A PLAN FOR PROVIDING CONSISTENT EVALUATIONS OF AREA SAMPLING FRAMES AMONG FIELD OFFICES

Jeffrey K. Geuder, Robert D. Tortora, USDA

Introduction

The Statistical Reporting Service (SRS) of the U.S. Department of Agriculture relies heavily on area sampling frame surveys to estimate a variety of agricultural characteristics. A separate sampling frame is maintained in each of 44 State Statistical Offices (SSO's). The characteristics of each frame (stratification, sample size, etc.) are dictated by the needs and priorities of the individual state and those of the agency as a whole. The construction of these frames requires a major investment of time and money, and the goal is to construct sampling frames that can be used efficiently for as long as twenty years or more. Over the course of its use, the sampling frame should be monitored to determine whether changes in land use or agricultural practices have caused it to become inefficient.

The responsibility for evaluating the sampling frames logically rests on statisticians in three units in SRS: the Sampling Frame Development Section (SFDS), the Methods Staff, and the State Statistical Offices. The SFDS is the unit responsible for constructing the frames, and statisticians in this unit work closely with statisticians in Methods Staff on the sample design and allocation. Statisticians in the SSO's use the area frame survey data to estimate a wide variety of agricultural characteristics for their states. They also have first-hand knowledge of the state's agriculture and have ready access to much of the data that is required for an adequate evaluation of the frame.

There are two primary reasons for providing the SSO's with a suggested review procedure. First, it saves time for the statisticians in the SSO's since they do not have to develop their own analysis plans. Second, the guidelines should provide consistent evaluations among SSO's. In the future, the decision as to when to construct a new frame for a state will depend, for the most part, on these evaluations. This decision should be an objective one, based on valid statistical analyses of survey data.

The suggested review procedure is based on the experience the SFDS has gained by evaluating new frames constructed by SRS in recent years. The procedure is intended to help the statistician detect possible nonsampling errors associated with the frame and to evaluate the effect of the sampling errors associated with the estimates. In situations in which problems are detected in the frame, the statistician is encouraged to work with statisticians in both the SFDS and Methods Staff to do a more complete analysis of the frame.

The June Enumerative Survey

The primary source of data for estimating agricultural characteristics is the June Enumerative Survey (JES), a large-scale survey, which in most states is of a multiple frame design. The area frame in these states is supplemented by a list frame of known producers of a particular commodity. That part of the population which overlaps both frames is "removed" from the area frame and is accounted for solely by the list frame sample. The estimates from the two samples are then combined to produce a single estimate for the state. The proportion of the state estimate accounted for by the area frame, then, varies by commodity. Some crop estimates are made solely from the area frame sample, so maintaining the quality of the frame is critical.

The area sampling frames used by SRS are stratified according to land use. The primary classification is agricultural land versus nonagricultural land, with further stratification in each category. The sampling unit is called a segment, which varies in size depending on the land use stratum. The reporting units within a segment are called tracts, which are classified as agricultural or nonagricultural. The agricultural tracts that are enumerated in the JES comprise a population of sampling units for other "follow-on" surveys.

The JES sample consists of several replications in each land use stratum. The replications are used in a rotation plan, in which approximately 80 percent of the sample is retained each year, and the other 20 percent is replaced. This provides a sufficient level of comparability each year while limiting respondent burden.

Sources of Data for the Evaluations

The general approach to evaluating the state's area sampling frame is to compare current year survey results to results from previous survey years. In making the comparisons, the statistician must be sure to consider any recent changes in the state's agriculture. The following sources of data are used in the review:

1. Current year survey data, consisting of the JES computer summary and the completed questionnaires;

2. An Area Frame Data Base (AFDB) containing previous years JES data, Crop Reporting Board (CRB) estimates, and other information describing the state's area frame;

3. In-house data, such as the state's annual Agricultural Statistics summary, the official estimates of commodities not included in the AFDB, records of "problem segments", and any other data that is related to the area frame and the estimates made from it.

Survey Items Included in the Evaluation

The statistician first identifies the major commodities in the state. Major commodities are those which: are important in the state's estimating program; and/or are important in the national estimating program; and can be estimated using an area sampling frame.

A commodity for which the JES area frame estimate provides the primary indication for the official state estimate would meet the first criterion. A commodity for which an individual state estimate accounts for a significant proportion of the national estimate would meet the second criterion. It is difficult to generalize the third criterion. "Specialty" crops (such as vegetables, tobacco, and sugarcane) are usually considered "rare items" and cannot be accurately estimated using an area sampling frame alone. In some states, when the sampling frame includes "crop specific" strata, the area frame estimate may be sufficiently precise to be usable.

In most states the following commodities are included in the analysis:

Winter wheat acres	Number of Cattle
Soybeans acres	Number of Hogs
Corn acres	

The statistician should also include in the analysis those survey items that can be used as indications of the quality of the frame. These would include the following variables, either at the segment level or at the stratum level:

Average segment size
Cropland
Idle land
Percentage of cultivation
Total tracts
Resident agricultural tracts
Nonresident agricultural tracts
Nonagricultural tracts

Overview of the Review Procedure

The suggested review procedure is focused on four aspects of the area frame:

1. the direct expansion estimates of the major commodities;
2. the precision of the estimates;
3. the number (or proportion) of agricultural tracts included in the sample; and
4. the number of "problem segments" in the sample. (A problem segment is one which presents severe enumeration problems, which can sometimes be corrected.)

In all four areas, current year survey results are compared with results from previous years. For example, the direct expansion estimate for a particular commodity for the current year is compared to the direct expansion estimates for recent years. If there is a change, the statistician must determine whether the change is consistent with current economic or market conditions in the state. If a change cannot be "explained" in terms of current conditions, there may be some problem in the area frame and a more detailed analysis would be in order. This approach (current year versus previous years) is followed for the other three areas as well.

The decision diagram in Figure 1 describes the review procedure. As shown in the diagram, a "possible problem" can be detected in each of the four areas mentioned above. Each time a possible problem is detected, a "flag" is raised, and the review continues until all major commodities, the proportion of agricultural tracts, and the number of problem segments have all been reviewed. If a possible problem is detected in one or more areas, the statistician can access the Area Frame Analysis Package (AFAP). This package of Statistical Analysis System (SAS) programs was developed by SRS specifically for analyzing area sampling frame survey data. The programs produce output to be used in a more detailed analysis attempting to locate the sources of the possible problems.

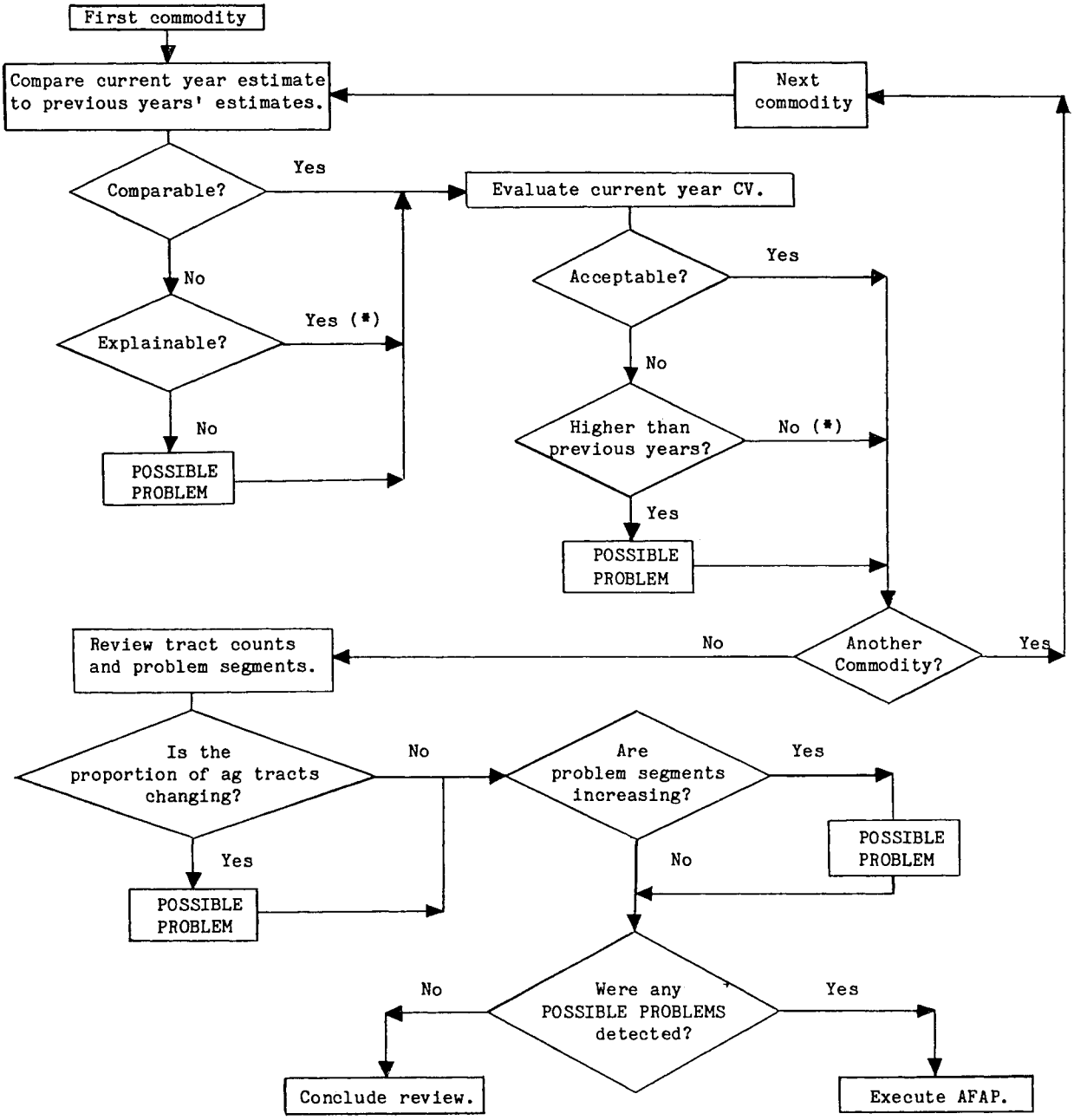
Review of Current Year Estimates

This section gives a detailed description of the review procedure shown in the decision diagram of Figure 1. The first major commodity identified earlier is evaluated as follows:

The current year direct expansion estimate is compared to the estimates for previous years. A rough rule-of-thumb used in this comparison is to construct confidence intervals of one standard error around the current year estimate and the previous estimate. The two intervals should overlap.

If the estimates are comparable, the Coefficient of Variation (CV) of the current year estimate is evaluated (see below).

If the current year estimate is not comparable, the statistician must determine whether the change in the level of the estimate can be "explained" by factors not associated with the sampling frame, such as changes in market or planting conditions.



(*) Although at these points, a large change in an estimate is explainable or a high CV is comparable to previous years, you may want to execute the AFAP for a more detailed analysis of the particular estimate and its variance.

Figure 1 - Decision diagram of the area frame review procedure.

If the change can be explained, the CV of the current year estimate is evaluated (see below). (Although a change in an estimate is "explainable", the statistician should execute the AFAP to evaluate how the frame is affected by the current conditions.)

If a change in an estimate can not be "explained", there is a possible problem in the frame. The commodity being reviewed should be included in the analysis when the AFAP is executed.

Any outliers (especially large reported values) are evaluated. The replications that have rotated into the sample for the first time in the current year are reviewed. The contribution to the total state estimate of each rotation group is evaluated.

Review of the Precision of the Estimates

The coefficient of variation of each major commodity estimate should be at an acceptable level. As a rough rule-of-thumb, the CV of a major commodity should be less than 8% to 10%.

If the CV is acceptable, the estimate of the next major commodity is reviewed. If all commodities have been reviewed, the problem segment records and the proportion of agricultural tracts in the sample are reviewed (see below).

If the CV is not acceptable, it is compared to the CV's obtained in previous years. If the CV hasn't changed from previous years, the lack of precision is not new to the frame in the current survey year. The next commodity should be reviewed, or if all commodities have been reviewed, problem segments and agricultural tracts are reviewed (see below).

If the CV is higher in the current year than in previous years, there is a possible problem in the frame. The commodity being reviewed should be included in the analysis when the AFAP is executed.

Outliers are evaluated for validity (there could be a serious reporting error in the data causing the high variance). Estimates for the various rotation groups are compared. The ESTIMATES program can be used to identify whether a particular stratum is considerably more variable than the others. The CROPDIST program shows the distribution of reported values of a commodity in the sample segments.

The next major commodity is then reviewed, or if all have been reviewed, problem segments and agricultural tracts are reviewed (see below).

Review of Tract Counts and Problem Segments

The proportion of agricultural tracts in each stratum in the current survey year is compared to the proportion in previous years. The number and/or proportion of agricultural tracts in each stratum should be about the same from year to year.

If the proportion of agricultural tracts has remained relatively stable, any problem segments in the current year are reviewed (see below).

If there is a trend, either increasing or decreasing, in the proportion of agricultural tracts, there is a possible problem, and the AFAP should be executed.

SSO records pertaining to problem segments are then reviewed. A few problem segments are to be expected, and are not, in themselves, an indication of problems in the frame. If the frequency of problem segments has not increased from previous years, the statistician should either continue with the AFAP review, if there were any indications of possible problems, or conclude the review.

If the number of problem segments is increasing, there is a possible problem, and the AFAP should be executed.

The Area Frame Analysis Package (AFAP)

If the review procedure outlined above indicates possible errors or problems in current year data, the statistician uses the AFAP to try to isolate the source of the problems. This section describes the analysis package.

GETOSAS program

This program converts the raw data to a SAS data set to be used in the remaining analysis programs. It provides a listing of every area frame segment and its associated survey data. These include: identifiers such as county code, land use stratum, paper stratum, and replication; reported data such as total acres in the segment, agricultural acres in the segment, nonagricultural acres, corn acres planted, wheat acres planted, number of head of cattle in the tract, etc; and other "summary" type data such as target segment size, expansion factor, and the year the segment rotated into the sample and the year it will rotate out.

FRAMECHECK program

This program indicates the proportion of segments in a stratum which conform to the stratum definition. (Most land use stratum definitions contain a specified range of percentage cultivated, such as 50% or more, or less than 15%, etc.) It indicates whether the distribution of percentage of cultivated land is clustered toward either extreme. That is, if the specified range for the stratum is 50% or more, whether there are more segments near 50% cultivated than near 100% cultivated.

The program provides a list of the segments that do not conform to the definition for the stratum. Extreme cases, such as a segment in a nonagricultural stratum that is 50% cultivated, or a segment in an "intensively cultivated" stratum that has no cultivation at all, are reviewed. In some cases, these will be the result of land use changes since the frame was constructed. In other cases, the lack of usable boundaries may have prevented the SFDS from stratifying a small area into a more suitable stratum. In a few cases, this listing may isolate an enumeration error.

The program compares the reported acreage for each segment with the digitized acreage, and the digitized acreage with the target segment size for segments in a given stratum, and lists the segments which have possible errors in size:

> Reported acreage differs from digitized acreage by more than 10 percent;
>
> Digitized acreage differs from target segment size by more than 25 percent;

These segments should be reviewed to verify the value of reported acreage. This is where enumeration errors that slip through the survey edit procedures will show up.

The program produces descriptive statistics by land use stratum. Average, minimum, and maximum reported values for segment sizes and number of tracts are reviewed.

ESTIMATES program

This program computes the direct expansion estimates of 5 commodities specified by the user along with two estimates of total acres - based on reported acres and based on digitized acres. Estimates, standard errors, expansion factors, CV's, and ranges of expanded values are printed out for review for each paper stratum within a land use stratum. This output will reveal situations where a land use stratum contributes a disproportionate amount of variance to the state level estimate.

Estimates and CV's are also computed for each replication within a land use stratum. Since each replication represents a random sample from the stratum as a whole, the estimates by replication for a particular commodity should be comparable. Large differences among replications lead to larger variances in the state estimate. If there has been a significant change in the estimate of a particular commodity, particular attention is given to the replications that rotated into the sample in the current survey year.

CROPDIST program

This program produces a summary (at the stratum level) of average, minimum, and maximum reported values for 5 commodities specified by the user. It produces frequency tables showing the distribution of reported values for the 5 commodities and charts that show the proportion of the state estimate accounted for by segments in the various rotation groups.

Summary

The review procedure described in this paper is aimed at evaluating four aspects of the area sampling frames used by SRS. The level and precision of the direct expansion estimates are evaluated for consistency (with previous years) and acceptability. The number of agricultural tracts is monitored, since these are the sampling units that are subsampled for other surveys. Finally, the frequency of "problem segments" is evaluated as an indication of frame deterioration.

It is hoped that by providing a review plan to the 44 field offices, SRS can obtain consistent evaluations to be used in determining when to update or replace the area sampling frames in individual states.