SMALL AREA ESTIMATION OF COVERAGE ERROR FOR THE 1997 CENSUS OF AGRICULTURE

Bruce Eklund, National Agricultural Statistics Service
Washington, D.C. 20250

KEY WORDS: small area estimates, synthetic estimates, coverage error, indirect estimates, composite weighting.

1 Introduction

Small area estimation is a statistician's compromise due to the demand for detailed data in the absence of money. The National Agricultural Statistics Service (NASS) is using survey data to estimate net coverage error for the 1997 Census of Agriculture at the state level. The corresponding 1992 estimates were for the four U. S. regions. The 1997 survey sample sizes did not increase sufficiently to make state estimates with comparable sampling error.

State indirect estimates are created by allocating regional (multistate) survey estimates to the state level. The regional estimates are allocated to the smaller estimating units with auxiliary variables.

Composite estimators will be formed by weighting and combining the direct and indirect estimates. A simple weighting scheme was concocted. The weighting accounts for both a state's sample size in relation to its region and the absolute state sample size.

2 Importance of Coverage Estimates

Coverage studies uncover and quantify list frame problems. The Census Bureau conducted coverage studies on each census of agriculture since 1945. Though the studies were designed to measure error in farm counts, inferences will be made to adjust commodity levels and farm demographics associated with the farm count errors as well. These adjustments should produce estimates closer to "truth." Adjusted commodity estimates will be published for the first time in Volume I, the "main" publication of the census of agriculture.

This coverage evaluation will have an additional internal use. As of 1997, the former Agriculture Division of the Census Bureau is now part of NASS. The coverage evaluation will be used to help understand differences between NASS survey estimates and census numbers.

Finally, the idea of using sampling to adjust census numbers is timely. This idea is a matter of contention for the 2002 U.S. Census.

3 Terminology

Consistent terminology is still evolving. Herein, small-area estimators are domain estimators that use data from other domain to augment insufficient sample sizes. Since the domain may be neither small nor an area, the modifiers "synthetic" and "indirect" might be more descriptive. "Smoothing" is the *act* of allocating or "borrowing strength" from other domain. The term "indirect" is handy because it has the counterpart "direct," which refers to estimators based solely on the survey data from the domain in question.

4 Small Area Estimation with Indirect Estimates

State estimates of coverage error will be derived for farms and commodities associated with farms found with coverage error. Indirect domain estimates will be a part of the estimating process.

Indirect estimators have relatively small variances. They may be biased. Still, their mean squared error can be small relative to the variance of unbiased direct estimators. The user of synthetic indirect estimators assumes similarity between domain (Singh, 1994). To the extent that this assumption is violated, the synthetic estimates are biased.

5 Estimation Outline

1) Direct estimates will be calculated at the state level.

2a) Direct estimates will also be calculated for the regions, which will be uniquely defined for each commodity.

b) Census numbers will be used as auxiliary variables to prorate regional estimates to the state level. The result is the state synthetic estimate. The choice of census auxiliary variables will match the survey data variables. For example, census soybean acres for soybean acres, cotton bales for cotton bales, and census cattle numbers to prorate the number of misclassified cattle.

3) The weighting of direct state estimates and synthetic state estimates will form composite state estimates.

6 Direct Estimates of Coverage Error

List frames are inherently difficult to maintain. Farms are continually entering and going out of business. Thus, survey data are used to account for the inevitable imperfection in the census mail list.

Two surveys are used to adjust key census numbers. Area frame surveys are used to estimate farms that were not on the census mail list. A sample of defined land areas is randomly selected. Concentrated effort is exerted to interview all households in these areas regarding agricultural activities. These sampled areas are used to estimate farms and farm characteristics missing from the mail list.

The Classification Error Survey is used to reinterview census respondents. More detailed questioning is designed to find farms that were misclassified as nonfarms, nonfarms misclassified as farms, and farms duplicated on the mail list. These data are also used to adjust census farm counts and farm characteristics. See Wolfgang (1997), Census of Agriculture (1992), and Coverage Evaluation (1992) for more about the Coverage Evaluation Program.

7 Grouping Domains

Since homogeneity among domain is important regarding bias, and since this homogeneity varies by commodity, separate regions (state groupings) are defined for each commodity. Several criteria are used to define regions according to the commodity to be estimated. Average production per farm is an important measure of farm homogeneity. Also considered is the percentage of farms with the commodity in question that are small. Three of the four types of coverage error measured occur more frequently among small farms.

Commodity experts suggested groupings based on production similarities. Dot maps show contiguous production areas. Also important is the number of farms with the commodity in question because states must be grouped to achieve reasonable region-effective sample sizes for the commodity in question. Some of these criteria conflict and necessitate compromise.

8 Simplifying Notation

The direct estimate of net coverage error (NCE) of the farm count is the estimated farms not on the mail list (NML) plus farms misclassified as nonfarms (incorrectly classified undercount, ICU) minus nonfarms misclassified as farms (incorrectly classified overcount, ICO) minus farms duplicated (DUP) on the mail list.

$$NCE = NML + ICU - ICO - DUP$$

Differences in computations of these components are outside the scope of this paper. For simplicity, the NML will be ignored. Although the NML component is calculated differently and separately from the other three, smoothing of the NML data is the same. For now:

$$NCE = ICU - ICO - DUP$$

Also ignored herein will be the difference in the expansion factor of the DUP and the two misclassification components.

Commodities associated with coverage error will be estimated. Let y be a commodity of interest observed from the survey data and let I_{1i}, I_{2i}, and I_{3i} be indicator variables; one if ICU, ICO, or DUP for record i is respectively positive, zero otherwise.

A component being positive means that there was the corresponding coverage error associated with that record. For simplicity, we will redefine y so that:

$$y_i = y_i I_{1i} - y_i I_{2i} - y_i I_{3i}$$

The direct state level estimate is now simply

$$\hat{NCE}_{ds} = \left(\frac{N_s}{n_s}\right) \sum_{i=1}^{ns} y_i$$

where,

NCE = Net Coverage Error
d = direct
N = the number of records on the mail list
n = the number of CES responses
y = observed quantity pertaining to commodity of interest. This could be a count of farms with that commodity, acres, or a production measure.
s = state

For the indirect state level estimate, first find the direct region estimate:

$$\hat{NCE}_{dr} = \left(\frac{N_r}{n_r}\right) \sum_{i=1}^{nr} y_i$$

where r = region, a group of states defined uniquely for the commodity of interest

The state indirect estimate of net coverage error is

$$\hat{NCE}_{Is} = \frac{X_s}{X_r} \hat{NCE}_{dr}$$

where,

I = indirect
X = auxiliary variable: Census number for corresponding item of interest

If the item of interest is soybean bushels, X is the census soybean bushels. If it is cotton acres, X is the census cotton acreage.

9 Composite Estimator

"The investigator should realize that ingenuity in putting together methods of sample selection and estimation is of the greatest importance in arriving at efficient designs for particular jobs" (Hansen, 1953). This job demands complex estimators (simplified here) for multiple commodities for each state. These data must be summarized not only in more detail but in an accelerated schedule unprecedented for the surveys involved. Thus, a simple approach to weighting the many composite estimators is desirable.

The direct and indirect state estimates will be combined using weights, w_d and w_I, so that they sum to one. The composite net coverage error, NCE_c, is (suppressing the subscript s for convenience):

$$NCE_c = w_d * NCE_d + w_I * NCE_I$$

The weights will be selected considering two concepts:

Ideally, with sufficient sample size, w_d would be one. Coverage error is closely related to the state's list frame, which each state office maintains. One would like, to the extent possible, to estimate for state s with state s data.

Realistically, with insufficient sample size, strength will be borrowed from similar states. The degree of borrowing (how close w_I is to one) will be in proportion to the state's sample size to that of the region.

In short, w_d will be close to one if the state sample size is close to the region sample size. However, if the state sample size is sufficiently large in absolute terms, then borrowing strength is no longer needed.

To capture both of these needs, we weigh the direct estimator as:

$$w_d = \min\{1, n_s/n_r + n_s/k\} \quad \text{and } w_I = 1 - w_d$$

where

k = a predefined constant

n_s = the state effective sample size (samples with the commodity in question)

n_r = the regional effective sample size (samples with the commodity in question)

The constant k is defined by deciding what sample size is sufficient to eschew smoothing.

10 Estimation Consistency

The practitioner may face interesting options regarding biased domain estimates. Obviously the domain estimates should sum to a published total. The practitioner could sum the biased domain estimators and publish the sum, or the practitioner could scale the domain to sum to the unbiased (direct) total estimate.

Some more interesting choices surface. For example, coverage error for both a common and rare event must be measured: state level farms operated by whites and those operated by other races. The former may not require data smoothing. One could scale both estimates to sum to all farms found with coverage error, or one could estimate coverage error for other races and than find coverage error for white operated farms by subtracting other races coverage error from the coverage error of total farms.

Now consider 1) cattle farms with coverage error and 2) total head of cattle on farms with coverage error. One may argue that more confidence be put in the former because the "effective sample size" of the latter comes from only those cattle farms found with coverage error. Hence, one could first use a small area estimation technique to simply estimate an *average* head per farm for farms with cattle that were found in coverage error. Then separately estimate cattle farms with coverage error. Head of cattle associated with farms with coverage error is then the product of (cattle farms)X(head per farm). This approach allows the cattle farm estimate to drive both estimates. It circumvents an undesirable possibility: the independent indirect estimation of cattle farms and total cattle head could move these two estimates in opposite directions.

11 Measure of Error

The mean square error of the indirect state coverage error (MSE_{Is}) is:

$$MSE_{Is} = (NCE_{Is} - truth)^2 - v(NCE_{ds})$$

where,

truth = the true state value

One can estimate the first term by hoping that the direct state estimate is a reasonable estimate of the truth:

$$Bias^2 = \left(NCE_{Is} - NCE_{ds} \right)^2$$

But if the direct state estimate came from a sufficient sample to produce a reasonable direct state estimate, there would be no need for indirect estimating. With insufficient state sample sizes, the estimated mean square error will be unstable.

An approach offered by Kott (1989) is to let the data specify a model to approximate the MSEs.

A specific model suggested upon the author's request, follows. Let's revert to more general notation with "y" rather than "NCE":

$$mse_{Is} = - (1 - 2X_s/X_R)v(y_{ds}) + (y_{Is} - y_{ds})^2$$

which accounts for the covariance of y_{Is} and y_{ds}

After finding the relative mean square error:

$$relmse_{Is} = mse_{Is}/Y_{is}^2$$

The relative mean square error can be modeled:

$$relmse_{Is} = \alpha - \beta X_s + \epsilon_s$$

This model was tested empirically. The model may work well with many observations, but grouping states to form homogeneous regions can result in few observations (states). With few observations, the term $(1 - 2X_s/X_R)$ can be large (even positive, if a state contributes more than half of the region total). Also, the term $(y_{Is} - y_{ds})^2$ can vary widely.

If one cannot approximate the mean square error with confidence, one might report the standard error of the direct estimate and clearly explain to data users that the reported numbers do not account for bias. If the application was a good candidate for small area estimation, then smoothing should have reduced the variance more than it introduced bias, and thus the standard error of the direct estimate is usually a conservative estimator for the root mean squared error of the composite estimate.

12 Highlights

There is a demand for commodity estimates with insufficient data: state estimates of coverage error to census numbers.

Even assuming that states can be grouped with homogeneity for a commodity, doing so is a necessary evil since the state lists vary in quality. Thus a composite weighting scheme was devised to more highly weigh direct estimates to the extent that sample sizes permit.

A method to model the mean square error was suggested to the author. This method may serve other cases well, but there are insufficient observations per region here.

13 References

1992 Census of Agriculture, Volume 1, Summary and State Data (in 51 parts), Commerce/Census, Washington, D.C.

1992 Census of Agriculture, Volume 2, Part 2, Coverage Evaluation, Commerce/Census, Washington, D.C.

Hansen, M. H., Hurwitz, W. N., & Madow, W.G. 1953. Sample Survey Methods and Theory. John Wiley & Sons, New York.

Kott, P.S., (June 1989), "Robust Small Domain Estimation Using Random Effects Modeling," *Survey Methodology,* 15 (1), Statistics Canada, pp. 3-12.

Singh, M.P., Gambino, J. & Mantel, H.J. (1994), "Issues and Strategies for Small Area Data," *Survey Methodology*, 20 (1), Statistics Canada, pp. 3-14.

Wolfgang, G. (1997), "Net Coverage Error in the 1997 Census of Agriculture," *Proceedings of the Section on Survey Research Methods*. pp. 407-412.