An Overview of Coverage Adjustment for the 2007 Census of Agriculture

Matthew J. Fetter[1]

U.S. Department of Agriculture, National Agricultural Statistics Service, 1400 Independence Avenue, SW, Washington, D.C., 20250

Abstract

The National Agriculture Statistics Service (NASS) conducts a census of agriculture every five years (years ending in 2 or 7). NASS maintains a list frame (referred to as the Census Mail List or simply the CML). A census form is mailed to every name on the list. However, there are many farms in the U.S. that are not on this list. This is referred to as list frame undercoverage. If no effort was made to account for this fact, many of the estimates produced solely from responding farms on the CML would be biased due to this undercoverage. To alleviate the effect of this bias, NASS employs an area frame sample of farms. The area frame contains all the land area for the contiguous states and Hawaii and thus has complete coverage of all farms in the U.S. (except Alaska which currently has no area frame). NASS uses the area frame sample to estimate the number of farms that are not on the CML. NASS uses this information to adjust the weights of the CML respondents. This paper explains the methodology used to make these adjustments.

Key Words: Weight calibration, census, coverage error, regression estimator

1. Introduction

The Census of Agriculture (henceforth simply referred to as "the census") produces thousands of estimates, many of these computed at the national, state, and county levels. These estimates are published in tabular format, with each table often containing hundreds, or even thousands of estimated cell totals. Contrary to what the lay person might think, there can be a significant degree of uncertainty in many of these estimates. This is especially true for the many cells that represent a relatively small number of farms.

Much of this uncertainty comes from two sources—nonresponse error and list coverage error. NASS maintains a list of names that represent farms or are highly likely to represent farms called the Census Mail List (CML). For the census, an attempt to contact every farm on the CML is made. Nonresponse error occurs when a farm on the CML receives a census questionnaire, but fails to report some or all of the requested data.

[1] This paper presents views of the author and does not represent any official position taken by the National Agricultural Statistics Service.

Coverage error occurs because the CML does not include every farm in the country, or, in other words, does not "cover" the entire U.S. farm population.

NASS uses manual and machine data imputation in order to reduce the effects of nonresponse error on the census. Nonresponse weights are computed and applied to the responding farm records to further reduce this error. CML response rates are fairly high, and as a consequence, nearly all of the nonresponse-adjusted weights for CML respondents lie in the closed interval [1, 2]. Nonresponse error, and how NASS compensates for it is outside the scope of this article.

NASS addresses the issue of coverage error by computing coverage adjustments that are applied to the nonresponse weights of responding CML farms. These coverage adjustments incorporate information concerning farms that are Not on the CML (referred to henceforth as "NML") with information obtained from responding CML farms in such a way that coverage error is reduced.

At the national level, nonresponse adjustment and coverage adjustment each account for about 15% of the total farm estimate. This means that census data was not obtained for an estimated 30% of the farms in the U.S. Although many of these farms tend to be economically small in size, the amount of data that is not obtained is still considerable and would lead to some severely biased estimates if an effort was not made to account for these missing data.

The CML does contain a large percentage of the U.S. farm population. Economically large farms have a very high coverage rate by the CML, and conversely, farms that are very small have a considerably lower rate of coverage. In addition to being smaller farms, operators of NML farms tend to be younger, are more likely to be female, and are racially and ethnically more diverse than operators on the CML. These unique characteristics of NML farms and their operators make it doubly important that they are reflected in the census estimates.

Clearly, estimates based solely on the data collected via the CML would have a built-in undercoverage bias because many farms would have no possibility of being contacted. Coverage error would occur because information pertaining to NML farms would be completely ignored in the estimates. To reduce the effect of this flaw, some method of permitting information concerning NML farms to influence the estimates produced by the census were devised. The method used to achieve this is the topic of this paper.

2. Compensating for Coverage Error: The Area Frame.

NASS has the capability to generate estimates of the number of NML farms that exist, and through these estimates, can obtain indications of their associated characteristics. These estimates are obtained by employing a land based sampling frame referred to as the area frame. The area frame is composed of land segments that cover the entire land mass of the U.S. (except Alaska, which currently has no area sampling frame). Theoretically, every farm in the country is on this frame. The area frame provides samples that are used to support the NASS survey program. In census years, area frame sample sizes are increased to provide better estimates of farm counts.

To obtain estimates for farms that are NML, a sample of land segments is drawn from this frame. The location of the sampled land segments is determined and enumerators are

sent to these locations to account for all the agriculture that resides in the selected segment. Once the farms that operate land in these segments are identified, a determination is made as to whether that farm is on the CML or is NML. Next, all identified NML farms are asked to complete a census questionnaire. The area frame sample makes it possible to produce estimates of the number of NML farms and an estimate of the standard error associated with these estimates. The data collected from NML farms on the census form is used in conjunction with the corresponding nonresponse adjusted CML data to compute dual frame estimates for key census items. These estimates are then used to create benchmarks that are fundamental to the coverage adjustment process. These key items are referred to as calibration variables, and the corresponding benchmark for each of these variables is referred to as the calibration target.

3. Coverage Models

Accounting for the NML contribution to census estimates through the use of CML data is necessary because the area frame sample sizes are not nearly large enough to provide reasonably precise direct NML estimates for the many farm population quantities for which estimates are required. For example, the census is relied upon to provide estimates of many quantities at the county level. Area frame sample sizes are generally not large enough to produce reliable estimates for the NML domain at such a low level of aggregation.

It is a nearly impossible task to develop a coverage model that will perform well for every quantity being estimated by the census. Emphasis is placed on utilizing a relatively small set of key items in the coverage model that relate to a broad range of farm sizes and types, as well as farms having particularly sensitive characteristics. The extent to which the coverage rate for an arbitrary item can be predicted by this set of model variables is the extent to which coverage adjustment will improve the estimate for that item. It is inevitable that for some items, the coverage adjustment model will not be very good, and this is expected, but for the key items, and items related to the key items, the coverage adjustment model should work fairly well.

Truncated linear weight calibration is used to create regression-type coverage-adjusted estimators for all published cell estimates. To create these estimates, CML respondent data for the calibration variables are used as predictor variables in the model, as well as the calibration targets for those variables.

The form of these coverage-adjusted estimators of an arbitrary population total, t_y is given by (1).

$$\hat{t}_y = \sum_i w_i y_i = \sum_i g_i d_i y_i \qquad (1)$$

where:

w_i = the coverage adjusted weight for CML responding farm i,

d_i = the nonresponse weight for CML responding farm i,

g_i = the coverage adjustment for CML responding farm i,

y_i = the value reported by CML responding farm *i*.

NASS requires that the coverage adjusted weight, w_i, be generally restricted to lie in the closed interval [1, 6]. The w_i can be thought of as representing the inverse of the probability that a farm is on the CML and responds to the census. This would then imply that this probability is no less than 1/6 and no greater than 1 for any farm in the population.

The coverage adjustment factor, g_i is obtained using the iterative equation expressed in (2).

$$g_i^{(r)} = g_i^{(r-1)} + x_i \left(X' \Gamma^{(r-1)} X\right)^{-1} \left(t_x - \sum_i w_i^{(r-1)} x_i' \right) \qquad (2)$$

where:

$r = 1,2, \ldots, R$ indexes the iterations.

R = the number of iterations required for $\left(t_x - \sum w_i^{(r-1)} x_i' \right)$ to be sufficiently small.

$w_i^{(r-1)} = g_i^{(r-1)} d_i$

t_x = vector of calibration targets = $\left\langle t_{x1}, t_{x2}, \ldots\ldots, t_{xp} \right\rangle'$

X = CML respondent data matrix of calibration variables = $\left\langle x_1, x_2, \ldots\ldots, x_N \right\rangle'$

x_i = vector of calibration variables for CML responding farm *i* = $\left\langle x_{i1}, x_{i2}, \ldots\ldots, x_{ip} \right\rangle$

$$\Gamma^{(r-1)} = \begin{bmatrix} \phi_1^{(r-1)} d_1 & & \\ & \ddots & \\ & & \phi_N^{(r-1)} d_N \end{bmatrix}$$

$$\begin{aligned} \phi_i^{(r-1)} &= 0 \text{ if } g_i^{(r-2)} d_i < 1, \\ &= 0 \text{ if } g_i^{(r-2)} d_i > 6, \\ &= \prod_{q=0}^{(r-2)} \phi_i^{(q)} \text{ otherwise.} \end{aligned}$$

$\phi_i^{(0)} = 1$

$g_i^{(0)} = 1;\ w_i^{(0)} = d_i$

$$w_i^{(r)} = 1 \text{ if } g_i^{(r)} d_i < 1$$

$$w_i^{(r)} = 6 \text{ if } g_i^{(r)} d_i > 6$$

$$w_i^{(r)} = g_i^{(r)} d_i \text{ if } 1 \le w_i^{(r)} \le 6$$

It should be noted that it is not unusual for $w_i < d_i$.

NASS carries out coverage adjustment independently for each state. The coverage models are very similar for each state and differ only to the extent that the set of predictor variables used in the models differ.

For purposes of publication, all coverage adjusted weights are integerized. Weights are randomly rounded up or down to the nearest integer using a probability that is proportional to the fractional portion of the raw coverage adjusted weight.

4. Farm Count Calibration Variables

NASS uses two types of calibration variables: farm count calibration variables and commodity calibration variables. The number of calibration variables utilized in a given state was often well over 100, with some states utilizing as many as 300 or more.

Farm count variables indicate whether a responding CML farm possesses a particular characteristic. A value of 1 for the particular farm count calibration variable indicates that the farm possesses the corresponding characteristic, a 0 value indicating that it does not. The set of farm count calibration variables is chosen to represent an array of farm operation and farm operator characteristics which have diverse CML coverage rates and/or characteristics that are deemed highly sensitive. Examples of the characteristics that are represented by farm count calibration variables are:

1. Simple farm indicator (corresponding calibration variable will equal 1 for all farms).
2. Farm value of sales falls into a particular sales category (or not).
3. Various farm commodity presence indicators- such as, farm has cattle (or not), farm has vegetables (or not), farm has field crops (or not).
4. Farm operator characteristic indicators- such as, principle operator is female (or not), principle operator is Hispanic (or not), principle operator falls in a certain age group (or not).

The same basic set of farm count calibration variables is used for each state. However, the number of farms that have a specific characteristic associated with a farm count calibration variable might be too small in a given state to warrant it being considered useful for coverage adjustment and consequentially removed from the model.

5. Farm Count Calibration Targets

Through the use of farm count calibration targets, census data collected from NML farms is directly employed in the coverage adjustment process. For each farm count calibration variable, x_j, there is a corresponding farm count calibration target, t_{xj}. The farm count targets are defined at the state level and based on dual frame estimators using data from both the CML and NML. These estimators are expressed below in (3).

$$t_{xj} = \sum_{CML} d_i x_{ij} + \sum_{NML} a_i x_{ij} \quad , \qquad (3)$$

with d_i representing the nonresponse-adjusted weight for farm i in the CML domain, and a_i representing an appropriate weight for farm i in the NML domain, x_{ij} =1 if farm i possesses key characteristic j, and zero otherwise.

6. Commodity Calibration Targets

A set of commodity calibration targets is defined for each state. Most major commodities for a given state will be included in the target set. The values for these targets are based on NASS published estimates that are adjusted through the use of data obtained from other sources such as the Farm Service Agency (FSA) and commodity processors. NML census data are not used in developing commodity calibration targets.

The CML generally has good coverage of most agricultural commodity production and inventory, but the contribution by commodity data to the coverage adjustment can still be significant in some cases.

Commodity targets are generally set at the state level, but are often set at district and county levels for some items.

Some examples of commodity calibration target variables are:

1) Cattle inventory.
2) Acres of corn harvested for grain.
3) Acres of land on farm.
4) Number of broilers sold.
5) Number of acres of farm land in a county.

7. Target Tolerance Ranges

A tolerance range is computed for each calibration target. Any value the coverage adjustment procedure attains that is within the tolerance range for the target is deemed acceptable. For farm count targets, this range is based on the estimated standard error of the NML contribution to the target value of the corresponding item. Tolerance ranges for commodity targets are determined by subject matter experts.

The weight calibration program is not asked to hit all the targets exactly, but only attempts to attain values that are within the tolerance range of the targets. This is a way of relaxing the benchmark constraints and can lead to the values of more calibration variables achieving acceptable levels. Tolerance ranges also reflect that fact that most of the targets are only estimates of unknown population values, as is the case particularly

with farm count targets. Tolerance ranges are also useful if a good target value for a particular item might not be available, but a realistic range of acceptable values can be determined. (for details on how the weight calibration program operates, see Fetter, Kott, 2003).

8. Excluding and Restricting Coverage Adjustment for Specific Farm Records

It is desirable to restrict the size of the coverage adjustment, or exclude completely from coverage adjustment certain farms that have characteristics for which the CML is deemed to have very good coverage, or produce rare commodities for which calibration targets are not available.

Coverage adjustment exclusion flags are set on farm records that are to be completely excluded from coverage adjustment. Most of these excluded farms are economically very large farms. Almost all of such farms will be covered by the CML. The coverage adjustment for such farms will be set to 1 and the coverage-adjusted weight for these farms will be set equal to its nonresponse-adjusted weight (recall that nonresponse-weights seldom exceed 2, with the large majority of records receiving a nonresponse-weight close to 1).

To reduce the possibility of the over-expansion of farms producing rare commodities, such farms may receive a coverage adjustment restriction flag. This flag tells the weight calibration program to restrict the coverage adjusted weight of the farm record to the closed interval [1, 3] (for all unrestricted records, coverage adjusted weights are permitted to lie in the closed interval [1, 6]).

Figure 1 shows a typical distribution of coverage adjusted weights at the state level.

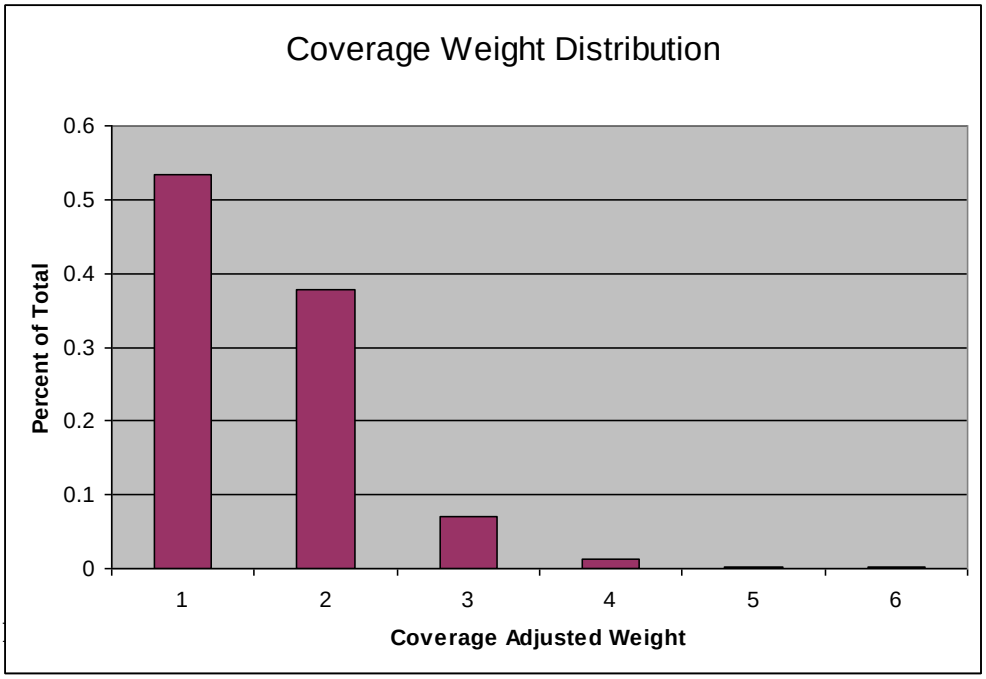

Figure 1.

9. The Coverage Adjustment Management System

Although the set of calibration variables are fairly similar for each state, the actual target values and tolerance ranges naturally vary significantly from state to state. Additionally, the types of records that are to be excluded from coverage adjustment or restricted in the amount of coverage adjustment permitted can vary significantly as well. It was therefore necessary to have a system in place that could be used to manage each state's coverage adjustment specifications. A coverage adjustment program parameter system was developed that accomplished this task.

The calibration program parameter system resides in a central database that is accessible to all field office and headquarters personnel. It is useful as a source of documentation, and can be used to interactively edit, insert, and remove calibration program parameters. It also serves as a central location from which the calibration program parameters can be extracted and input into the calibration programs.

10. Concluding Remarks

Full coverage adjustment of all published census estimates represents a recent major change to census methodology. The 2002 census was the first time that coverage adjustments were applied to all published estimates. Based on the 2002 experience, improvements were made to the methodology and applied to the 2007 census. As we look forward to the next census in 2012, there are still improvements that can be made. Some of the improvements might involve the utilization of more calibration targets being defined at the county level. Other improvements might come about through investigating the applicability of using non-linear calibration in the coverage model. Additionally, benefits might be gained through a re-evaluation of the set of calibration variables being used and the consideration of adding new variables and possibly removing existing ones. Re-considering the allowable size of the coverage weights might also lead to some improvements.

There are also computational issues that come to bear such as algorithm convergence failures, how these failures are handled (or avoided) and the amount of computer time required to compute weights for the roughly 1.5 million records being summarize in the census. Currently, a complete set of coverage adjusted weights can be calculated in about 24 hours of run time . The amount of computer time required to compute the coverage adjusted weights for a particular state is primarily linked to the number of records requiring weights, and the number of calibration targets employed. States having a large number of records and hundreds of calibration variables can require many hours to complete. Although great strides have been made in reducing the computer time required to coverage adjust the weights between 2002 and 2007, the possibility of using more county level targets could potentially increase the over-all number of calibration targets, and, consequentially, greatly increase the required run time.

References

Fetter, M. J. and P. S. Kott (2003). Developing a coverage adjustment strategy for the 2002 census of agriculture, *FCSM proceedings*.

Fetter, M.. J., and P. S. Kott (2004). Mean squared error estimation for the coverage and nonresponse adjusted census of agriculture, *JSM proceedings.*

Fetter, M. J. and P. S. Kott (2007). A strategy for estimating the number of minority operated farms, *JSM proceedings*.

Singh, A. C., and C. A. Mohl (1996). Understanding calibration estimators in survey sampling, *Survey Methodology* **22**, 107–115.