125 YEARS OF AGRICULTURAL ESTIMATES

FREDERIC A. VOGEL
USDA - NASS - SSD
AUGUST 1988

125 YEARS OF AGRICULTURAL ESTIMATES

I. INTRODUCTION

On July 10, 1863, the U.S. Department of Agriculture initiated monthly crop reports on the condition of crops in 21 States loyal to the Union, plus the Nebraska Territory. These early reports were based on collected data that were subject to various biases of judgment and sampling. The returns could not be taken at face value and the estimation process was crude and subjective. The ideas of probability sampling and accompanying estimation procedures had not yet been born.

Benchmark estimates were provided by the decennial censuses. Forecasts and estimates for the intervening years were based upon farmers indicating a percentage change from the preceding year. The published estimates were based on the subjective evaluation of voluntary reponses from farmers and an ingenious use of data from other independent sources. For example, administrative data such as carlot shipments of fruit and vegetables, receipts at mills and elevators, sales of livestock, etc., were used to revise the preliminary estimates. Past comparisons between data reported by farmers and final revised estimates became an increasingly important basis for interpreting and converting current reports from farmers into estimates.

Thus, the early forecasts and estimates were primarily based on the art of subjectively evaluating survey data, interpreting how survey data fit with knowledge of current weather and marketing trends, and anticipating how the survey data might later match up to administrative data. These early estimates, however, quickly became known for their general accuracy and had an influence on the markets. As the markets became more sensitive to the reports, there was an increasing need to make the estimates and forecasts more accurate. Therefore, from the beginning, improvements in statistical methodology were being continually sought.

The use of improved estimating procedures followed the development of statistical theory in general. The introduction of regression techniques and the later concept of probability sampling were milestone events in the development of estimation procedures.

As the estimation procedures developed over time, using modern science and statistical theory, the art form of arriving at the official estimates has remained essentially unchanged. A subjective apppraisal of the results of several data collection activities and administrative data are used to produce the official estimates. Charts are used to "read" current survey indications to evaluate their historic performance against administrative data. Balance sheets are used in the estimation process to compare survey estimates of crop and livestock inventories with data on grain utilization or livestock slaughtered. Official estimates will depart from survey indications, if necessary, to maintain a reasonable balance with the administrative data. This is becoming a serious problem as the probability surveys are strengthened, yet do not always show results that agree with administrative data.

The following sections will trace the evolution of the estimating procedures. The summary will revisit the dilemma between the use of administrative and survey data and discuss how it can be resolved.

II. IN THE BEGINNING - 1863

On the tenth day of each month, (May through October) a circular was mailed to a corps of 2,000 crop correspondents, whose names came from members of Congress. The questions related to two matters: The average amount sown in 1863 compared with 1862 and the current appearance of the crop. The correspondents were asked to report for their locality rather than their own farms to ensure a greater geographic coverage. For each crop, numerical answers were given with 10 representing an average of the amount of area sown making each number above or below 10 represent one-tenth of an increase or decrease. The number 10 was also used to represent an average appearance or condition of the crop. The assumption was that farmers would be knowledgeable about their locality and could report whether acreage was increasing or decreasing and whether crop conditions as affected by weather, insects, disease, etc., were above or below average. The following table was extracted from the July 1863 report.

May 1863 Report - Corn

	Average Amount of Corn Sown Compared with 1862	Appearance of Crop At This Date
Connecticut	10	11
Delaware	12	9
Illinois	11	9
Indiana	10	10
Iowa	12	11
Kansas	10	11
Kentucky	8	10
Maine	9	10
Maryland	10	8
Massachusetts	10	9
Michigan	10	10
Minnesota	13	10
Missouri	11	10
New Hampshire	9	10
New Jersey	11	10
New York	10	10
Ohio	11	10
Pennsylvania	11	9
Rhode Island	10	10
Vermont	10	11
Wisconsin	11	10
Nebraska Territory	8	10
GENERAL AVERAGE	**10 1/9**	**9 1/2**

These averages were basically simple straight averages. Some early analysis discussed the weighted average vs. the straight average.

By 1866, annual reports were initiated that included estimates of acreage, yield per acre and production of important crops, and numbers of livestock on farms on January 1. In general, the estimates through the 19th century continued to be linked to the decennial census with correspondents reporting on their viewpoint of year-to-year changes in their locality.

During the 19th century, primary estimation efforts went into enlarging the number of voluntary crop reporters. In 1882, State Agents were appointed in each State to work on a part-time basis and to build up the list of crop correspondents who would report directly to Washington, D.C. By 1914, full-time agricultural statisticians had been appointed in nearly every State. These State Statisticians began developing their own lists of farm reporters who reported to the State Statisticians. Meanwhile, the Washington, D.C. lists of correspondents were also maintained. This meant that for each survey, both lists received the same inquiry with the result from each list used as a check against the other.

Statisticians in the headquarters office had several sources of information to use to establish the official estimates. These included:

- Survey results from the headquarters list
- Survey results from each State list
- The State Statisticians' interpretation of the results of the State Survey

The process of reconciling all information into official estimates led to the creation of the Crop Reporting Board in 1905. The chief statistician would invite two headquarters statisticians and two State Statisticians to sit with him as a committee to review the data and make the final estimates. This was a subjective process requiring thorough knowledge of the items being estimated and of how the survey data would later relate to administrative data. The Chairman of the Crop Reporting Board had the full authority to "set" the estimate at the point he deemed to best represent the current situation.

III. THE 20TH CENTURY BEFORE PROBABILITY SAMPLING

In the absence of probability sampling theory much effort went into improving estimating procedures to measure crop acreages and to forecast crop yields. Although the basic objective was to measure crop production, forecasts of the crop production prior to harvest created the most interest. To forecast production during the growing season, two components are used -- estimates of acreage to be harvested and forecasts of probable yield. These components are discussed below and in a chronological order.

Par Method

In 1912, the "Par Method" was adopted to translate farmer reported crop condition values into a probable yield per acre. The par method to forecast yield ($\bar{y}$) consisted of the following components:

$$\bar{y} = \frac{C \times Ym}{Cm} \quad \text{where}$$

Cm = The previous 10-year average condition.
Ym = The previous 10-year average yield per acre.
C = Current condition for a given month.

The forecasting model was simply a line passing through the origin and (c, $\bar{y}$). A separate par was established for each State, Crop, and Month. In actual practice, subjective modification of the pars was considered necessary to remove the effects of atypical conditions. To aid in these adjustments, 100 percent equivalent yields were computed for each month of each year and 5 and 10-year moving averages were computed to identify unusual situations or trends.

Regression Techniques

The development of simple graphic solutions for regression and correlation was a major breakthrough as a practical means to forecast crop yields. Data for a sufficient number of years had been accumulated so final revised estimates of yields could be plotted against averages of reports from farmers.

$\bar{y}$ = Final revised yield
c = Condition for given month
$\hat{y}$ = $a + bc$

The regression techniques provided a consistent method to translate survey data into estimates while adjusting for persistent bias in the data caused by the purposive sampling procedures. This method quickly replaced the par method and was adopted rapidly.

Mathematical methods were not used to fit the regression lines. Instead, graphical methods were used to fit lines freehand because:

-- The method was not limited to linear relationships.
-- Years that fall "off the line" could be studied separately.

This provided the agricultural statistician some flexibility in determining the official estimate.

Beginning in 1926, farmers were also asked to report a probable yield on their farms on the inquiry used for the last forecast of the season. The probable yields were also plotted graphically to arrive at the official estimates.

The following discussions describe early attempts to estimate the acreage to be harvested:

Ratio Relative

In the early days, farmers were asked to report their judgment of the annual percentage change in crop acreages in their locality. Starting in 1888, farmers were asked to report acreages on their individual farms. By 1912, this method had completely replaced the judgment inquiry. The change in acreage computed as a percentage of the previous year was multiplied by the previous year's estimate to obtain the current estimate.

While it was considered to be a significant improvement, this method was subject to a serious bias caused by the selectivity of the sample. In an effort to make an allowance for this bias, a relative indication of the acreage was developed in 1922. This indication became known as the ratio relative and contained the following components:

R_1 = Sample ratio of the acreage of a given crop to the acreage of all land in farms (or crops) for the current year.

R_2 = Sample ratio of the items for the previous year.

$\hat{y}$ = (R_1/R_2)*Previous year's acres in given crop.

The belief was that this ratio held the bias resulting from the purposive sampling constant from one year to the next. A reported limitation was the extreme variability in the acreage ratios between the sample units. This was countered by increasing sample sizes and weighting sample results by size of farm.

In 1928, matched sample units reporting in both years were used to compute the ratio relative. This reduced the influence of the variability between sample units. When looking back at the ratio relative estimator from a current perspective, one is compelled to examine the estimate of Rel-variance (also assuming probability sampling).

$$CV^2(\hat{y}) = CV^2(R_1) + CV^2(R_2) - 2\ COV\ (R_1R_2)$$

This quickly shows why using matching reports improved the ratio relative estimator. However, this did not solve the problem because by using matching reports, farms going into or out of production of a particular crop were not properly represented. Therefore, statisticians continued their efforts in searching for a more objective method of gathering and summarizing survey data.

Pole Count

Some statisticians would travel a defined route on the rural roads and record the number of telephone or telegraph poles opposite fields planted to each crop. The relative change in the pole count for each crop from year to year provided a measure of the change in crop acreage.

Crop Meter

A more refined method of estimating acreage was developed by the Mississippi Agricultural Statistician. A "crop meter" was developed and attached to an automobile speedometer to measure the linear frontage of crops along a specified route. The same routes were covered each year. This made possible a direct comparison of the number of feet in various crops along identical routes for the current year and the previous year.

Objective Measurements of Yield Using Route Sampling

Some early work was done to use objective methods to replace the practice of relying on grower reported yields. In 1925, a North Carolina statistician submitted a plan for counting the number of cotton plants, bolls, etc., in field plots consisting of 15 feet in a row of cotton. One aspect missing from this early work was an objective random method of sampling fields to remove the selectivity bias. A significant attempt in 1939 and 1940 was to select wheat fields at random along a specified route using the crop meter. From Texas to North Dakota, samples of grain from the selected fields were obtained for computing yield and quality estimates.

Dilemma of Non-Probability Surveys

Because of the selective/purposive nature of the surveys, the determination of the "official" estimates relied heavily upon a subjective appraisal of the survey data as plotted on charts and a reconciliation with whatever supplemental data were available.

In the 1930's, demands for more accurate data rapidly increased. The depression, the "Dust Bowl", Agricultural Adjustment Act programs, and a rapid change in farming practices challenged the traditional estimating procedures. In 1938, a cooperative research program was initiated with the Statistical Laboratory at Iowa State University to develop theory of sampling and estimation to deal with these challenges. Reliable methods that were not dependent on historical relationships as bases were needed for estimation -- especially for single-time surveys or periodic surveys.

IV. THE 20TH CENTURY AFTER PROBABILITY SAMPLING

A milestone in the evolution of statistical methodology was the development of the master sample of agriculture. This was a cooperative project involving Iowa State University, the U. S. Department of Agriculture, and the Bureau of the Census. This area sampling frame demonstrated the advantages of probability sampling.

A difficulty was that with this improved sampling, estimating methodology was considerably more expensive than using the voluntary mail responses of farm operators. Thus, the national sample was only used periodically for generally single-time surveys. It was not until 1961 that Congress appropriated funds allowing the implementation of annual area frame sample surveys.

During the 1950's, however, some research had been conducted to evaluate area frame estimating procedures. These will be discussed in later sections. This period also saw a rapid change in the structure of agriculture. Farms became more specialized and much larger. This introduced more variability that could be handled by the master sample only by increasing sample sizes. The situation that was occurring can best be explained by the following relationship:

$$CV^2(y) = CV^2(\bar{y}_p) + \frac{(1-P)}{np} \quad \text{where}$$

$\bar{y}_p$ = Average of sample units having the characteristic being measured.

P = Proportion of sample units having the characteristics.

The proportion of farms having livestock was decreasing rapidly during this period. The variation in size of the farms with livestock also had increased dramatically. The combination of these two factors meant that either resources for an extremely large area frame sample would be needed or alternative sampling frames were needed. In the early 1960's, Dr. H.O. Hartley at Iowa State University was approached about this problem. The result was his 1962 "landmark" paper laying out the basic theory of multiple frame sampling and estimation which involved the joint use of area and list sampling frames. Basically, a list frame of unusually large livestock operators would be used along with the area frame which would be sampled to estimate for the incompleteness of the list sample. This basic methodology has survived the test of time. Considerable changes have been made in sampling methodologies within sampling frames and the content of the surveys , but the early fundamental Hartley estimators still are the backbone of the estimating procedures for major crop and livestock estimates. The following paragraphs briefly describe the estimators being used. The bibliography contains an extensive set of references citing research on the area and multiple frame estimators. A brief discussion of the different estimators follows. The most thorough discussion is in Nealon and Cotter (1987).

Area Frame Estimators - The sampling unit for the area sample frame is a segment of land -- usually identified on an aerial photograph for enumeration. During the frame development process, the segment boundaries are determined without knowledge of farm or field boundaries. Therefore, an early (and continuing) difficulty was how to associate farms with sample segments during data collection. Three methods have evolved which are both referred to as methods of association and as estimators.

- Farm (Open): The criteria for determining whether a farm is in the sample or not is whether its headquarters are located within the boundaries of the sample segment. This was the method used at the inception of the use of the master sample.

- Tract (Closed): This concept was first tried in 1954. The tract estimator is based on a rigorous accounting of all land, livestock, crops, etc., within the segment boundaries regardless of what part of a farm may be located within the boundaries of the segment. The method offered a significant reduction in both sample and nonsampling errors over the farm method. The difficulty was that some types of information, such as economic, could only be reported on a whole-farm basis. This led to the development of the weighted procedure in the late 1960's.

- Weighted: In this approach, data are obtained on a whole-farm basis for each farm with a portion of its land inside a sample segment. The whole farm data are prorated to the segment based on the proporation of each farm's land that is inside the segment. This estimator provided the advantage of a smaller sampling error than either the farm or tract procedures. On the minus side, data collection costs increased 15-20 percent, and intractable nonsampling errors are associated with determining the weights.

Ratio - The area frame sample was designed so that 50-80 percent of the segments were in the sample from year to year. This allowed the computation of the usual ratio estimators. The commodity statisticians, in their desire to move away from the nonprobability survey ratio estimates and any reliance on previous base data, did not give much consideration to the ratio estimates.

Multiple-Frame Estimator - As farm size continued to increase and as farms became more specialized, the efficiency of the area frame design was pressed to the limit. The presence or absence of a single large operation could significantly impact estimates at the State and regional level. The problem was that a complete list of farms with accompanying measures of size did not exist -- nor has ever existed. Farms go into and out of business, combine with others, and dissolve from multiple into single entities. Therefore, an Agency priority beginning in 1968 was to make full use of both list and area frames. The list frame would estimate for the large and unusual farms and for other farms on the list using mail and telephone techniques to reduce survey costs. The area frame would be and is used for the incompleteness of the list using more expensive face-to-face interview techniques.

A difficulty caused by the use of the multiple frame estimator is that the area frame reporting units must be divided into two domains --

- o Farms that also are members of the list frame.
- o Farms that are not in the list frame.

The domain determination has been the most difficult operational aspect of developing, implementing, and using multiple frame methodology. As the structure of farms becomes more complicated with complex corporate and partnership arrangements, the survey procedures require a substantial effort to minimize nonsampling errors asociated with domain determination.

A multiple frame ratio estimator has had limited use because of the number of changes that occur in list frame units over time. Ratio estimators are used for surveys within a survey year, but are not used between years.

As the probability survey system developed and became more consistent, the use of all of the above estimators continued. While sometimes unstable at the State level, the Farm and Tract estimators were reliable at the U.S. level. The multiple frame estimator was the most reliable estimator at the State level. The weighted estimator was used for the nonoverlap domain in the multiple frame estimator.

The estimating procedure involves plotting three and sometimes four probability estimates. The current survey estimates were then reviewed relative to their performance in earlier years and their relationship with administrative data in order to "set" the official estimate.

In 1971, Houseman suggested a composite estimator that would consistently produce the least variance combination of the different estimators. Commodity statisticians have resisted the use of this estimator to have more freedom to "set" the estimate they consider the best compromise of the survey and administrative data.

V. SUMMARY OF YIELD FORECASTING METHODOLOGY

Perhaps the most market sensitive report published by the Agricultural Statistics Board is the August 1 forecast of crop production (May 1 for wheat). These reports provide the first comprehensive evaluation of the size of the current year's crop. The impact of this report is reflected in world-wide markets and closely observed by everyone from farm operators to exporters/importers to government officials around the world. To make these forecasts as accurate as possible, many techniques have been and are being tested, evaluated, and used.

Objective Yield Surveys - Objective yield surveys provide information to make forecasts and estimates of crop yield based directly on counts, measurements, and weights obtained from small plots in a random sample of fields. Sample units are located in fields identified during the June Enumerative Survey as having the crop of interest. Self weighting samples are selected. Observations within fields are made in two randomly located plots. Plots for most crops include two adjacent rows of predetermined length. Appropriate counts, measurements, and other observations are made in each sample plot.

Simple linear and multiple regression models are used to describe past relationships between the prediction variables and the final observations at maturity. Typically, early season counts and end of season harvest weights and counts unit are used. They are first screened statistically for outlier and leverage points. Once these atypical data are identified and removed, the remaining data are used to create current forecast equations.

The basic forecast models for all crops are essentially the same in that they consist of three components: the number of fruit, average fruit weight, and harvest loss.

The net yield per acre as estimated for each sample plot is computed as follows:

$\bar{y}_i = (F_i x\ C_i\ x\ W_i) - L_i$ where

F_i = Number of fruit harvested or forecast to be harvested in the i^{th} sample plot.

C_i = Conversion factor using the row space measurement to inflate the plot counts to a per acre basis.

W_i = Average weight of fruit harvested or forecast to be harvested.

L_i = Harvest loss as measured from post-harvest gleanings (the historic average is used during the forecast season).

$\hat{\bar{Y}} = (\Sigma \bar{Y}_i/n)$ for the n sample fields.

Separate models are used to forecast the number of fruit (F_i) to be harvested and the final head weight (W_i). The variables used in each model vary over the season depending upon the growth stage at the time of each survey.

At the end of the crop season, F_i and W_i are actual counts and weights of fruit for harvest. Table A shows the variables used to forecast the number of fruit to be harvested and the average fruit weight for several crops.

Table A - Forecast Components for Number of Fruit and Weight Per Fruit for Selected Crops 1/

	Number of Fruit		Average Fruit Weight	
Crop	Component	Variable Measured	Component	Variable Measured
Wheat	# Heads	# Stalks	Wt/Head	# Fertile Spikelets 2/
		# Heads in Boot		Grains/Head
		# Emerged Heads		Wt/Head
Corn	# Ears	# Stalks	Wt/Ear	Length of Husk
		# Ears with Kernels		Kernel Row 2/
				Length
				Wt/Ear
Cotton	# Bolls	# Squares	Wt/Boll	Boll Wt for Large Bolls
		# Blooms		
		# Small Bolls		
Soybeans	Pods/ Plant	# Plants	Wt/Pod	5-year Historic Average
		# Blooms		
		# Pods with Beans		
Potatoes 3/	# Hills	# Hills		Actual Wt/Hill

1/ Variables measured depend upon stage of maturity.

2/ During the growing season, counts and weights from heads or ears adjacent to the sample plot are obtained to forecast wt/head within the unit.

3/ The potato survey is only to estimate final yield -- it is not used to forecast production.

The determination of variables to use in the forecast equations is an ongoing effort. Factors affecting the choice of variables to be measured are:

a) The ability of the component to forecast final number of fruit or fruit weight.

b) The relationship between the different components being measured. For example, if two variables that forecast number of fruit are highly correlated, then consideration is given to only using one.

c) The ability to measure or observe the variable for each sample plot. This has generally precluded the use of precipitation, soil moisture, soil temperature, etc., on a sample plot basis.

d) The plot size and number of plots per field required to measure the yield component. For example, experience has shown that there is generally more variability in number of fruit than in fruit weight. Therefore, sampling considerations emphasize measuring components for number of fruit.

e) The enumerator effect. Some measurements may affect the plants' growth for the remainder of the season. Since end of season counts are paired with early season counts to develop forecast models, it is important to avoid affecting plant growth within the sample plot.

f) Destructive Sampling. The number of grains per head are used to forecast final head weight in wheat. The grains can only be counted by dissecting the head. Therefore, this prohibits the use of heads within the sample plot because they must remain until harvest so that final numbers of heads and actual head weights are available to regress against early season data to develop forecast models in future years. In the case of wheat, heads outside the unit are used in the forecast equation — this does induce another source of variability sometimes referred to as "errors in variables."

The objective yield surveys are conducted monthly during the growing season. The first survey for each crop usually begins at about the time the crop is reaching the fruiting stage. The survey results are an integral part of the crop production forecasts issued around the tenth of each month during the growing season. The surveys continue on a monthly basis until the crop is ready for harvest. The same sample plots are visited each month. At maturity, the plots are harvested and actual fruit weights are obtained. After the entire field has been harvested, additional sample plots are gleaned to measure the actual harvest loss. All data collected are not only used for the current year forecasts and estimates, but they then become part of the data base for model development for future years.

The major contributor to the forecast error is the difficulty of forecasting fruit weight early in the season. Many factors such as planting date, soil moisture, temperatures at pollination time, etc., acutely affect a plant's potential to produce fruit. While the number of fruit can be counted early in the season, the plant does not always display characteristics that provide an indication of final fruit weight. While each plant's potential to produce fruit is affected by previous circumstances, that information is locked inside the plant — often until fruit maturity. For that reason, some of the research efforts underway are to improve the early season forecasts of fruit weight.

VI. THE USE OF REMOTE SENSING

An ambitious effort was undertaken to explore the use of satellite data. Data from the Landsat satellites have been used to improve the estimates of the area planted to major crops. First, it is used in the basic construction of the area frame for the June Enumerative Survey.

Another use of Landsat data has been to use it along with the ground data collected during the June Enumerative survey to obtain improved estimates of planted acres as shown by Hanuschak (1982) and Sigman, et al., (1978).

A regression estimator utilizes both ground data from the June Enumerative Survey and classified Landsat pixels.

The variance of the regression estimator can be considerably less than that from the direct expansion estimator if there is a good correlation betwen satellite data and ground data. This procedure assumes that ground data are available to do the initial development of discriminant functions. The NASS experience indicates that the use of Landsat data without ground cover information is of limited value for estimation purposes. The use of Landsat data without corresponding ground cover data is of value for general land use stratification purposes; however, there are other limitations to the use of the Landsat data which require additional research for improvements.

Cloud Cover - Each satellite passes over a given area once every 16 days. If there are two satellites, the frequency is once every 8 days. It is possible for an entire crop season to pass and not obtain a single Landsat scene for a region without cloud cover.

Timeliness - Due to delays in receiving data because of cloud cover and the time required for processing, estimates of acres planted based on Landsat data are not received until November. By that time, their primary use is to improve the estimates of acres planted based on survey data. It is still necessary to rely upon sample survey data to measure acres harvested.

Bias in Estimators - The same data from the sample segments is used to develop the discriminant functions and to estimate the regression parameters. If the number of sample segments in a Landsat scene is small, the bias can become large. Considerable research is underway to evaluate this bias. In summary, there is tremendous potential for Landsat data to improve estimates of the area planted to each crop. A topic needing further exploration is the use of Landsat data in small or local area estimation. The estimates were produced on a limited basis in 1984. Battese and Fuller (1981) have developed a small area estimator that is being evaluated.

VI. MODERN DAY PROBLEMS

As farms became larger and more specialized, two estimation problems became more critical. These involve imputation for missing data and adjustments for outliers. A third problem involves variance estimation for the complex sample designs being used. The imputation problem has received more attention and will be discussed first.

Imputation - In the early 1970's, the "hot deck" procedure was developed and implemented into the Quarterly Agricultural Labor Survey. This survey provided quarterly estimates of numbers of farm workers by type of work, method of payment, and wages paid. The "hot deck" was basically a large matrix consisting of moving averages, number of workers and wages paid from previous reports. The matrix had separate cells for type of work and method of payment. The most obvious weakness of this method was that the sampling errors of the resulting estimates were understated because imputation was for individual farms which were further processed assuming the data had been actually reported. Also, the imputation method did not take into account the complex multiple frame design. The largest farm (if a nonrespondent) could receive the average of the most recent three reports regardless of their size or type.

The next imputation procedure used was first tested in 1978 (Crank). Imputation was not on an individual farm basis, but estimates for non-respondents were obtained by treating them as a group or domain.

The estimator for the nonresponse domain was based on two assumptions:

1. It will be possible to determine for nonrespondents whether or not they have the item of interest.
2. The distribution for respondents with the item of interest will also represent the nonrespondent.

The following paragraphs provide a short overview of how the imputation occurred.

The direct expansion for the h^{th} stratum can be written as follows:

$$y_h = \frac{N_h}{n_h}\left(n_h^{p} \cdot \bar{y}_h^{p} + n_h^{rk} \cdot \bar{y}^{p} + n_h^{ru} \cdot \bar{y}_h^{u}\right)$$

Contributions to estimate from sample units reporting the item of interest.

Contributions to estimate from refusals, etc., who are known to have the item of interest.

Contributions to estimate from refusals, etc., whose status is unknown.

One can see, after careful examination of the components, that the overall estimate is sensitive to the breakdown between refusals whose status is known and those whose status is unknown in addition to the values used to estimate for them. Another procedure that should be developed would involve an estimate standardized for a number of refusals. In other words, how would the $\hat{y}_h$ react if the number of refusals were constant from survey to survey?

The use of a new sample or a change in survey procedures can change the number of refusals and also the number identified to have the item of interest.

A refinement of the Crank estimator has been developed by Atkisson which, similar to the hot deck procedure, imputes for missing farms. It relies on the assumption underlying the Crank estimator that it will be possible to determine a minimum amount of information for the missing records, i.e., whether or not they have the item of interest.

Reported data within each sampled stratum is post stratified by crop reporting districts which are contiguous groupings of homogeneous counties. A typical State will have 7-9 districts. Means for positive reports $\bar{y}_h^p$ and usable reports $\bar{y}_h^u$ are computed as before, but by each separate district. These means then are used to impute for a missing record.

For example, a missing record known to have item of interest receives the mean for all positive reports lying in the same stratum and crop reporting district as the missing record.

Variance estimates are computed using reported and missing records alike. It is asumed this understates the variance, but at a minimal level because of the post stratified means induce variability. Additional analysis is needed to settle this issue.

A closely related problem, but also one becoming more critical as farms become larger and more diverse, is the problem of outliers or extreme observations.

Outliers - Outliers are observations that have an undue influence on the survey estimate and sampling error. In agricultural surveys, they generally occur in one of two ways:

- An extremely large operation that was incorrectly classified or missed in the sample design process and assigned to a sampled stratum.

- An ordinary operation that is assigned or falls into a stratum or Primary Sampling Unit that has an extremely small probability of selection (large expansion factor). A typical example is an urban segment that unexpectedly contains an agricultural operation.

The basic procedure to identify outliers is similar to the ESD (Extreme Studentized Deviate Rule) which is:

$$R_1 = \text{Max} \, |x_i - \bar{x}| / S \qquad \text{The } R_1$$

Value is computed from historic survey data. Any survey value exceeding the R_1 value is considered to be an outlier.

One estimate that is generated is:

$$\hat{Y}_1 = \sum_{i=1}^{t} y_i + \frac{N-t}{n-t} \sum_{t+1}^{n} y_i$$

This involves identifying the outliers and assigning them a weight of 1 assuming they were pre-selected. The remaining observations are expanded using expansion factors adjusted by the number of outliers. Another estimate is:

$$\hat{Y}_2 = r \sum_{i=1}^{t} y_i + \frac{(N-rt)}{(n-t)} \sum_{t+1}^{n} y_i$$

This is similar to Y except that a weight (r) is applied to the outlier units.

The estimator Y_1 is appropriate when the outlier is caused by an extremely large report while Y_2 is appropriate when the outlier is caused by large expansion factors. Then the (2) value can be the weight the unit should have received if it had been classified correctly.

Variance Estimation -- The sample designs used for the multiple frame surveys and objective yield surveys are based on complex, stratified, multiple stage sampling within sample frame. These designs are described elsewhere (Bosecker). These designs lead to unbiased and relatively efficient estimators. The variances of these estimators are difficult to estimate -- in some cases design unbiased estimation of the variances is impossible.

The survey design involves a combination of cluster sampling, post stratification and subsampling. The first attempts at variance estimation assumed simple random sampling with no replacement. Some early work on variance estimation was done by R. Cochrane and H. Huddleston. At the same time, Hartley also proposed a variance estimator. These estimators were appropriate for the sample designs used at that time which were more single frame oriented. Kott has shown that these understimate variances for current sample designs, and suggests new estimators.

Recent contributions by Fuller and Francisco also show that the variances being used for the objective yield estimates were understated. They suggested an improved estimator and also suggested changing the sample design to permit unbiased estimation of the variance.

VII. LOOK TO THE FUTURE

The paper so far has traced the history of the estimation methodology used. During this entire time, a "Board" process has been used to determine the official estimate. This issue has been subject to much internal debate and has not been resolved.

Since the early 1960's, significant improvements have been made in sampling and survey methodology. Despite significant developments in statistical methodology, basic Board procedures to determine official estimates have remained essentially unchanged. The Board has viewed its purpose to mainly utilize the results of various data collection activities and State Statistician recommendations as its basis to produce the bet estimate. The Board has relied upon charts to "read" current survey indications. The Board has placed much reliance on the use of administrative data and balance sheets to evaluate survey indications.

A major issue is the Board's subjective analysis of survey and check data to arrive at the official estimates as opposed to more statistical analysis based upon composite estimation procedures. The Board's position, for example is that statistical reports on production and stocks should also be in balance with administrative data. The dilemma is what to do when survey indications differ from balance sheets or administrative data. NASS has full control over data collected from its own survey program and knows its strengths and weaknesses and sampling errors. Some knowledge of nonsampling errors is also available. NASS has no control over the check data, yet is compelled to review its survey data in light of the information available from administrative sources.

A quote by Houseman concerning composite estimation in a 1970 paper still has merit.

"Probability sampling and estimation are so intertwined and related that we cannot say, logically, that we have fully embraced probability sampling until the principles of composite estimation have been embraced."

A major issue to be resolved or one that will be subjected to considerable debate will involve the role of the Board and the use of the composite estimation.

Several other estimation problems are receiving considerable attention and are discussed below.

Robust Estimators - Estimators that remain stable in the presence of outliers are needed. Agricultural operations will continue to become larger, more complex and more specialized. Structure will change faster than sample frames can be updated.

Measures of Change - Since the implementation of probability sampling, primary reliance has been on the direct expansion based on the probabilities of selection. Estimators to evaluate change from year to year need to be developed and used along with measures of level.

Crop Yield Forecasting - Historically, currently, and in the future, the most market sensitive statistics are the crop production forecasts. As satellite weather data produce better weather forecasts and more timely weather data, forecast models to improve the accuracy of the forecasts will be needed.

Small Area Estimates - The probability sample designs, survey, and estimating procedures have been developed to produce State and National estimates. What county and local area estimates are available are still based upon large scale non-probability survey data. A bridge between these two data sources is needed to produce improved county estimates.

Timeliness of Estimates - As the "information float" shortens the time span in which data are most useful and as markets continue to become even more data sensitive, there will be an increasing need to shorten the time span between data collection and dissemination of the results.

Data Analyses - The current practice is to publish official estimates --period. However, information is embedded in the survey data that would explain changes -- ups and downs -- in livestock inventories, crop acreages, etc. For example, was an increase in livestock inventories caused by new producers, or existing producers' increasing herd sizes. Each has an implication of future inventory levels. Improved procedures to "mine" the data are needed.

Conclusion - From a statistical estimation standpoint, agriculture involves many challenges. It has very diverse content and size distributions. Farms change size on a seasonal basis. Many of the commodities that are produced are perishable which presents difficulties in tracking the flow through the marketing system. Because of spoilage, grading, etc., amounts finally processed or marketed will differ considerably from the amount actually produced.

The next decade and the next century will continue to offer challenges.

BIBLIOGRAPHY

Arkin, G. F., R. L. Vanderlip, and J. T. Ritchie (1976)
"A Dynamic Grain Sorghum Growth Model," Transactions of the ASAE 19 (4): 622-630

Battese, G.E., and W. A. Fuller (1981)
"Prediction of Small Area Crop Estimation Techniques Using Survey and Satellite Data," Survey Research Section Proceedings, ASA Annual Meeting, Detroit, Michigan.

Becker, Joseph A. and C. L. Harlan (1939)
"Developments in the Crop and Livestock Reporting Service Since 1920," Journal of Farm Economics 21:799-827.

Beckman, R. J. and R. D. Cook (1983)
"Outliers" Technometrics, 25, 119-149.

Bosecker, Raymond R. (October 1977)
"Data Imputation Study on Oklahoma DES." U.S. Department of Agriculture, Statistical Reporting Service.

Bosecker, Raymond R. and Barry L. Ford (1976)
"Multiple Frame Estimation with Stratified Overlap Domain," Proceedings of the Social Statistics Section, Annual Meeting of the American Statistical Association.

Cochran, Robert and Harold Huddleston (1969)
"Unbiased Estimates of Agriculture," Statistical Reporting Service.

Cochran, William G. (1977)
"Sampling Techniques" (3rd Edition) New York, NY: John Wiley and Sons, Inc.

Cook, P. W. (1982)
"Landsat Registration Methodology Used by the U.S. Department of Agriculture's Statistical Reporting Service 1972-82," Proceedings of the ASA Annual Meeting.

Crank, Keith N. (April 1979)
"The Use of Current Partial Information to Adjust for Nonrespondents." U.S. Department of Agriculture, Statistical Reporting Service.

Fellegi, I. P., and D. Holt (1976)
"A Systematic Approach to Automatic Edit and Imputation," Journal of the American Statistical Assocation, Volume 71.

Ford, Barry L. (1976)
"Missing Data Procedures: A Comparative Study," U.S. Department of Agriculture, Statistical Reporting Service.

Ford, Barry L. (1978)
"Nonresponse to the June Enumerative Survey." U.S. Department of Agriculture, Statistical Reporting Service.

Francisco, Carol A. and Wayne A. Fuller (1986)
"Statistical Properties of Crop Production Estimators," Report on Cooperative Research with the U.S. Department of Agriculture, Statistical Reporting Service.

Fuller, Wayne A. and Leon F. Burmeister (1972)
"Estimators for Samples Selected From Two Overlapping Frames," Proceedings of the Social Science Section of the Montreal Meetings of the American Statistical Association.

Gleason, Chapman P. (1982)
"Large Area Yield Estimation/Forecasting using Plant Process Models," Presented at the Winter Meeting of the American Society of Agricultural Engineers.

Gunst, Richard F. and Robert L. Mason (1980)
"Regression Analysis and Its Application," Marcel Dekker, Inc., New York.

Hartley, H.O. (1962)
"Multiple Frames Surveys" Paper given at Minneapolis Meeetings of the American Statistical Association.

Hanuschak, George, Richard Sigman, Michael Craig, Martin Ozga, Raymond Luebbe, Paul Cook, David Kleweno, and Charles Miller (1979)
"Obtaining Timely Crop Area Estimates Using Ground-Gathered and Landsat Data," Technical Bulleton No. 1609, U.S. Department of Agriculture, Washingtohn, D.C.

Hendricks, Walter A. (February 1942)
"Theoretical Aspects of the use of the Crop Meter." Agricultural Marketing Service, USDA.

Hidiroglou, Michael A. and Kodaba P. Spirnath (1981)
"Some Estimators of a Population Total from Sample Random Samples Containing Large Units," JASA, 76, 690-695.

Houseman, Earl E. (1971)
"Composite Estimation," SRS, USDA.

King, A. J., D. E. McCarty, and Miles McPeek (1942)
"An Objective Method of Sampling Wheat Fields to Estimate Production and Quality of Wheat." U. S. Department of Agriculture, Technical Bulletin No. 814.

King, A. J. and G. D. Simpson (1940)
"New Developments in Agricultural Sampling." Journal of Farm Economics. 22:341-349.

Kott, Phillip S. and Read Johnston (1988)
"Estimating the Non-Overlap Variance Component for Multiple Frame Agricultural Surveys." National Agricultural Statistics Service, U.S. Department of Agriculture Staff Report SRB-88-05.

McClung, Gretchen (January 1988)
"A Commodity Weighted Estimator." R&AD Staff Report No. SRB-NERS-8802, U.S. Department of Agriculture, National Agricultural Statistics Service.

Nealon, Jack and Jim Cotter (1987)
"Area Frame Design for Agricultural Surveys", U.S. Department of Agriculture, National Agricultural Statistics Service.

Platek, R. and G. B. Gray (June 1985)
"Some Aspects of Nonresponse Adjustments," Survey Methodology, Volume 11, Number 1.

Roser, Bernard (1983)
"Percentage Points for a Generalized ESD Many Outlier Procedure", Technometrics, 25, 165-173.

Searls, Donald T. (1966)
"An Estimator for a Population Mean which Reduces the Effect of Large True Observations," JASA, 61-1200-1204.

Survey Reserach Methods, Annual ASA Meeting (1978)
"Implementation and Editing of Faulty or Missing Data," Selected Papers Compiled and Edited by Social Security Administration, Printed by the Bureau of the Census.

Wigton, W. H. and R. W. Van der Vaart (1973)
"A Note on Combining Estimates," Statistical Reporting Service, U.S. Department of Agriculture.