 considers the effects of the weighting adjustment on covariates that are not part of the scheme used in forming the adjustment classes. Finally, Section 9 summarizes some unresolved questions and areas for future research.

2. OBJECTIVES AND MATHEMATICAL FORMULATION OF A WEIGHTING PLAN

It is an essential goal of the proposed plan that the population of the block be assigned to valid household units, so that statistics for which the unit is the household are unambiguously defined. Thus, weights are assigned to *households*; the same weights apply to every *person* within the household.

In order that the counts in the weighted roster be those which are given by the predetermined adjustment, the following constraints must be satisfied:

(A1) Within each block, the sum of household weights equals the adjusted number of households.

(A2) Within each adjustment class and each block, the sum of weights for persons equals the adjusted number of persons.

In order that the weighted block roster be as similar as possible to the original block roster, we further require that:

(B) The weights should be, in some sense, as close to each other as possible.

With unit (or equal) weights, the composition of the block remains unchanged. If the weights are not very unequal, the census composition of the block is nearly preserved by the weighting scheme. To the extent that information about the undercount does not require a drastic revision of our view of the makeup of the block such a drastic revision should be avoided, consistently with good survey practise regarding weights.

We now turn to the mathematical formulation of these criteria. Suppose that in the block under consideration, there are S adjustment classes and I enumerated households, and household i contains C_{is} members from class s. Suppose that H is the desired total number of households in the adjusted roster for the block and D_s is the desired total number of persons in class s. Let $\mathrm{W_i}$, $i = 1,2, \ldots I$, be the weights corresponding to the households. (A1) requires that

$$\sum_{i=1}^{I} W_i = H \tag{1}$$

and (A2) requires that

$$\sum_{i=1}^{I} W_i C_{is} = D_s, \; s = 1,2 \ldots S. \tag{2}$$

These constraints can be represented by a matrix equation of the form $AW = B$, where

$$A = \begin{bmatrix} \mathbf{1} \\ C' \end{bmatrix}, B = \begin{bmatrix} H \\ D \end{bmatrix}, \; W' = [W_1 \; W_2 \ldots \; W_I] \text{ and } D' = [D_1 \; D_2 \ldots D_s] \tag{3}$$

and $\mathbf{1}$ is a row of $1's$.

Objective (B) is represented by selecting some objective function that represents the distance between the weights W and uniform weighting, and minimizing it. We will use the objective function $T = \sum W_i \log (W_i)$. This measure is proportional to the discriminant information (Kullback-Liebler information) of the discrete probability distribution (over households) with relative weights W_i with respect to the probability distribution with equal weights, and is the same objective function that underlies the traditional "raking" (iterative proportional fitting) procedure for adjusting contingency tables (Deming and Stephan 1940; Ireland and Kullback 1968; Oh and Scheuren 1978 have a larger bibliography). Thus, our procedure may be regarded as an extension of raking. Scheuren (1973) applies raking to reweighting of households; Cilke and Wyscarver (1988) reweight to linear constraints but use a different objective function than those considered here. Methods similar to those presented here were developed independently by Alexander (1987).

In the context of raking, initial counts X are given for cells in a contingency table, and new cell counts Y are calculated to minimize the objective function $\sum Y_i \log (Y_i/X_i)$. Then the weights of the original observations are the ratios $W_i = Y_i/X_i$. In our context, if X_i households happened to have exactly the same composition we could regard them, in the same way, as forming a single entry in the roster with initial count X_i and fit an adjusted count Y_i. However, with a large number of adjustment classes, it would be unusual for several households in the same block to have exactly the same composition. Thus we will not attempt to group households; rather, it is notationally and computationally simpler to list the households separately so that for each enumerated household composition the initial count $X_l = 1$ and $Y_l = W_l$. Aside from this notational difference, the mathematical formulation here differs from that of a raking adjustment only in that the linear constraints do not have the special structure of margins in a contingency table. For brevity in the presentation of examples, we will sometimes include a count on a line to represent that number of identical lines in the roster of households.

In the contingency table setting, raking preserves cross-product ratios of cells, and preserves independence of variables when it holds in the original table. For these reasons, it has been called "structure-preserving estimation" in small-area estimation applications (Purcell 1979; Purcell and Kish 1979). See Section 10.1 for a further discussion of objective functions.

Our procedure differs from raking in that the linear constraints do not necessarily refer to margins in a contingency table. Our methodology includes raking as a special case, as well as the raking generalization of Oh and Scheuren (1978) in which different tables are used to fit each margin. In fact, constraints may be imposed on continuous as well as discrete covariates; applications of this sort are proposed in Section 8.3. Furthermore, the algorithms that are set forth allow direct determination of whether there are in fact any weights that are consistent with all of the given constraints. It is possible then to select constraints that must be relaxed in order to fit weights. These features give these methods potential applicability extending beyond the area of representing undercount.

3. FITTING THE WEIGHTS

The problem before us now is to determine weights satisfying the constraints $AW = B$, $W \geq 0$, minimizing the objective function $T = \sum W_i \log (W_i)$. To make T a continuous function of W, we adopt the usual convention $0 \log 0 = 0$.

We will call any weight vector that satisfies the linear constraints (the equations and the inequalities) a *feasible solution*. As long as there is a constraint on the total weight of the households, the set of feasible solutions is bounded and therefore T assumes a minimum value on it; furthermore, since T is strictly convex, the solution is unique.

The problem of calculating weights then naturally is divided into three tasks: (1) determining whether the linear constraints $AW = B$ are consistent; (2) determining whether there are any feasible solutions; and (3) finding the feasible solution minimizing T. We will suppose that there are I households and p constraints, so A is a $p \times I$ matrix.

Example 2: *Fitting weights.*

Table 3 illustrates the roster of households in a block in which three classes are represented, as in Example 1; we may think of the classes as "men," "women," and

Table 3

A household roster

Line #	Count per household by class			Number of households
	Class 1 (men)	Class 2 (women)	Class 3 (children)	
1	0	1	0	50
2	0	1	1	40
3	1	0	0	40
4	1	0	2	15
5	1	1	0	50
6	1	1	1	60
7	1	1	2	40

Table 4

Adjusted totals

	Raw count	Adjustment rate	Adjusted count
Class 1	205	.05	215
Class 2	240	.03	247
Class 3	210	.04	218
Households	295	.02	301

and "children." This table may be regarded as a condensed version of a table with 295 lines, each representing one household.

The unadjusted and adjusted counts of households and of persons in each class are found in Table 4. The adjusted counts are calculated by applying the listed adjustment rates and rounding. The method by which the adjusted counts are obtained is immaterial, however, to the rest of the process.

3.1 Consistency of Linear Constraints

As long as the rows of A are independent, the constraints $AW = B$ will be consistent. If any row is dependent on the others, the corresponding constraint is either inconsistent or redundant, depending on the values in B. Dependent rows can be identified by applying the Q-R decomposition to A'. If the corresponding constraints are redundant, they may be deleted without any loss of information; if they are inconsistent, the constraints must be reformulated in some way.

Example 2: (*continued*).

The A matrix for this example has independent rows, and hence the constraints are consistent.

In Section 5, we consider circumstances in which inconsistent constraints are likely to appear and some methods for dealing with them.

3.2 Existence of Feasible Solutions

Determining the existence of feasible solutions is equivalent to determining an initial feasible solution in a linear programming problem, and the standard algorithms can be used. Suppose our problem is to find a positive solution W to $AW = B$, where $B \geq 0$. (If the latter condition does not hold it can be made true by reversing the sign of negative elements of B and the corresponding rows in A.) Then create an augmented problem $[A \mid I]\ [W' \mid Z']' = B$, $W, Z \geq 0$, where I is a $p \times p$ identity matrix and Z is a p element vector variable. This problem automatically has an initial solution $W = 0, Z = B$. Then apply the simplex method (as in Gass (1964) or any other linear programming text) to minimize $\sum Z_i$. If that sum can be reduced to 0, the corresponding W values are a solution to the original problem, while if it cannot, the original problem has no solution.

Example 2: *(continued)*.

A feasible (but not optimal) solution for this example gives total weighted counts of 86, 54, 29, and 132 to the household compositions in lines 2, 3, 5, and 6 respectively of Table 3. It may be verified that these counts yield the desired adjusted totals for households and for individuals in each class.

The problem of infeasibility is similar to that of inconsistency and is also discussed in Section 5.

3.3 Optimizing the Objective Function.

By the method of Lagrange multipliers, the minimizing solution must satisfy the equations $\partial T/\partial W_i = \log W_i + 1 = a_i'\lambda$, where a_i is the i-th column of A and $\lambda' = (\lambda_1, \lambda_2, \ldots \lambda_p)$. Then $W_i = \exp(a_i'\lambda - 1)$; thus the model for the weights is log-linear in form, like that for a conventional raking adjustment. λ_s represents the additional log-weight increment associated with a unit increment in the corresponding constraint coefficient a_{is}, *i.e.* adding an additional household member from adjustment class s to the household.

We can solve for λ by Newton's method to satisfy $AW = B$. The iterative scheme we use is

$$\lambda^{(t+1)} = \lambda^{(t)} - (AW^*A')^{-1}(AW - B), \tag{4}$$

where W^* is the matrix with the elements of $W = W(\lambda^{(t)})$ on the diagonal. A good starting value for λ is $\lambda^{(0)} = (AA')^{-1}B$, which can be derived from a linear approximation around equal starting weights. A cyclic descent procedure for solving these equations, which is a generalization of iterative proportional fitting, is described in Section 10.2.

Example 2: *(continued)*.

The weights per household and total weighted counts (weight times raw count) for each line in Table 3 are shown in Table 5. No household is upweighted by more than 8% or downweighted by more than 5%.

Table 5

Optimal weights for Example 2

Line #	Weight	Weighted counts
1	0.9554	47.77
2	0.9557	38.23
3	0.9816	39.27
4	0.9823	14.73
5	1.0730	53.65
6	1.0734	64.40
7	1.0737	42.95

4. WHOLE-AND WITHIN-HOUSEHOLD ADJUSTMENTS

We now consider the distinction between within-household adjustments (that is, adjustments for omissions of persons within enumerated households) and whole-household adjustments (that is, adjustments for omissions of whole households). This distinction has previously been made for purposes of analysing the causes of undercount (Fay 1986). Our concern here is to use it to more accurately represent the undercount by an adjustment.

Within-household adjustments do not involve adding any households to the roster, but only shifting weight between households to increase the weighted totals of persons in the various classes. That is, households with few or no persons in a particular class are downweighted and those with many are upweighted, so that the total household weight remains constant. Thus, in this portion of the adjustment, some households will inevitably have their weights reduced. Whole-household adjustments, on the other hand, correspond to households that were omitted entirely from the census. These adjustments do not reflect on the accuracy of the enumerated households; thus they should be represented by adding households to the roster without taking weight away from the households that were enumerated.

We propose to separate these two portions of the adjustment. One set of constraints represents the within-household adjustment. The total household weights are here constrained to equal the enumerated count of households, while the total weights assigned to persons in each class are constrained to equal the enumerated count in that class plus the within-household adjustment for that class. $A W_1 = B_1$ where B_1 consists of the *enumerated* household count and the counts of persons by class adjusted for *within-household* undercount.

A second set of constraints represents the whole-household adjustment. The total household weights are here constrained to equal the estimated omitted households, and the total person weights in each class are constrained to equal the estimated omitted persons in those households. $A W_2 = B_2$ where B_2 consists of the count of added households and the counts of added persons by class for the adjustment for *whole-household* undercount.

After fitting two sets of weights corresponding to the two sets of constraints, the two weights for each household are added to obtain weights that incorporate both parts of the adjustment ($W = W_1 + W_2$). The distinction between whole- and within-household adjustments contains information which may lead to a different set of adjusted weights than would be calculated if the adjustments were combined, as is illustrated in Example 3. However, if this distinction is not made in the estimation of the undercount, an adjustment can still be calculated in a single step.

Example 3: *adjustments for whole-household omissions.*

Suppose there are only two adjustment classes, and a hypothetical block has the composition described in the first three columns of Table 6.

Suppose now that to the 30,010 households enumerated, we must add 231 persons each in Class 1 and Class 2, and 121 households. The last three columns of Table 6 show the adjusted counts under alternative assumptions: (1) the omitted persons may belong to any household, enumerated or omitted, and (2) all of the omitted persons were in the omitted households.

When the omitted persons could have been in any household, the algorithm downweights the households with only one person from each class (1,1) and upweights households with two from one class and one from the other (1,2 and 2,1). While the households with two persons from each class are substantially upweighted (by a factor of 1.354), only a small portion of the added persons appear in those households since

Table 6

Hypothetical raw and adjusted household counts for Example 3

Household composition		Raw count (number of households)	(1) Omitted persons in any household	(2) Omitted persons in omitted households only	
Class 1 persons	Class 2 persons		Adjusted counts	Counts of omitted households	Adjusted totals, omitted and enumerated households
1	1	10,000	9904.54	.01	10,000.01
1	2	10,000	10106.46	10.99	10,010.99
2	1	10,000	10106.46	10.99	10,010.99
2	2	10	13.54	99.01	109.01

the original count for that composition is so small.

When the omitted persons appear only in the omitted households, weights are calculated first to fit 231 × 2 = 462 persons into 121 additional households, and then these weights are added to the unit weights in the raw counts. While no household composition is downweighted, the (2,2) households are upweighted extremely (by a factor of 10.901). In fact, it is mathematically impossible to accommodate 462 persons in 121 households of two to four persons each without having at least 99 households with 4 members. Thus, the information that the added persons (or some known fraction of them) belong in the omitted households substantially changes our view of the appropriate adjustment.

5. FEASIBILITY OF CONSTRAINTS

In the preceding sections we have assumed that feasible solutions exist to the constrained optimization problem. Here we will consider situations in which the solutions will not exist or will be unsatisfactory, and some alternative methods to deal with these situations.

5.1 When Will Constraints be Non-feasible?

There are three ways in which the constraints may fail to allow of satisfactory solutions: (1) when the constraints are actually inconsistent, (2) when the constraints are consistent but there are no positive weights that satisfy them, and (3) when there is a feasible solution but it involves an extreme adjustment to some household weights. The issues associated with these three failure modes are fairly similar.

One could write down constraints that are intrinsically inconsistent, for example that all classes of men are adjusted upward by 2% while men in total are adjusted upward by 4%. In our procedure each constraint applies to the number of persons in a distinct adjustment class and so there are no inconsistencies of this sort. However, a contingent inconsistency is still possible, that is to say one that depends on the particular collection of household compositions that appears in a block. The following are examples of contingent inconsistency, infeasibility, or unsatisfactory weights:

(1) Proposed undercount estimation methods envision defining over 100 adjustment classes. In a small but diverse block the number of classes represented might be larger than the

number of households; hence the number of constraints would be larger than the number of weights to be fitted. An inconsistency is then almost inevitable.

(2) If all households in the block have exactly the same number of members from a particular adjustment class (*e.g.* every household has one young Hispanic girl), then the number of members of this class represented is unaffected by the distribution of weights.

(3) The adjustment of the number of households may be too large or small to accommodate the adjustment of persons in some class. This may represent a failure of the model for adjustment of the number of households. For example, suppose that the number of men to be added by the whole-household adjustment is greater than the number of households to be added, but no household in the block has more than one man. The constraints then might be consistent but infeasible, since they could be satisfied only by assigning negative weights to some households without men.

(4) The block may have had omission rates atypical of blocks in the PES on which omission rates were estimated. For example, suppose that in most blocks (including most of the PES sample blocks), adult males with certain characteristics tend to be heavily undercounted, but the block being adjusted is atypical in having adult males of this class present in most households and well counted. The class undercount estimate might lead to an extreme upward adjustment that could not be accommodated within the existing households.

(5) Some adjustment may require giving substantial additional weight to households containing persons from a combination of adjustment classes that appears in only one household, so that household receives an extreme weight. In this case the problem is feasible but the solution is not very satisfactory.

Problems of infeasibility may also arise where the difficulty cannot be so easily traced to a particular inconsistency in the adjustment.

5.2 Making the Constraints Feasible

Regardless of the stage of the fitting procedure at which the infeasibility is discovered, several methods are available to relax the constraints and make them feasible. In this section, we survey several such methods, drawing out both the intuitive logic of each choice and the computational methods required.

5.2.1 Methods Based on Dropping Rows (constraints) of A

When checking for consistency of constraints, some rows may be found to be linearly dependent on the previous rows and hence either redundant or inconsistent. If these rows are simply dropped from the A matrix, a consistent set of constraints is obtained; thus, no further computational effort is required.

If the constraints are arranged in sequence from the most important to the least important, than the less important constraints will be dropped when they are inconsistent with the more important ones. This ordering makes the most sense if the original constraints on distinct adjustment classes (defined by a multi-way classification of the population) are reframed in an ANOVA-like manner as constraints on total population ("grand mean"), classes defined by one classification variable ("main effects"), and classes defined by interactions. For example, if there are ten adjustment classes defined by two sexes and five age ranges, the reframed constraints in order of importance might be: total population (1 constraint), population by sex (1 more constraint), population by age (4 more constraints), age-sex interactions (the remaining 4 constraints). The 4 age constraints could be further broken down as old-vs.-young (1 constraint) and 3 further constraints within those larger groups.

A similar procedure can be applied at the stage of checking feasibility of the constraints. If it is not possible to make all of the $Z_i = 0$, the objective function in the linear programming problem can be modified to be $\sum c_i Z_i$, with the coefficients $c_i > 0$ corresponding to the most important constraints made larger. Then a maximal set of feasible constraints can be identified, and the remaining constraints dropped.

The outcome of this procedure would be weights that give the correct block totals on the coarser classifications of persons, while failing to be correct on all cross-tabulations.

5.2.2 Methods Based on Adding Columns (households) to A

When constraints are only contingently infeasible (in the previous sense that infeasibility depends on the particular set of household compositions in the block), they become feasible when households are added that have the required composition. The simplest application of this principle is to work at a higher level of geographical aggregation than a block. A few adjacent blocks may be combined when problems arise in fitting, or the entire roster may be grouped at, for example, the enumeration district level before weighting. The larger the unit, the broader the range of household compositions that will be represented and the less likely that problems of infeasibility will arise.

A more sophisticated procedure would use a hot-deck of households from adjacent "donor" blocks to enrich the pool of households to which weight can be assigned. Computational simplicity is important here since it may be necessary to scan through a long list of households to find the one or ones which will make the constraints feasible. In the consistency-checking stage, if row j of A is dependent on the previous rows, then if the column for the added household is independent of the columns of A (with regard only to the first j rows), row j of the augmented A will be independent. In the stage of checking for feasibility, if the algorithm halts because no reduction can be made in the objective function $\sum Z_i$, the search for basic columns can be extended to columns corresponding to households in the hot deck. Finally, if some household's fitted weight is extremely high, the hot deck can be scanned for other households that would also receive high weights with the current values of λ (that is, columns a such that $a'\lambda$ is large). If these are added to the block they will draw off some of the weight from the overweighted households when the weights are refitted, since they are likely to also have members in the same adjustment classes.

The intuition behind this method is that the household compositions that are enumerated in a block are only a sample of those which actually could have appeared there had the enumeration been complete. The observed distribution of household compositions is smoothed by mixing it with the distribution for adjacent blocks, which contain households that are also typical for that area. Thus, conceptually this method is related to Bayesian smoothing methods that improve estimation of some quantity for one unit by borrowing strength from its distribution in similar units. This Bayesian rationale is developed in terms of a block-level random-effects model by Zaslavsky (1989).

The donor blocks could be chosen by a sequential hot deck procedure; then, the donor blocks would tend to be geographically close to the adjustment block and no particular set of blocks would have undue influence on the entire census. By detailed stratification of blocks, the donor blocks could be selected to be similar to the block being adjusted on characteristics such as mean income, types of housing units, and racial balance.

5.2.3 Combined Methods

The two types of methods outlined above can be combined by an appropriate reframing of constraints. The principle here is to satisfy *all* constraints in the larger geographical units, while satisfying only the more important constraints in the smaller units. This type of compromise may make it possible to get a fairly good fit to the desired distribution without having to add additional records to the roster.

Suppose that the A matrices for several blocks have been reframed similarly as sequences of rows representing main and interaction constraints. Then a single large A matrix representing all of the constraints can be formed. The rows for the more important constraints can be kept separate, while rows for subsidiary constraints can be combined across blocks. For example, suppose there are ten adjustment classes, defined by sex (2 levels) and age (5 levels), and two blocks. Altogether there are eleven constraints (one for number of households and one for each adjustment class) in each block. If these are combined into a single matrix, keeping main effects and two-way interactions, the constraints are: block household counts (2 constraints), block populations (2 constraints), sex (1 constraint), age (4 constraints), block × sex interaction (1 constraint), block × age interaction (4 constraints), and sex × age interaction (4 constraints) in the combined blocks. Here 4 constraints have been eliminated (block × sex × age interaction); in a more realistic problem with more blocks, classification variables, and levels, the reduction would be much greater.

6. SIMULATION RESULTS

Simulations were performed to answer two classes of questions:

(1) The first set of questions is concerned with evaluation of the success of the algorithm in terms of its own constraints and objectives. Does the reweighting algorithm give an answer? In real problems, is there a solution to the weighting constraints? How much do the weights vary? Is the amount of computation required within reasonable limits?

To answer these questions, "feasibility simulations" were performed in which the weighting algorithm was applied to simulated blocks made up of real households, using real adjustment rates. This procedure thus closely parallels the practical application of the algorithm.

(2) The second set of questions is concerned with evaluation of the success of the algorithm in improving the quality of inferences based on a micro-data set: does the weighted micro-data set more accurately describe the real world than the raw, unweighted data?

To answer these questions, simulated blocks made up of real households were drawn, representing the true (but unobserved) compositions of households in blocks. For each "true" block, omissions were imposed using real estimated undercount rates and a plausible model for the distribution of undercount among households. The weighting algorithm was applied to the "enumerated" blocks generated in this way. Summary statistics describing household composition were calculated for the simulated "true" blocks and for the simulated observed blocks with undercount, both unweighted and weighted for undercount adjustment. The goal of these "inference simulations" was to determine whether the reweighting brought the statistics closer to their values in the "true" blocks; in other words, did reweighting correct the biases caused by the undercount?

The source of households for all simulations was the 1% "B" Public Use Microdata Sample (PUMS) from the 1980 Census (Bureau of the Census 1985). Households were extracted from sections of Los Angeles County, California that include the site of the Test of Adjustment Related Operations (TARO) of the 1986 Test Census.

Undercount rates were those calculated from the 1986 TARO (Diffendal 1988, Table 7) for adjustment classes defined by sex, age (five levels), race (Hispanic, Asian, or "other race"), and tenure (owner or renter). Adjustment factors calculated from the given undercount rates ranged from 0.982 to 1.211.

Each household was coded as a vector of counts representing the number of individuals in that household from each of the 60 adjustment classes.

Further details on the simulation procedures and on a larger set of simulations are in Zaslavsky (1989).

6.1 Feasibility Simulations

For each of four block sizes (20, 50, 100, and 200 households), 50 simulated blocks were drawn from the full sample and 50 were drawn from only those households with no Asian members. For each block, simulations were attempted using two levels of the household adjustment rate (the factor by which the number of households in the block is adjusted).

The algorithms of Section 3 were applied. To recapitulate, the linear constraints were checked first for consistency, and then for feasibility (existence of a positive solution); finally, weights were calculated using Newton's method. As no data were available distinguishing within-household and whole-household omissions, no effort was made to separate them in these or other simulations.

The results of these simulations are summarized in Table 7.

Consistency and feasibility:

The columns headed "incons", "infeas", and "OK" represent the number of simulated blocks (out of the 50 trials) in each simulation that fell into each of the following categories respectively: (1) the constraints were inconsistent (could not be satisfied by any weights), (2) the constraints were consistent but not feasible (could not be satisfied by any positive weights), or (3) the constraints were both consistent and feasible.

In the "non-Asian" simulations there are 41 constraints to be satisfied (some of which may be trivial, *i.e.* when the corresponding adjustment classes are unrepresented in the block). Thus with 20-household blocks, the constraints were never consistent; with 50-household blocks, the constraints were sometimes consistent and then usually feasible. The constraints were usually feasible in 100-household blocks, and always in 200-household blocks.

The numbered columns at the right represent the order of the simplest marginal constraint that could not be satisfied, in the sense of the heirarchical reparametrization in Section 5.2.1. Thus, column (1) indicates the number of simulated blocks for which a "main effect" constraint (marginal total of persons classified by one stratifying variable) could not be satisfied, column (2) indicates the number of trials for which a two-way interaction constraint could not be satisfied, etc. Even when the constraints were inconsistent with 50- or 100- household blocks, the main-effect constraints and often the two-way or even three-way interactions were feasible. This suggests that pooling of blocks for higher-order interactions, as described in Section 5.2.3, might be a successful strategy for dealing with problems of infeasibility.

The results were less encouraging for simulations using the full samples. Even with 200-household blocks, only rarely were the constraints consistent and feasible. With increasing block size the lower-order constraints were more likely to be feasible. This is explained by the small number of households with Asian members (approximately 5% in each sample). Out of 200 households, the expected number of Asian households would be about 10, an insufficient number to satisfy the 20 possible constraints for the Asian adjustment classes. Such a situation in which some groups of adjustment classes are poorly represented in a certain region or in particular blocks would surely not be unusual in practise. This would require pooling of blocks on a large scale for the corresponding constraints, while the constraints for the better-

Table 7

Feasibility simulation results

Non-Asian Households

size	HH	rate	incons	infeas	OK	maxW	minW	varW	iters	(1)	(2)	(3)	(4)
10		1.00	50	0	0	NA	NA	NA	NA	22	28	0	0
10		1.05	50	0	0	NA	NA	NA	NA	8	42	0	0
20		1.00	50	0	0	NA	NA	NA	NA	0	50	0	0
20		1.05	50	0	0	NA	NA	NA	NA	0	50	0	0
50		1.00	47	1	2	1.921	0.200	0.142	3.00	0	3	37	8
50		1.05	47	0	3	1.550	0.620	0.036	1.33	0	3	36	8
100		1.00	10	0	40	2.068	0.429	0.088	2.03	0	0	8	2
100		1.05	10	0	40	1.573	0.753	0.020	1.90	0	0	8	2
200		1.00	0	0	50	2.434	0.543	0.063	2.18	0	0	0	0
200		1.05	0	0	50	1.749	0.821	0.015	2.00	0	0	0	0

Full Sample

size	HH	rate	incons	infeas	OK	maxW	minW	varW	iters	(1)	(2)	(3)	(4)
100		1.00	49	0	1	--	--	--	--	0	34	15	0
200		1.00	49	0	1	--	--	--	--	0	2	43	4

represented classes might be satisfied on a smaller scale.

Weights:

The maximum and minimum household weights and the variance of the weights were calculated for each simulated block for which the constraints were consistent and feasible. For each simulation condition, the average value of these quantities (across simulated blocks) is displayed under the heads "maxW", "minW", and "varW." The following observations characterize some of the effects of the simulation design factors on the fitted weights.

(1) For simulations with household count adjustment factor of 1.05, in every case, the average variance of the weights was smaller, and the average of the minimum weights and of the maximum weights were closer to unity, than with household adjustment factor 1. This is intuitively reasonable since almost all class adjustment factors exceed 1, and it requires a more extreme adjustment to add individuals to existing households than to add individuals and households to accommodate them. For example, if the adjustment factors for households and for every adjustment class are all equal, every household would be upweighted equally.

(2) Fixing other factors, the variance of the weights becomes smaller as the number of households per block increases. Again, this is intuitively reasonable because the pool of households is richer in a larger block; the probability of finding exactly the households needed to represent undercounted individuals is higher. The trends for the extreme weights are less clear-cut than for variances; here, the narrowing of the variance is offset by the larger sample over which the extreme is calculated in the larger blocks.

(3) The average variances for simulations with 200-household blocks were at most .063. Thus the reweighting is generally not extreme.

Computational costs:

The mean number of Newton steps required to fit the weights (from the starting values given in Section 3.3), shown under the heading "iters", is usually about two. These iterations were sufficient to satisfy all constraints with error no greater than .001. Using this information, a rough estimate can be given of the number of floating point operations required to apply the algorithm. Computational costs of the modified raking algorithm are discussed in Section 10.2.

Assume that blocks are of sufficient size that it is not necessary to check consistency and feasibility of the constraints in every case (but perhaps only when the weight fitting does not succeed in a few steps). Then the key calculation is fitting the weights. For production runs, data structures and programs should be devised which take advantage of the sparseness of the A matrix (due to the fact that only a few classes are represented in each household). Then if S_1 is the total number of nonzero entries in A and S_2 is the sum (through the block) of the *squares* of the number of nonzero entries for each household, each Newton step requires about $5S_1/2 + S_2/2$ multiplications (plus a term independent of the number of households per block). In the samples studied here, $S_2 \approx 5S_1$; S_1 is bounded by the total population of the block. Thus the bound on the number of multiplications is approximately 15 × population total (counting the start as an iteration); the number of additions is comparable.

In an era in which even microcomputers have megaflop arithmetic capability, 8×10^9 floating point operations to reweight an entire census does not seem unreasonable. The calculation of weights might well take less computer resources than the "bookkeeping" data processing required in any method of incorporating undercount. Of course, if the procedure were applied to a sampled database, as in forming a public-use sample, the costs would be reduced correspondingly.

6.2 Inference Simulations

For the inference simulations, pseudo-blocks of 50 households each with only Hispanic members were drawn. These were treated as if they represented true blocks. Then simulated omissions were imposed on the these households, assuming that each member was (independently) omitted with probability equal to the undercount rate from Diffendal (1988), with two negative undercount rates truncated to 0.

The entire distribution of the "enumerated" block was represented by including in the pseudo-Census roster the true composition and the possible compositions obtained by omission of one or more household members, each weighted by its probability under the model.

The pseudo-Census roster with undercount was then reweighted to the original pseudo-block totals for number of households and of individuals in each adjustment class. Both the pseudo-Census roster and the reweighted roster were compared to the original pseudo-block.

The purpose of organizing the simulation in this manner was to remove variability due to randomness in the rate of omissions in a block (around the mean undercount rate) and in the distribution of the omissions among the households in the block. Furthermore, feasibility is guaranteed because the original households are always included (with weights) in the pseudo-Census roster. One way of regarding this setup is that each simulated block represents a very large population in which observed undercount rates and the distribution of observed compositions approach their expectations.

Several sets of statistics were used in evaluation of the reweighting procedure. These were all chosen because they summarized household characteristics that are not functions of the populations by adjustment class. The first set was the distribution of sizes (number of members) of households. Note that the mean number of persons per household, like any function of the class totals and household count, will automatically be adjusted to the correct (pre-undercount) values; the distribution of sizes, however, is not controlled by the adjustment procedure.

The second set of statistics was the distribution of number of *adult* (over 14 years old) members in households with one or more *children* (up to 14 years old). In this case, the mean is not automatically adjusted to the correct value, since it depends on the joint distribution of counts from different classes within households as well as on marginal totals.

The last two sets of statistics were the distribution of the age group (coded from 1 to 5 as in the formation of the adjustment classes) of the *oldest male* in the household (coded 0 if no male is present), and the same distribution for households with one or more children. Again, neither the distribution nor its mean are directly constrained to their true values.

The results of these simulations are summarized in Table 8. Because almost all of the differences noted here are highly significant (relative to between-pseudo-block variances of the differences), standard errors are not shown in the tables. The lines of each table are labelled "true" (for the original pseudo-blocks), "enum" (for the simulated enumerated blocks, *i.e.* after omissions due to undercount), and "adjust" (enumerated blocks after adjustment for undercount). Every column except the means should be read as a percentage of households in the block.

Table 8

Inference simulation results

Size distribution

	size 1	size 2	size 3	size 4	size 5+	mean
true	7.240	16.200	20.240	22.600	33.720	3.971
enum	10.349	19.631	21.772	20.690	27.558	3.632
adjust	7.372	16.421	20.596	21.392	34.219	3.971

Size distribution (number of adults) for households with children

	size 0	size 1	size 2	size 3	size 4	size 5+	mean
true	0.000	6.925	58.404	17.214	9.125	8.332	2.585
enum	1.736	18.309	49.874	15.965	7.677	6.439	2.323
adjust	0.924	13.277	48.557	18.223	9.810	9.209	2.562

Age of oldest male (by five age classifications)

	none	age 1	age 2	age 3	age 4	age 5	mean
true	7.080	4.000	28.680	33.800	21.960	4.480	2.730
enum	9.981	7.388	26.296	30.972	21.160	4.203	2.585
adjust	7.853	5.989	26.307	33.439	21.931	4.480	2.690

Age of oldest male (by five age classifications) for households with children

	none	age 1	age 2	age 3	age 4	age 5	mean
true	3.602	6.214	30.744	42.649	15.843	0.949	2.638
enum	5.809	11.723	27.321	39.096	15.158	0.894	2.488
adjust	4.272	9.069	27.242	42.038	16.418	0.962	2.601

Household size distribution was biased downwards in the enumerated blocks. As well as correcting the mean, adjustment brought the estimated percentage for every size substantially closer to the true percentage.

The distribution of number of adults in households with children was also biased downwards. The majority of these households had contained two adults, so this size category was most understated by the enumerated statistics. Due to the log-linear structure of the adjustment, however, the most extreme adjustments were made to the largest and smallest households. Thus, the highest size categories were slightly overadjusted and intermediate categories were underadjusted; the "size 2" category was adjusted a small amount in the wrong direction. Nonetheless, the mean of the adjusted distribution was much closer to the "true" value than the adjusted mean was.

The story is similar for the distributions of age of oldest male. Although these statistics are only indirectly related to the counts by class, in almost every case the adjusted distributions and means are closer to the "truth" than are the unadjusted distributions and means.

In summary, these simulations suggest that these weighting adjustments can improve estimates of measures of household structure as well as the aggregate counts for which they were intended. However, reweighting does not provide accurate adjustments with certain configurations of the data, such as the many households with two adults noted above; to deal with these situations may require a model-based imputation method such as that outlined by Zaslavsky (1989).

7. THE USE OF WEIGHTED DATA

The product of the methods of the preceding sections would be a census roster in which households have weights, persons in households have weights adopted from their households, and institutionalized persons have individually assigned weights. This section outlines the use of these rosters for various Census purposes.

7.1 Formation of Tables of Counts

As with any data set of weighted observations, the sum of weights replaces the simple count of observations in forming tables. The only problem created by the use of weights is that of obtaining integer entries in the tables. This problem arises even before the calculation of household weights: when the estimated omissions are calculated, the counts in each class will not in general be integers.

If the adjusted totals by class are rounded to be integers, any table that aggregates classes (for example, a count of adult males that is a sum of counts of adult males from different classes) will also contain integers, since it must be consistent with those totals. For tables that are not based on those totals, summing the weights in a particular group may not necessarily generate integer counts. For example, if a class combines women of ages 20-40, a sum of weights for women aged 20-30 would not necessarily be an integer. In any case, it seems unlikely that all class weights would be rounded since this might well lose the entire adjustment to roundoff error. However, it should be possible to use existing Census Bureau integerizing methods ("controlled rounding") to deal with these problems, especially where non-disclosure requires that published counts be rounded anyway (Cox *et al.* 1986; Cox 1987).

7.2 Formation of tables of sums and means

Generally, sums (of continuous quantities) and means are not expected to be integers, so the issue of rounding does not arise. Also, tables based on long-form information are already derived from a sample so an additional source of weights should not change the process much. A deeper issue is that of the values of non-classification covariates to be assigned to households that are "weighted in" to the census; this is discussed in Section 8.

7.3 Public Use Samples

The public use tapes are a sample of census records that are released for further analysis by consumers of census data.

To generate these samples from weighted census rosters requires only that the sampling procedure be modified slightly to make sampling probabilities proportional to weights. Even on the 5% tape (the highest sampling rate), the weighted sampling probabilities should be smaller than 1. Once these tapes are produced, the user would not have to be aware of the adjustment and weighting process that had gone into generating them.

The public use tapes are the source of data for many of the more complicated analyses by sociologists, economists, planners, etc. in which the details of household composition, as well as counts of persons, are of importance. It is important that these tapes could be generated easily and used like raw census data.

As a service to those users of the public use tapes who wish to check the sensitivity of their analyses to the undercount adjustment, the tape should include factors (the inverse of the adjustment weights attached to the household records in the original census rosters) that would allow the user to reconstruct the equivalent of the unadjusted census.

8. ADJUSTMENT OF COVARIATES THAT ARE NOT USED IN CLASSIFICATION

The methods described above guarantee that weighted block totals by variables used in classification, such as sex, race, and age group, will equal the adjusted block totals. However, these lists will also be used to accumulate totals or counts for variables such as income and education that might not be used in the classification scheme. This section will consider the effect of these adjustment methods on such statistics. For concreteness of exposition, income will be used as the main example. Income is an important non-classification variable; some research suggests that revenue allocation programs may be most affected by errors in measurement of income. (National Academy of Sciences 1985).

In general, there are two possible sources of bias in the estimation of a non-classification covariate: (1) bias in adjustment of household composition, and (2) systematic differences between fully enumerated households and households with similar composition that are omitted (entirely or in part). However, if we have an estimate of mean income for the block, we can make the weighted mean for households in the block equal the estimated (adjusted) mean in much the same manner we make the weighted counts of individuals in the block equal the estimated (adjusted) counts.

8.1 Household Composition Bias

In this section we will assume that the average income level associated with a certain household composition is the same for fully enumerated households and those which are partly or wholly omitted from the enumeration. In other words, we consider here the case in which omission is noninformative for income.

Suppose that household income is a sum of independent contributions from persons of each class in the household (*i.e.* suppose that the contribution to income from persons in each class are independent of what other members are in the household). Then weighted household income totals would be an unbiased estimate of the true income totals (when adjustment rates are correct), since the sum of incomes would be a linear function of class counts for the block. However, under the more realistic assumption that linearity does not hold, misallocation of persons between households (and corresponding misrepresentation of household composition in the adjustment) could lead to bias in income estimates. Thus, for example, the average income of households with two children might not be the mean of the average income of one-child and three-child households (with the same composition of adult members). Then the weighting procedure might introduce the correct number of children but if, on the average, too many (compared to the truth) two-child households were created relative to one- and three-child households, estimates of household income would be biased.

Our procedure tends to fit weights that make the "adjusted-in" households similar in composition to those that are common in the enumeration. However, the adjustment is described only by adjustment class totals, which do not carry detailed information on the composition of the omitted households. Thus, if certain household compositions are disproportionately undercounted they may be underrepresented in the weighted lists, and if these compositions are associated, for example, with lower incomes, the total income estimates will be biased upwards.

This is essentially a problem of potential lack of fit of the model used in adjustment to the patterns in the data. The most severe biases might appear in statistics that refer specifically to household composition, such as the number of single-parent families.

If composition bias were found to be a serious problem, one approach to controlling it would be to augment the class adjustment rates with additional information that describes the joint omissions of persons from different classes (or grouped classes).

8.2 Response Bias

It is not unreasonable to think that, of a group of households with the same composition, those which are missed in the census will differ systematically in some characteristics from those that are enumerated. In other words, omission may be a form of nonignorable nonresponse. For example, households with lower incomes and educational levels may be more likely to be missed altogether, or to omit some members from their roster; income and education are not classification variables and therefore are not directly adjusted.

Whole-household adjustments are represented in the proposed methods by upweighting households, preserving the values of all covariates. The implicit assumption is that the omitted households do not differ on these covariates from enumerated households with similar composition. There is no information available in the block being adjusted to contradict this assumption. However, it should be possible to collect information in the PES on the differences between enumerated and missed households, which could be incorporated into the adjustment. For example, the income of wholly omitted households might be related to the mean income of enumerated households with the same composition by a linear regression; then the added (weighted-in) households could be imputed the income obtained by applying the linear regression function to the income of the enumerated donor household. Little and Rubin (1987) discuss relevant methods for missing data problems with informative nonresponse. Another approach that is integrated with the weighting adjustment methodology is described in the next section.

Within-household adjustments are represented by downweighting a household with certain enumerated characteristics and upweighting another household that contains an additional member or members. In the absence of further adjustment, the characteristics of the upweighted household, rather than those of the enumerated household from which the weight was taken, will apply to the "weighted-in" component.

This poses problems that cannot be resolved without collecting some data (from a subsample of the PES). For example, if a *child* were omitted from the household roster, there is no reason to think this would lead to misreporting of income. If households with more children had a higher mean income than those with fewer children, then the weighting would tend to over-estimate mean incomes.

If an *adult* were omitted from the roster, this might also mean that the same adult's income (if any) would be left out of the reported household income. It is plausible that the mean unreported income in this situation would be positive but less than the mean income of the corresponding adults in households where all adult members appear on the roster. For a stereotypical example, consider a family on public assistance that does not report an adult male member, whose income would otherwise be deducted from the assistance level, and whose residence is somewhat inconsistent. That member's income is likely to be less than that of a permanently resident adult male in a family that does not depend on public assistance. Thus, neither the income of the enumerated household nor that of the "weighted-up" household would be an accurate imputation for the adjusted household.

No direct correspondence is established between households that are down-weighted and those that receive additional weight. Thus an unadjusted income cannot be carried over directly from the enumerated household to the "weighted up" household. However, with some research comparing the incomes of enumerated and missed households, the incomes of down-weighted households could be used in adjusting incomes. For example, the mean household income of the reweighted block could be constrained to be equal to that of the block before adjustment.

8.3 Weighting Adjustment of Non-classification Characteristics

Suppose that adjusted summary information (by block) is available on some characteristics of households other than counts of individuals by adjustment class. For example, we might have an adjusted estimate of mean income or of the proportion of single-parent families, possibly from a regression model. As long as the summary statistic can be represented as a weighted sum of covariate values for each household, then conformity to the desired adjusted value can be imposed by a linear constraint on weights which can be made part of the weighting adjustment methodology of this paper. Thus, in the income example, we would constrain the weighted sum of incomes to equal the product of the number of households and the adjusted mean income. To adjust the proportion of single-parent families, we would constrain the weighted sum of 0-1 indicators for that status to the desired total count.

8.4 Summary and Implications

The methodology proposed will upweight households, and without further consideration of possible biases, will carry along the characteristics of the upweighted households. If the size of the adjustment and the biases introduced in household characteristics are both of small order, the overall bias in estimated block characteristics will be of second order and should not be a major problem. Some simple regression adjustments might make it possible to reduce the biases by an additional order of magnitude.

9. SUGGESTIONS FOR FUTURE RESEARCH AND DEVELOPMENT OF METHODOLOGY

This section summarizes a number of suggestions for implementation and further development of this adjustment methodology.

9.1 Post Enumeration Survey (PES) Data-gathering and Statistical Modeling

Omissions of persons in enumerated and omitted households should be distinguished in the PES and the two omission rates modeled separately for each adjustment class. Rates of omissions of whole households should also be modeled (Section 4). A variety of measures (as in Section 6.2) could be used to compare the composition of ''weighted-in'' households to that of omitted households found in the PES; if research found that ''composition bias'' was a significant problem, higher-order statistics should be developed (Section 8.1). A sample of PES households that were omitted in the Census should be administered the long form, so that the relationship between omission and covariates such as income and education could be modeled for the adjustment (Sections 8.2, 8.3).

9.2 Feasibility of Adjustments

The methods of Section 5 should be tested and compared using PES data.

9.3 Multiple Imputation

Although the procedures proposed in this paper operate deterministically, there are a number of sources of uncertainty in statistics based on the weighted records. These include: uncertainty in estimation of undercount rates; variability in class undercount rates from block to block around the national mean; binomial variability in the actual number of omitted households or individuals around the expected number given the undercount rate; uncertainty regarding differences between covariate values for omitted households and for enumerated households that are weighted up to replace them.

For research uses, files could be prepared that would represent all of these forms of uncertainty by multiple imputation (Rubin 1987). Two or more versions of the reweighted data set could be represented by including several sets of weights on the file. Researchers could repeat their analyses using each set of weights in turn. The variability among the statistics calculated on the different versions gives an estimate of the variability introduced by the process of undercount adjustment. Zaslavsky (1989) discusses procedures for multiple imputation in this setting.

10. SUPPLEMENTS

10.1 Choice of Objective Function for Weighting

A number of objective functions have been proposed for calculating an optimal fitted table (usually in the context of contingency tables, *cf.* Fagan and Greenberg 1988). In each case the function takes the form $T = \sum T_1(W_i)$, where T_1 takes one of the forms displayed in Table 9. Each of these functions can be standardized to an equivalent function T_0 by multiplication by a constant coefficient and adding a linear function of W, so that $T_0(1) = 0$, $T_0'(1) = 0$, $T_0''(1) = 1$. Since $\sum W_i$ is constrained to a given value, the optimum weights will be unaffected. Then the standardized objective functions agree through the second term of their Taylor expansions about 1, and should give similar results when the weights are close to 1.

Table 9

Comparison of objective functions for table fitting

Name of fitting procedure	Objective function $T_I(W)$, usual form	Objective function $T_0(W)$, standardized form	Second derivative $T_0''(W)$
Least squares (minimum variance)	$(W-1)^2$	$(W-1)^2/2$	1
Raking	$W \log W$	$(W \log W) - W + 1$	$1/W$
Maximum likelihood	$-\log W$	$W - 1 - \log W$	$1/W^2$
Minimum χ^2	$(W-1)^2/W$	$(W-1)^2/2W$	$1/W^3$

in the degree of asymmetry between the costs of downweighting and upweighting, determined by the exponent of W in the second derivative, $T_0''(W) = W^{-k}$. The least squares procedure ($k = 0$) treats up-and down-weighting completely symmetrically and therefore may yield zero or negative weights. As k increases, the cost of upweighting becomes smaller relative to that of downweighting. All of the other objective functions ($k > 0$) give every observation in the raw data a positive weight; in the case of the "raking" function, this is obvious from the form of the weights as shown in Section 3.3. The use of the "raking" function here in preference to maximum likelihood or minimum χ^2 is motivated by the simple form of its solution and by the analogy to raking in contingency tables. Cressie and Read (1984) systematically study the properties of this family of measures of fit.

10.2 A Cyclic Descent Methodology for Fitting Weights

In this section we present a fitting methodology analogous to iterative proportional fitting (IPF) in contingency tables. In IPF, the cell counts are transformed multiplicatively in such a way that the cross-products are preserved (the condition for minimization of the objective function) while the table is made to conform to each set of marginal constraints in turn. The algorithm converges to a table that satisfies all of the constraints, and perforce preserves the cross-products as well (Bishop, Feinberg and Holland 1974; Ireland and Kullback 1968).

In our setting, the weights are required to have the log-linear form $W_i = \exp(a_i'\lambda - 1)$ derived in Section 3.3 while satisfying the constraints $AW = B$. In this exposition we will assume that the total weight constraint $\sum W_i = H$ is omitted from $AW = B$, and that A is of dimension p (constraints) × I (number of household compositions). We will proceed through a series of steps in each of which each weight W_i is multiplied by $c\rho^{a_{ji}}$ to obtain a new weight W_i', thus preserving the log-linear structure; c and ρ are chosen so that the constraints $\sum W_i' = H$ and $\sum W_i' a_{ji} = b_j$ are satisfied. By proceeding cyclically so that $j = 1, 2, \ldots p$ indexes each constraint in turn, the algorithm eventually converges to weights that satisfy all of the constraints.

On step j of cycle t, the new weights are given by $W_i^{(t,j)} = c\rho^{a_{ji}} W_i^{(t,j-1)}$ (initialized for $j = 1$ by using the last weights from the last cycle, $W_i^{(t,0)} = W_i^{(t-1,p)}$). Then c and ρ must satisfy

$$\sum_i c\rho^{a_{ji}} W^{(t,j-1)} = H, \quad \sum_i a_{ji} c\rho^{a_{ji}} W_i^{(t,j-1)} = b_j. \tag{5}$$

Eliminating c from these equations, ρ is a root of

$$\sum_i \left(Ha_{ji} - b_j\right) W_i^{(t,j-1)} \rho^{a_{ji}} = 0. \tag{6}$$

We must have $Ha_{j,min} \leq b_j \leq Ha_{j,max}$ where $a_{j,min}$ and $a_{j,max}$ are respectively the minimum and maximum values of a_{ji}. If this were not the case, constraint j could not be satisfied with any weights. Thus there must be at least one root ρ, and if the a_{ji} are non-negative, the expression is increasing in ρ so this root is unique. The actual value of ρ is determined then by Newton's method, or by a closed-form formula for the roots of a polynomial (since with the original A, a_{ji} is the number of class j members in a household, which is an integer rarely exceeding 2).

While we have not yet proven that this algorithm always converges, we have found it to be successful in practice. This algorithm does not require any matrix inversion, and if the a_{ji} are small integers, then at each step, the recalculation of the weights involves calculating only a few integral powers. Furthermore, if some constraints take the form of simple marginals, the adjustment for those constraints takes the form of a conventional raking step.

If the original constraint matrix A is used, the procedure may take advantage of the sparseness of A (which is a consequence of the fact that only a few classes are represented in each household). At each step (say, adjusting to fit margin b_j), only the weights corresponding to non-zero a_{ji} need be modified; thus only S_1 multiplications (the number of nonzero entries in A, which is less than the population of the block) and perhaps $3S_1$ additions are required per cycle, as compared to $5S_1 + S_2$ operations per iteration with Newton's method. On the other hand, the rows of A tend to be highly dependent, so convergence may be slow (typically 20 cycles in our simulations); orthogonalization of A destroys the sparse structure of the coefficients. Thus, unless S_2 is much larger than S_1 (or unless some other method is devised to accelerate the algorithm), raking is not faster than Newton's method.

ACKNOWLEDGEMENTS

This research was supported by Joint Statistical Agreements 86-8 and 87-7 between the U.S. Bureau of the Census and Harvard University. The author has benefited from comments by Donald B. Rubin and other participants in the seminar on census undercount at the Harvard University Department of Statistics, and Nathaniel Schencker, Nash Monsour, and other members of the Undercount Research Staff in the U.S. Bureau of the Census. Simulations used data made available by the Inter-university Consortium for Political and Social Research and originally collected by the U.S. Department of Commerce, Bureau of the Census.

REFERENCES

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

BISHOP, Y.M.M., FIENBERG, S.E., and HOLLAND, P.W. (1974). *Discrete Multivariate Analysis.* Cambridge: M.I.T. Press.

BUREAU OF THE CENSUS (1985). Census of Population and Housing, 1980: Public Use Microdata Samples.

CILKE, J.M., and WYSCARVER, R.A. (1988). The Individual Income Tax Simulation Model, in Office of Tax Analysis, *Compendium of Tax Research* 1987, Washington: Government Printing Office.

COX, L. (1987). A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association*, 82, 520-524.

COX, L., FAGAN, J., GREENBURG, B., and HEMMIG, R. (1986). Research at the Census Bureau into disclosure avoidance techniques for tabular data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 388-393.

CRESSIE, N., and READ, T.R.C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society*, Series B, 46, 440-464.

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

DIFFENDAL, G. (1988). The 1986 Test of Adjustment Related Operations in Central Los Angeles county. *Survey Methodology*, 14, 71-86.

FAY, R.E. (1986). Implications of the 1980 PEP for future census coverage evaluation. U.S. Bureau of the Census, unpublished.

FAGAN, J.T., and GREENBERG, B. (1988). Algorithms for making tables additive: Raking, Maximum Likelihood, and Minimum Chi-square. *Proceedings of the Section on Survey Research Methods, American Statistical Association* (forthcoming).

GASS, S.I. (1964). *Linear Programming: Methods and Applications*. New York: McGraw-Hill.

IRELAND, C.T., and KULLBACK, S. (1968). Contingency tables with given marginals. *Biometrika*, 55, 179-188.

LITTLE, R.A., and RUBIN, D.B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

NATIONAL ACADEMY OF SCIENCES (1985). *The Bicentennial Census: New Directions for Methodology in* 1990. Washington: National Academy Press.

OH, H.L., and SCHEUREN, F.J. (1978). Multivariate ratio raking estimation in the 1973 Exact Match Study. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 716-722.

PURCELL, N.J. (1979). Efficient estimation for small domains: a categorical data analysis approach. Ph. D. dissertation, University of Michigan.

PURCELL, N.J., and KISH, L. (1979). Estimation for small domains. *Biometrics*, 35:365-384.

RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

SCHEUREN, F.J. (1981). Methods of estimation for the 1973 exact match study. In *Studies from Interagency Data Linkages*, Washington: Social Security Administration.

ZASLAVSKY, A.M. (1989). Representing Census undercount at the household level. Ph. D. thesis, Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts.