Developing A Coverage Adjustment Strategy for the 2002 Census of Agriculture

Matthew J. Fetter and Phillip S. Kott

USDA/National Agricultural Statistics Service
RDD Division
3251 Old Lee Highway, Room 305
Fairfax, Va. 22030
Matt_Fetter@nass.usda.gov Phil_Kott@nass.usda.gov

Keywords: Undercoverage, Calibration, Range Restrictions

Introduction.
The National Agricultural Statistics Service (NASS) assembled an electronic list called the Agricultural Census Mail List (CML) of approximately 2.8 million potential farm operators for the 2002 Census of Agriculture in the United States. Although much effort has been expended on making this list as complete as possible, the coverage of farms is not complete. In addition, though strong efforts are made to insure that no farm is represented more than once on the list, farm duplication does occur resulting in a certain amount of overcoverage of certain farms. NASS is primarily interested in producing Agricultural Census estimates that are fully adjusted for list undercoverage problems at the county level as undercoverage is considered the more severe problem then overcoverage. To this end, explicit estimates of the CML undercoverage for a specified set of demographic variables, called calibration variables, will be computed from an area frame sample. These estimates, when combined with the CML estimates for the same items will result in target values. After initial weights are assigned to CML respondents to account for nonresponse, these weights will be further adjusted in an attempt to reproduce the target values for each of the calibration variables. Compensation for the *state*-level CML undercoverage for each of the calibration variables can be accomplished in this fashion. Since each farm with Census data will be given a fully-adjusted weight by this process, it becomes a simple matter to estimate county-level totals for every Census variable, not just the calibration variables.

In addition to correcting for undercoverage, NASS wants to adjust Census-based estimates for measurement errors (duplication, erroneous collections, biased imputations, etc.) by not allowing major commodity totals to vary too far from established benchmarks. This prompts the use of a second set of calibration variables with targeted totals based on these established benchmarks rather than the simple sums of census list and undercoverage estimates used for the first set. For convenience, we call the calibration process involving both sets of variables "coverage adjustment."

Most calibration targets will be determined at the state level. Some small states will be combined into "calibration regions". The calibration multi-state regions are likely to be: AZ/NM, UT/NV, DE/MD, and CT/ME/MA/NH/RI/VT. These groupings are based on 1997 data and will be grouped together to increase the precision of the undercoverage estimates. In what follows, a "state" refers to a calibration region where appropriate.

The methodology for undercoverage adjustment for the 2002 Agricultural Census was developed by working with the 1997 Agricultural Census data. The remainder of this paper will describe the methodology as applied to the 1997 data. Its application to the 2002 Agricultural Census data should be very similar.

Working with 1997 Data--Determining Targets to Correct for Undercoverage.
Computing targets for the variables used to correct for Census undercoverage involved a number of steps. First, estimates of the CML undercoverage had to be made using an area frame sample. The 1997 June Agricultural Survey (an area frame sample composed of 10,821 land segments) was used for this purpose. Each sampled land segment contained a number of farm tracts (the number could be zero). A tract is that part of a farm operation wholly within the land segment. NASS checked whether the farm associated with each tract on a sampled land segment could be linked to a CML operation. The tracts with no link to the CML were deemed "not on the mail list (NML)." Data values associated with NML tracts were used to estimate the state level undercoverage of the CML for the first set of calibration variables. The state level totals for these variables were then summed to yield national totals.

State level NML estimates for the number of farms in a state could be used directly in determining calibration targets (CML + NML). The other calibration targets to be used for that purpose were the number of farms of a certain type (e.g., in a particular sales class or with a primary operator of a particular race). Most of these had unacceptably high state level standard errors. As a result, more reliable national level NML estimates were used to *smooth* state estimates. The smoothed state NML estimate was computed by taking a weighted average of the actual state estimate and a prediction of the state total based on national and state level numbers (e.g., the number of NML farms in the state, the fraction of farms with black owners on the state's CML, and the national relative difference between the fraction of black owners on the NML and CML). The weighting factor was chosen to approximately minimize mean squared error under a random effects model. The smoothed NML estimates were then added to corresponding CML estimates to obtain coverage adjusted state level totals, which served as calibration targets. Details for the smoothing procedure can be found in Kott 2002 [3].

The decision on which demographic items to select as calibration variables was largely based on the size of the differential between the estimated proportion of farms having a particular characteristic on the CML and the same estimate for the NML. Characteristics for which this differential was large would be good candidates for inclusion on the list of calibration variables. Commodity items were selected based on cash receipts at the U.S. level, and their relative importance at the state level (see table1).

Table 1.

Calibration Variables

Total Value of Production (Dollars)	Operator Age	Gender	Operator Race	All Farms	Commodity Presence
0	<25	Male	Black	All	Cattle
1-999	25-34	Female	AI/ASN/Other		Dairy
1K-2.5K	35-44				Sheep/Goats
2.5K-5K	45-54				Poultry
5K-25K	55 +				Hogs
25K-100K					Fruit/Nut/Berry
100K-500K					Vegetables
500K+					Nursery/Hort
Extreme Ops (EO)					Tobacco
Commodity Inv/Prod					Horse/Mules
Corn Acres HV	Sugarbeets Acres HV	Broiler Production			Cropland
Soybean Acres HV	Tomatoes Acres HV	Turkey Production			CRP
Wheat Acres HV	Hay Acres HV	Cattle Inventory			
Cotton Bales	Apple Acres HV	Dairy Cow Inventory			
Potato Acres HV	Orange Acres HV	Hog Inventory			
Sugarcane Acres HV	Grape Acres HV	Layer Inventory			
Tobacco Acres HV	Lettuce Acres HV	Rice Acres HV			
Commodities varied by state. HV=harvested, AI =American Indian, ASN=Asian					

The Calibration Procedure.
One approach to weight calibration is to compute an adjusted weight, w_k, for each of the k farms, $k=1,\dots, n$, that minimizes the sum of the relative squared differences from the original weights, d_k, subject to the following set of constraints:

A: The P target values t_p, $p=1,\dots,P$ are achieved using the new weights.
B: The new weights lie within some range of values that are deemed acceptable. For the Agricultural Census this interval would be [1,6] for the final calibrated weights, w_k.

Problem 1.
In mathematical notation this can be restated as:

Minimize: $\sum_1^n (w_k - d_k)^2 / d_k$, $1 <= d_k <= 2$

subject to:

A) $\sum_1^n w_k y_{kp} = t_p$, $p=1,2,\dots, P$, $y_{kp} >= 0$

B) $1 <= w_k <= 6$

Note that the initial weights for all farms are in the interval [1,2] and that all calibration variables are zero or positive for each farm.

A problem of this type can be solved using non-linear programming algorithms that are available in software packages such as SAS– if a solution exists. In our early work using the program CALJACK (Crouse 1999) , it was apparent that in many cases, any set of weights that could be found that was consistent with the first set of constraints (A) ended up violating the second set of constraints (B). This resulted in a failure to produce a useful set of weights and left us with the time consuming problem of determining which constraints needed to be relaxed so that a solution could be found. What we needed was a computer program that would determine which constraints to drop, drop them, and then produce the best set of weights that it could find while guaranteeing that all weights would be in the desired interval . It was decided that we would write a program in SAS that would accomplish this. What we next needed to determine was the priority of the constraints.

It was decided that the calibrated weights would not be allowed to stray from the original interval constraint of [1,6]. This bound on the resultant weights would be the highest priority. Any set of weights obtained from the program would be required to reflect this constraint. It was also decided that it was not actually necessary to hit all the target values exactly. Calibrated weights that produced values that were within a tolerable range about the targets would be considered satisfactory. In fact, because the target values were in many cases estimates themselves, subject to uncertainty and error, this approach seemed natural. Hitting the target value exactly was then given the lowest priority and would be the first constraint to be relaxed. Only if a value within the tolerance interval for a calibration variable could not be achieved would that variable be removed from the calibration altogether and thus in effect, force the relaxation of the tolerance constraint for the target.

Problem 2.
Expressed mathematically we would :

Minimize : $\sum_1^n (w_k - d_k)^2 / d_k$

Subject to:

A) $\sum_1^n w_k y_{kq} = t_q$, $q=1,2,\dots,Q$, $y_{kq} >= 0$.

B) $1 <= w_k <= 6$.

resulting in

C) $t_r - a_r <= \sum_1^n w_k y_{kr} <= t_r + b_r$, $r=Q+1,\dots,R$, $a_r, b_r > 0$, $y_{kr} >= 0$

D) $\sum_1^n w_k y_{kp} > t_p + b_p$ or $\sum_1^n w_k y_{kp} < t_p - a_p$, $p=R+1,\dots, P$, $y_{kp} >= 0$.

The major difference between *Problem 1* and *Problem 2* is that the set of active calibration variables whose target values are attained in A is unknown in *Problem 2.* The Q targets that are attained in A, the P-R target constraints dropped in D , and the values of Q and R themselves, would need to be determined by the program. The calibrated weights produced by the program would always be the result of solving *Problem 2* for the set of Q active calibration variables. The initial targets would be specified by the user, along with the endpoints of the tolerance interval given by $[t_j - a_j , t_j + b_j]$. The targets would not be required to be centered in the interval. As will be explained below, in certain situations, the target could switch to an end point of the interval.

The R-Q calibration variables in C, for which an acceptable value within the tolerance interval is attained without being in the active set of calibration variables (referred to henceforth as "floaters") is in part a function of their relationship to those calibration variables in the active set.

More on Tolerance Ranges.
One of the problems with this approach is the somewhat subjective determination of what the values a_j and b_j should be. In fact, choosing these values did cause some difficulty and consternation. The number of farms target (All Farms) had no tolerance range beyond the estimated state level total (CML + NML) as well as the extreme operations (EO) for which the NML estimate was zero. The tolerance range for all other demographic variables was the estimated state total for the variable (CML + smoothed NML) plus or minus one-half of one estimated standard error. This choice limited the cumulative deviation from the estimated total for a variable when state-level totals were combined to create a US-level total. The state level tolerance ranges for commodity targets were provided by NASS commodity specialists.

Calibration Program Overview.
The program we developed was written in SAS and will be referred to here as *CalBuilder.* CalBuilder's approach to solving Problem 2 is to solve Problem 1 recursively, adding to the set of active calibration variables one variable at a time. The basic algorithm used by the program is described by Singh and Mohl (1996) where it is referred to as the Linear Truncated Method.

At each iteration, a variable is added as a temporary member to the active calibration set and a search for a solution is made. If the search is successful then that variable becomes a permanent member of the active calibration set and another variable is introduced to the active set for the next iteration. If the search for a solution is not successful, that variable is removed from the active set and another variable is entered into the active set. Variables that have been set aside on previous iterations are eligible to be re-entered into the active set so that multiple attempts can be made to obtain an acceptable value for that variable. At the end of each iteration, the calibration variable for which the current value is furthest out of range (in relative terms) is brought into the active set of calibration variables. Any variable for which the current value is within range is not eligible to be added to the calibration system. It is typical that for many calibration variables, acceptable values are obtained without becoming members of the active calibration set.

Program Details.
The program starts by calibrating weights so that the two most precise targets would be achieved– number of farms, and number of extreme operator farms (EO) at the state level. Once a solution is obtained , these two variables become the first two permanent members of the active calibration set. Obtaining the exact target values for these two variables is considered tantamount to any continuation of the calibration process.

CalBuilder makes up to two passes through the calibration variable list. On the first pass, CalBuilder will attempt to hit the central target value for each variable contained in the active calibration set. If necessary, CalBuilder will make a second pass through the list of calibration variables for which an acceptable value was not obtained on the first pass. However, on the second pass, CalBuilder will switch the target from the central value it used in the first pass to one of the end points of the tolerance interval.

The first pass through the calibration variable list terminates when either acceptable values have been achieved for all calibration variables, or target values can not be obtained for any calibration variable whose current value is outside of the tolerance range. CalBuilder will then attempt to hit the endpoint nearest to the current value for all calibration variables whose value is still out of range. This second pass uses the same one variable at a time approach that was used in the first pass. The second pass terminates when it can not achieve endpoint values for any calibration variables that were out of range after the first pass.

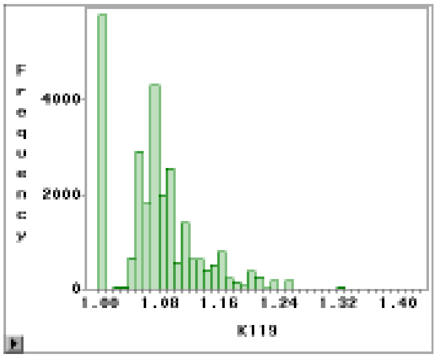
The result of running CalBuilder is the attainment of a set of "fuzzily" calibrated weights which produces most of the target values or range endpoints exactly, produces values within the tolerance range for some variables without being entered into the system, and produces a few "disappointing" values outside the tolerance range for some calibration variables.

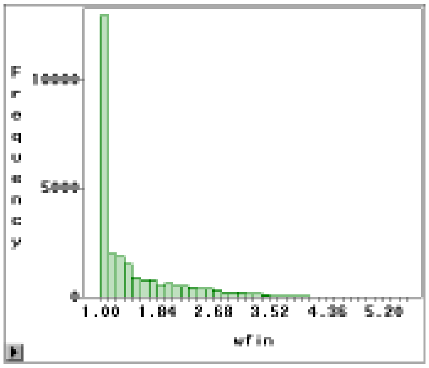
Results.
A test run of CalBuilder was made on the 1997 Agricultural Census data. There were approximately 34 calibration variables for every state. Each state had the same set of approximately 29 demographic variables and about 5 or 6 commodity variables. The commodity variables varied by state. There were 40 states, about 34 variables per state, which gives about 1360 total variables being calibrated over all. The following table shows the results of running the CalBuilder program for all states.

1997 Test Calibration Variable Counts Summed to U.S. Level			
Central Target	Floaters	End Point Target	Out of Range
903	380	19	58

In only 8 of the 40 states did CalBuilder find acceptable values for all calibration variables. The most variables that were out of range for a state was 7 (one state). The most floaters in a state was 13 (4 states). The most variables that were successfully calibrated to an end point in a state was 3 (1 state). Although only 13 states benefitted from the endpoint search, it actually appears to have reduced the number of out of range variables by nearly 25% (19/[19+58]).

The graphs below show the distributions of the weights prior to calibration (K119) and after calibration (wfin) for a particular state. The calibrated weights in the graph on the right shows that many weights are being pushed to the lower boundary of 1 while greatly increasing the size of many of the weights as well. Note that for this state, few, if any weights are close to the upper boundary of 5 .





Because the method does not require the targets to be achieved exactly, a loss function that gives no loss for any value achieved within tolerance seems reasonable. It is difficult to say , however, if the solutions obtained are optimal in a mathematical sense. At the state level, the method did achieve acceptable values for most calibration variables (96%). It should be kept in mind that, especially for the demographic variables, the targets and ranges are merely estimates. The inability for CalBuilder to obtain a value within the stated range might actually result in a better indication in some cases. Better results might be achievable, particularly with the commodity variables, if calibrated weights were permitted to take on values less than 1. Commodity variables with census indications above the upper endpoint of the tolerance interval were especially hard to calibrate successfully due to the lower bound of 1 imposed on the calibrated weights.

When aggregated to the U.S. level, the calibrated demographic variables were well with in two estimated standard errors of the target with one exception. The aggregated results for the commodity variables were good for many items, but results were a little more disappointing than the demographic items results.

With respect to speed, computational time was reasonable. The calibration was carried out one state at a time. Typically a state has between 20,000 and 75,000 weights, although some have nearly 100,000 or more, and will take from 15 minutes to several hours to complete. Computation time is not only a function of the number of weights, but also a function of the relationship of the data and the targets. In all, nearly 1,700,000 weights were calibrated within a 24 hour period using one Pentium IV 2.0 Ghz processor with 512 M RAM and required no user interaction other than executing the program itself.

References:

[1] Crouse, C.(1999) "Evaluating a One Number Approach to the Agricultural Census". *Proceedings of the Federal Committee on Statistical Methodology Research Conference*, pp. 46-55.

[2] Deville,J-C.,and Sarndal,C-E.(1992). "Calibration Estimators in Survey Sampling". *Journal of the American Statistical Association*, 87, 376-382.

[3] Kott, P.S. (Unpublished, 2002). "A Components of Variance Model Estimator for State Level Not-on-Mail List Proportions". USDA/NASS/RDD working paper.

[4]Singh, A.C. and Mohl, C.A.(1996). "Understanding Calibration Estimators in Survey Sampling". *Survey Methodology*, 22, 107-115.