Calibration for the Census of Agriculture

Andrea Lamas[1], Kelly Toppin[1], Matthew Williams[2]
Linda J Young[1], Cliff Spiegelman[3]
[1]National Agricultural Statistics Service, USDA, 1400 Independence Ave, SW, Washington, DC 20250
[2]Department of Health and Human Services, 1 Choke Cherry Road, Rockville, MD 20857
[3]Texas A&M University, College Station, TX 77843

Abstract
The National Agricultural Statistics Service (NASS) conducts a Census of Agriculture every 5 years, in years ending in 2 and 7. For the 2012 Census of Agriculture, NASS used capture-recapture methods to adjust the Census for under-coverage, non-response, and misclassification of farms/non-farms. After these adjustments, the weights were calibrated and integerized. Calibration was conducted to ensure that state and national totals were unbiased for variables where administrative data were available. The integerization process rounded weights but did not change marginal totals. NASS researched alternative calibration methods applied to the Census. Here the constraints and limitations of those methods are discussed.

Key Words: Capture-recapture, Calibration, Non-response, Under-coverage

1. Introduction

The USDA's National Agricultural Statistics Service (NASS) conducts hundreds of surveys and prepares reports that cover every aspect of U.S. agriculture. The majority of the reports provide estimates that impact U.S. markets and price of commodities. Some examples of these include corn, soybeans, wheat and upland cotton. The largest survey that NASS conducts is its Census of Agriculture. The census provides information on characteristics of U.S. farms and ranches and people who operate them. It is used by federal, state and local governments and others who provide services to farms and rural communities. Its estimates are produced at the national, state and county levels. The estimates impact community planning, availability of operational loans and other funding, location and staffing of service centers, and farm programs and policies.

Estimates produced by the Census of Agriculture are adjusted in two ways. First the estimates are adjusted through capture-recapture. Second, the estimates are adjusted through calibration. This ensures census estimates are consistent with available information on commodity production. However, through the current calibration methodology, all targets are rarely met. Other issues also arise through this process. The purpose of this work is to discuss the constraints and limitations of the current calibration methodology and to propose an alternative methodology.

2. Census of Agriculture

NASS conducts the Census of Agriculture every five years (years ending in 2 and 7). The census is a count of U.S. farms and ranches and the people who operate them. As established by Congress in 1974, a farm is any place from which $1,000 or more of agricultural products were produced and sold or normally would have been sold during the year. During the census, data are collected on land use and ownership, operator characteristics, production practices, income and expenditures, and numerous other characteristics. The census provides the most uniform, comprehensive agricultural data for every county in the nation. It is a list-based survey; the Census Mail List (CML) is the list of all operations mailed a census questionnaire.

2.1 Census Estimates

Several sources of error are known to exist on the census of agriculture. The CML contains agricultural operations that are farms and agricultural operations that are non-farms. Some farming operations are not on the CML, due to incompleteness of the list. Because of this, there is list under-coverage on the census. Also, not all agricultural operations on the CML respond, resulting in non-response. Lastly, misclassification occurs on the census due to errors in census reporting. This occurs when, based on their response to the census questionnaire, some non-farms are classified as farms, or when some farms are classified as non-farms. To adjust for these sources of error, the census estimates are adjusted through capture-recapture.

NASS also obtains administrative data on commodity production. After the census estimates are adjusted through capture-recapture, they are then calibrated to ensure the estimate are consistent with the administrative data.

2.2 June Agricultural Survey

To adjust for errors due to under-coverage, non-response, and misclassification, using capture-recapture, two independent surveys are required. The census of agriculture is the first survey and the June Agricultural Survey (JAS) is used as the second survey. The June Agricultural Survey (JAS) has an area frame and is conducted annually. It collects information on U.S. crops, livestock, grain storage capacity and type and size of farms. Because the distribution of crops and livestock can vary widely across a state in the U.S., land is divided, in preparation for sampling, into homogeneous groups or strata, such as intensively cultivated land, urban areas and range land. The general strata definitions are similar from state to state; however, minor definitional adjustments may be made depending on the specific needs of a state. Each land-use stratum is further divided into substrata by grouping areas that are agriculturally similar. This yields greater precision for state-level estimates of individual commodities. Within each substratum, the land is divided into primary sampling units (PSUs). A sample of PSUs is selected and smaller, similar-sized segments of land are delineated within these selected PSUs. Finally, one segment is randomly selected from each selected PSU to be fully enumerated. Through in-person canvassing, field interviewers divide all of the land in the selected segments into tracts, where each tract represents a unique land-operating arrangement. Each tract is screened and classified as agricultural or non-agricultural. Non-agricultural tracts belong to one of three categories: (1) non-agricultural with potential, (2) non-agricultural with unknown potential, or (3) non-agricultural with no potential. A tract is considered agricultural if it has qualifying agricultural activity either inside or outside the segment. Otherwise, it is non-agricultural. An agricultural tract will subsequently be classified as a farm if its entire operation (land operated both inside and outside the segment) qualifies with at least $1,000 in sales or potential sales. All non-agricultural tracts and agricultural tracts with less than $1,000 in sales are classified as non-farms.

2.3 Capture-Recapture Weight

For a farm to be captured by the census, the farm must first be on the CML, respond to the census, and be classified as a farm based on the response to the questionnaire. Therefore, the probability of capture is

$$\pi_C = \pi\left(CML, responded, farm\ on\ census \middle| farm\right)$$
$$= \pi(CML|Farm)\, \pi\left(Responded \middle| CML,\ Farm\right) \pi\left(Census\ Farm \middle| CML,\ Responded, Farm\right)$$

where

c = capture.

This probability of capture accounts for under-coverage, non-response and misclassification of non-farms as farms. However, the misclassification of farms as non-farms, or correct census farm classification is not included in the probability of capture. Therefore, the probability of correct census farm classification is

$$\pi_{CCFC} = \pi(Farm|Farm\ on\ census)$$

where

CCFC = correct census farm classification.

A matched dataset of CML and JAS records is created and logistic regression models are developed for each probability. Therefore, the capture-recapture weight is:

$$\frac{\hat{\pi}_{CCFC}}{\hat{\pi}_C}$$

and the capture-recapture estimate is

$$CR = \sum_{i \in F} \frac{\hat{\pi}_{CCFC_i}}{\hat{\pi}_{C_i}}$$

where

F = set of all CML records classified as a farm based on their responses to the Census questionnaire.

3. Calibration

3.1 Census Calibration

After the capture-recapture adjustment, calibration is conducted to ensure that census estimates are consistent with administrative data on commodity production. NASS obtains information on most commodities from administrative sources or from NASS surveys of non-farm populations. Some examples are USDA Farm Service Agency program data, Agriculture Marketing Service market orders, livestock slaughter data and cotton ginning data.

The targets used in the census calibration are the commodity administrative data and 65 of the capture-recapture estimates. The 65 capture-recapture estimates are estimated for each state. They are the number of farms, land in farms, and the number of farms by the following characteristics: 8 categories of value of agricultural sales, age of farm operator, female operators, race of farm operator, Hispanic origin of the principal farm operator, 10 major commodities by their 4 sales categories and 7 farm type categories.

3.2 Current Calibration Methodology

The current calibration methodology is truncated linear calibration with weights between 1 and 6 (Fetter, 2009). Each state is calibrated separately. High priority targets are calibrated first and are treated as hard (fixed) targets. Within the set of priority targets, the target having its estimate furthest from the target value is included in calibration first. Once that target is hit, the next target with the estimate furthest from the target value is included. If a target cannot be hit, it is removed from the list for targets, and the next target with the estimate furthest from the target value is included. Once this process is complete with the high priority targets, a stepwise algorithm is used to calibrate the remaining. In the stepwise algorithm, all variables are treated with equal priority and as hard targets. However, once a target has been entered and has been hit, it is then treated as a soft target (within an interval) as other variables are entered in the stepwise algorithm. Each soft target is calibrated within a pre-specified tolerance range (generally less than 2% of the target). Output weights from calibration are to several decimals but census results are published at the integer level. Therefore weights are integerized (or rounded) to ensure all tables and breakdowns are summed to correct totals.

Not all records are treated similarly during calibration. For large and unique farms, census data collection was assumed to be complete. Weights are controlled to be one during the calibration adjustment process for these records. Specialty operations have weight restriction of the interval [1, 3]. For all other farms, calibration adjustments begin with the capture-recapture adjusted weights but are truncated to [1, 6].

With this methodology, all targets are rarely met through calibration. The fact that all variables are treated as hard targets when being entered into the stepwise algorithm constrains feasible solutions. Also, after calibration, all estimates are rounded to integers in a manner that preserves farm totals. This rounding for large farm producers becomes problematic.

3.3 Proposed Calibration Methodology

A new calibration methodology was developed. It is similar to the current methodology in that it first hits high priority targets as hard targets. However, it then includes all targets at once as soft targets using the LASSO (least absolute shrinkage and selection operator) penalty. Also, in the proposed calibration methodology, the weight restriction scheme and truncation of DSE weights input into calibration were evaluated. The LASSO methodology was run with and without weight restrictions and with DSE weights input and output from calibration between [1, 6] or [0.9, 6]. The calibration methodologies, weight restrictions schemes and weight truncations were compared based on the number of targets missed. Results for Michigan, North Carolina, and Texas were obtained using 2012 Census of Agriculture data.

4. Results and Conclusions

For all states (Michigan, North Carolina and Texas), four variations of calibration were compared. The first is the current methodology used by NASS, where truncated linear methodology is used, no weight restriction changes were made, and DSE input and output weights were between [1, 6]. The second approach uses the proposed LASSO methodology, however keeps all weight restrictions and truncations the same. The third again uses the LASSO and changes the weight restriction scheme. Instead of large and unique farms whose census data collection was assumed to be complete having a weight of one, only records who do not have a non-response, under-coverage and misclassification adjustment receive a weight of one. Lastly, the fourth approach uses the LASSO, does not change the weight restriction scheme. However, DSE input weights to calibration are changed from [1, 6] to [0.9 to 6]. Output weights of calibration are also changed from [1, 6] to [0.9 to 6]. These changes allow the input weights to calibration and output weights from calibration to be less than 1.

The results for Michigan, which has 175 targets, are in Table 1 below. The truncated linear methodology missed 9 targets. Alternatively, the LASSO methodology missed 6 targets. Changing the weight restriction scheme (only records who do not have a non-response, under-coverage and misclassification adjustment receive a weight of one) and using the LASSO methodology, still missed 6 targets. Using LASSO and allowing the DSE input weights to calibration and output weights from calibration to range from [0.9, 6] reduced the number of missed targets to 4.

Table 1: Michigan Calibration Results

Calibration Methodology	Weight Restriction Change	DSE Input Weight to Calibration	Output Weights from Calibration	Targets Missed
Truncated Linear	No	[1,6]	[1,6]	9
LASSO	No	[1,6]	[1,6]	6
LASSO	Yes	[1,6]	[1,6]	6
LASSO	No	[0.9,6]	[0.9,6]	4

North Carolina has 184 targets and the results for this state are in Table 2. The truncated linear methodology missed 4 targets. Alternatively, the LASSO methodology missed 3 targets. Changing the weight restriction scheme (to only records who do not have a non-response, under-coverage and misclassification adjustment receive a weight of one) in the LASSO methodology, reduced the number of missed targets to 1. Using LASSO, and allowing the DSE input weights to calibration and output weights from calibration to range from [0.9, 6] resulted in 3 missed targets.

Table 2: North Carolina Calibration Results

Calibration Methodology	Weight Restriction Change	DSE Input Weight to Calibration	Output Weights of Calibration	Targets Missed
Truncated Linear	No	[1,6]	[1,6]	4
LASSO	No	[1,6]	[1,6]	3
LASSO	Yes	[1,6]	[1,6]	1
LASSO	No	[0.9,6]	[0.9,6]	3

The results for Texas with 346 targets are in Table 3 below. The truncated linear methodology missed 14 targets. Alternatively, the LASSO methodology missed 12 targets. Changing the weight restriction scheme (only records who do not have a non-response, under-coverage and misclassification adjustment receive a weight of one) in the LASSO methodology, reduced the number of missed targets to 5. Using LASSO and allowing the DSE input weights to calibration and output weights from calibration to range from [0.9, 6] resulted in 11 missed targets.

Table 3: Texas Calibration Results

Calibration Methodology	Weight Restriction Change	DSE Input Weight to Calibration	Output Weights of Calibration	Targets Missed
Truncated Linear	No	[1,6]	[1,6]	14
LASSO	No	[1,6]	[1,6]	12
LASSO	Yes	[1,6]	[1,6]	5
LASSO	No	[0.9,6]	[0.9,6]	11

In conclusion, the proposed code does as well or better than the operational code. Changing the weight restriction scheme decreases the number of targets missed. Changing the input and output weights from calibration decreases the number of targets missed, but this requires changes to the methods used for rounding. Therefore, how best to round the values is the next research project.

References

Fetter, Matthew J. 2009. An Overview of Coverage Adjustment for the 2007 Census of Agriculture. Proceedings of the Government Statistics Section, JSM 2009. Pp 3228-3236.

Singh, A. C., and C. A. Mohl (1996). Understanding Calibration Estimators in Survey Sampling, Survey Methodology 22, 107–115.