United States
Department of
Agriculture

National
Agricultural
Statistics
Service

Statistical
Research
Division

Staff Report
Number YRB-86-08

August 1986

An Assessment of Crop Production Estimators Used With The Objective Yield Surveys

Ron Fecso
Carol A. Francisco
Wayne A. Fuller

AN ASSESSMENT OF CROP PRODUCTION ESTIMATORS USED WITH THE OBJECTIVE YIELD SURVEYS. By Ron Fecso, Carol A. Francisco, and Wayne A. Fuller*, Statistical Research Division; National Agricultural Statistics Service; U.S. Department of Agriculture. August, 1986. Staff Report No. YRB-86-08.

ABSTRACT The National Agricultural Statistics Service (NASS), formerly the Statistical Reporting Service conducts, yield surveys for a variety of field crops. While plot sizes, methods of locating plots within fields, and vegetative and fruit measurement techniques differ by crop, the surveys rely on the same basic survey design. The survey design, current estimators of average yield, and current estimators of variance were reviewed under a cooperative research agreement with Iowa State University. This paper presents a summary of the research and gives recommendations based on the research.

*Ron Fecso is Head of the Yield Assessment Section in the Statistical Research Division of NASS, USDA. Carol A. Francisco is a Research Assistant and Wayne A. Fuller is a Distinguished Professor in the Department of Statistics, Iowa State University.

TABLE OF CONTENTS

SUMMARY

Currently used estimators of yield and production are adequate for use with the existing objective yield survey design. However, the currently used estimators of variance are seriously biased downward and should be replaced. Variance estimators assume simple random sampling, while the actual procedure is a complex stratified, probability proportional to size sample. Alternative variance estimators are suggested, and gains in accuracy are shown through simulation. It is also suggested that the sample be drawn in a manner which allows for a straightforward variance estimator.

Results of the study also indicated the potential for improving the quality of yield estimates through reallocation of resources between the June Enumerative Survey and the Objective Yield Survey, improved stratification of the area frame, plot level data editing procedures, and reduction of nonresponse problems.

AN ASSESSMENT OF CROP PRODUCTION ESTIMATORS USED WITH THE OBJECTIVE YIELD SURVEYS

By Ron Fecso, Carol A. Francisco and Wayne A. Fuller

INTRODUCTION

The National Agricultural Statistics Service (NASS), formerly the Statistical Reporting Service, USDA conducts surveys of corn, cotton, soybeans, and wheat yields in states which are major producers of these field crops. Recently, NASS has developed objective yield surveys for rice, grain sorghum, and sunflowers (10). The basic design for these surveys, called objective yield surveys, was developed in the 1950's. The objective yield survey data are used to forecast yield and production during the growing season and to estimate these values at harvest.

Procedures for estimating the variance of yield and production estimates computed from the complex objective yield survey design have been debated by NASS researchers for several years. Recent cooperative research has shown serious inadequacies in the currently used variance estimators (3). This paper gives a nontechnical description of the cooperative research, presents suggestions for operational action by NASS, and gives recommendations for future objective yield survey research.

OBJECTIVE YIELD SURVEY DESIGN

The basic survey design incorporates a four-step sampling procedure where the first two steps in the procedure come from the area-frame sampling used for the June Enumerative Survey (JES) (10). The third step is the selection of a sample of the fields that are enumerated in the JES. In a few states a list supplement is used because the crop occurs too rarely in the area samples. This special case is not considered in this report. The fourth step is the selection of plots within fields on which to make yield measurements. While procedures for

sample plot selection within fields differ for each crop in terms of plot sizes, methods of locating plots within fields, and measurement techniques, all surveys rely on the same basic survey design to select plots within sampled fields. Detailed information on the area frame design used for the JES is available in (2), (6) and (7).

The area frame for the conterminous 48 states is stratified by land use within each state. Land-use strata within states typically include urban, agri-urban, range, and an agricultural stratum which is further subdivided on percent of land cultivated. In some states, land-use strata are defined for specific commodities which are dominant in well defined areas. Examples of this include fruit and vegetable crops in California, wheat in Washington, and cotton in Texas.

The stratification process begins with all land within each state being divided on maps into large blocks with well defined boundaries and with each block conforming to one of the predefined land-use strata. Each of these blocks is subdivided into areas of land called frame units. The size of frame units varies depending upon factors such as available boundary designations, available ancillary information, and political boundaries. Frame units typically contain between one and 30 area segments. Once frame units are established, the number of area segments in each frame unit is determined. This is done by dividing the total area of each frame unit by the target (the desired) segment size for the given land-use stratum in which the frame unit is defined. For example, in California the target size for area segments is one-half square mile in the orchard and vegetable strata and one square mile in the all-other cropland strata.

Each land-use stratum is geographically substratified. To develop the geographic substratification, called paper

stratification, frame units within each land-use stratum are ordered by county in such a manner that adjacent counties which are agriculturally similar are placed together (1). Paper strata are formed using sequential groups of area segments from this listing. Within a given land-use stratum, paper substrata are of approximately equal area, have, within rounding, an equal number of area segments, and generally contain area segments which are agriculturally similar and geographically close together. A recently written explanation of paper strata can be found in (6). Paper stratification within land-use strata provides NASS with an important means of forming relatively homogeneous strata for the area frame. This is the sampling frame for the JES which is used to estimate acreages of most major crops, various livestock totals, and other economic information related to agriculture. For purposes of variance estimation in the JES and objective yield surveys, it is the paper strata within land use strata that are the sampling strata.

Area segments are selected using a two-step procedure. The first step in the procedure involves the selection of a frame unit within each paper stratum. Selection is done randomly with probability proportional to the number of area segments within the frame units of each paper stratum. One area segment within the selected frame unit is randomly selected at the second step. Thus, all segments within a paper stratum have an equal probability of selection. This two-step process of selecting area segments is repeated until a preassigned number of distinct segments has been selected within each paper stratum. Typically, in cropland strata, 8 to 15 segments are drawn; whereas, in agri-urban, city, and nonagricultural strata only 4 to 5 segments are drawn. The first segment selected in each paper stratum is designated as replication one, the second segment is designated as replication two, and so forth.

Approximately one-fifth of the replicates in each land-use stratum are replaced annually.

This two-step procedure for selecting segments can be considered a single step for design considerations in the estimation of totals and variances, since all segments in each paper stratum have an equal probability of selection. As is done in most studies of JES estimators, area segments are treated as the primary sampling unit, and the sample of area segments is treated as a stratified sample with simple random selection within land-use/paper strata.

The third and fourth steps in the sampling procedure involve the selection of fields and the sampling of plots within selected fields. As part of the JES, selected area segments are visited and fields which have been planted, or are scheduled to be planted, with the crop of interest are identified. The selection of fields involves several steps but can be characterized by the following process. Fields which are, or are to be, planted with the crop of interest are arrayed by segment number and order of enumeration within segment. A systematic sample of size K is selected from the array of fields with selection probabilities proportional to the product of the field acreage and the inverse of the probability of selection of the area segment in which the field is contained. Hence, the number of sampled fields in each segment varies from zero to several, and large fields within a segment can be selected more than once.

Two plots are randomly located within each selected field using a random row and pace method starting in an accessible corner of each field. Where rows are not readily distinguishable and in wheat, a random number of paces along the field edge and a random number of paces into the field are used to locate the plots. A further exception to these procedures occurs in the

wheat objective yield survey. For this survey, the first plot is randomly located; the second plot is placed in a fixed position relative to the first plot.

In the event that a large field is selected more than once during the third step of the sampling procedure, additional sets of two plots are independently sampled. Unless a field is drawn more than four times, the total number of pairs of plots observed in a field is equal to the number of times the field is selected. In the rare case where more than four sets of plots would be in one field, yield values from the first four sets are used to impute for the additional sets.

CURRENTLY USED ESTIMATORS

Estimators of the state average yield and the estimated variance of the estimated yield are currently computed as though the sample were an equal probability simple random sample of K pairs of plots. That is, the estimated average yield per acre is

$$\bar{y}_n = K^{-1} \sum_{i=1}^{K} \bar{y}_{i.} , \qquad \text{(EQ-1)}$$

and the variance of the estimated yield is estimated by

$$\hat{V}(\bar{y}_n) = [K(K-1)]^{-1} \left[\sum_{i=1}^{K} (\bar{y}_{i.} - \bar{y}_n)^2 \right] , \qquad \text{(EQ-2)}$$

where $\bar{y}_{i.}$ is the yield per acre computed from field and laboratory measurements associated with the two plots in sample i, and K is the number of pairs of plots selected. The computational form of $\bar{y}_{i.}$ varies somewhat by crop, but for all crops a single yield per acre value is computed for each pair of plots. Thus, the estimator of average yield per acre and the estimator of its variance are computed as though the sample was a simple random sample of K values of $\bar{y}_{i.}$ (the sample field estimate of yield per acre).

The estimate of total production in a state, $\hat{Y}_n$, is the product of an adjusted JES acreage estimate, $\hat{A}$, and the objective yield estimate of yield per acre, $\bar{y}_n$,

$$\hat{Y}_n = \hat{A}\,\bar{y}_n \,. \qquad \text{(EQ-3)}$$

Alternatively, $\hat{Y}_n$ can be written as $N\hat{\bar{Y}}_n$, where N is the total number of segments in the population and $\hat{\bar{Y}}_n = \hat{\bar{A}}\,\bar{y}_n$ is the average production per segment. The quantity $\hat{\bar{A}}$ is a stratified estimator of the average per segment acres of the crop. It is calculated from data collected during the June Enumerative Survey:

$$\hat{\bar{A}} = \sum_{h=1}^{L} W_h n_h^{-1} \sum_{i=1}^{n_h} A_{hi} \,,$$

where $W_h = N^{-1}N_h$ is the ratio of the number of segments in stratum h to the number of segments in the total population, L is the number of land-use/paper strata in the population, n_h is the number of segments in stratum h selected during the June Enumerative Survey, and A_{hi} is the number of acres of the crop in segment i of stratum h.

The currently used estimator of the variance of the estimated total state production was developed from a Taylor Series approximation that assumed independence of the yield $(\bar{y}_n)$ and acreage estimators. The estimator is

$$\hat{V}(\hat{Y}_n) = \hat{A}^2\hat{V}(\bar{y}_n) + \bar{y}_n^2\hat{V}(\hat{A}) + \hat{V}(\hat{A})\hat{V}(\bar{y}_n) \,.$$

It is interesting to note that this is not an unbiased estimator even if the simple random sampling approximation to the design were true. The unbiased estimator has a minus sign before the right most term, the variance product (5). Thus, this estimator is biased high with respect to the simple random sample approximation, yet we will show that it is biased low

compared to estimators which more adequately account for the sample design. The above estimator is equivalent to $N^2\hat{V}(\hat{\bar{Y}}_n)$, where

$$\hat{V}(\hat{\bar{Y}}_n) = \hat{\bar{A}}^2\hat{V}(\bar{y}_n) + \bar{y}_n^2\hat{V}(\hat{\bar{A}}) + \hat{V}(\hat{\bar{A}})\hat{V}(\bar{y}_n) \ . \qquad \text{(EQ-4)}$$

$\hat{V}(\hat{\bar{A}})$ is the usual variance estimator for a stratified mean

$$\hat{V}(\hat{\bar{A}}) = \sum_{h=1}^{L} W_h^2 n_h^{-1}(n_h - 1)^{-1} \sum_{i=1}^{n_h} (A_{hi} - \bar{A}_h)^2 \ ,$$

and

$$\bar{A}_h = n_h^{-1} \sum_{i=1}^{n_h} A_{hi} \ .$$

EVALUATION OF THE CURRENTLY USED ESTIMATORS

The estimator of average per acre crop yield (EQ-1) currently used by NASS is a type of combined ratio estimator. As can be seen in (EQ-3), $\bar{y}_n$ is the ratio of an unbiased estimator of the mean segment production and a stratified estimator of the mean number of acres per segment. As is typical with ratio estimators $\bar{y}_n$ is biased, although the bias is negligible in large samples.

Because the National Agricultural Statistics Service uses systematic subsampling of the fields selected in the JES, additional assumptions concerning the sampling scheme must be imposed to allow estimation of the variance. Replacement sampling of segments with probabilities proportional to the number of acres of the crop within each segment is assumed. This is an approximation to the probability proportional to size systematic subsampling scheme used to select objective yield sample fields.

The variance estimator currently used by the NASS (EQ-2) is an unbiased estimator of the conditional variance of $\bar{y}_n$ under the assumed replacement sampling and given the sample of segments selected for the JES (3). It, thus, is a biased estimator of the unconditional variance of $\bar{y}_n$. The unconditional variance of $\bar{y}_n$, under the assumptions of probability proportional to size sampling with replacement of segments found to have the crop during the JES, is the sum of two components: (1) the variance of the conditional expected value and (2) the expected value of the conditional variance. Estimation of the first component of the unconditional variance, the variance of the conditional expected value, is intractable, even under the simplifying assumption of probability proportional to size sampling of segments from the JES. The currently used variance estimator (EQ-2) is unbiased for estimating the second component and, therefore, seriously underestimates the variance of $\bar{y}_n$. The magnitude of the bias is a function of the effects of the use of systematic nonreplacement sampling and of the use of conditional probabilities at the second step of the sampling procedure. A Monte Carlo study of one population, which is described later, found (EQ-2) to be a 38 percent underestimate of the unconditional variance.

ALTERNATIVE ESTIMATORS

A number of formulas can be developed for estimating the conditional variance of $\bar{y}_n$. Such formulas must be approximations because of the probability proportional to size, systematic design used for yield sampling. All estimators of the conditional variance will have limited applicability in the construction of the unconditional variance of $\bar{y}_n$. This was illustrated in the previous section where the currently used estimator was shown to be equivalent to an estimator of the conditional variance of $\bar{y}_n$ under an assumption of probability

proportional to size replacement sampling of segments from the JES and, thus, seriously underestimates $V(\bar{y}_n)$.

An alternative method of developing a variance formula is to view the sampling procedure as a two-phase process. The first phase is a stratified, simple random sample of area segments. The second phase is composed of a subsample of area segments selected during phase one. The primary sampling units of the second phase are segments. For purposes of variance estimation, secondary sampling units of the second phase are pairs of plots. Two types of area segments are observed at phase two, those that have zero acres of the crop and those that have nonzero acres. The acres and the total production are known (both equal to zero) for an observed segment with zero acres. The acreage is known, but a subsample is needed to estimate production in segments with positive crop acres.

Viewing the sample as a two phase sample assumes that the unconditional probability of selecting a segment to receive a pair of plots is proportional to the conditional probability of selecting a segment for the second phase given the first phase sample of segments. Let π_{hi} be the conditional probability that segment i of paper stratum h is selected to receive a pair of plots on a draw, given the sample of segments selected during phase one of the sampling procedure (the June Enumerative Survey). We have

$$\pi_{hi} = \hat{A}^{-1} W_h n_h^{-1} A_{hi}$$

Let π^*_{hi} be the unconditional probability that an observation is made on segment i in stratum h . If $A_{hi} > 0$, then π^*_{hi} is the unconditional probability that segment hi is selected to receive at least one set of two plots. If $A_{hi} = 0$, then π^*_{hi} is equal to the probability that segment hi is selected at the first phase of sampling. Assume

$$\pi^*_{hi} = \frac{n_h}{N_h} \qquad \text{if } K\pi_{hi} = 0$$

$$= K\pi_{hi}\frac{n_h}{N_h} \qquad \text{if } 0 < K\pi_{hi} \leq 1$$

$$= \frac{n_h}{N_h} \qquad \text{if } K\pi_{hi} > 1 \ ,$$

and assume that π^*_{hi} are fixed prior to sampling.

An unequal probability combined ratio estimator of the mean yield per acre is given by

$$\bar{y}_{n,r} = \hat{A}_r^{-1} \sum_{h=1}^{L} \sum_{i=1}^{d_h} \pi^{*-1}_{hi} A_{hi} \bar{y}_{hi.} \ , \qquad \text{(EQ-5)}$$

where

$$\bar{y}_{hi} = m_{hi}^{-1} \sum_{j=1}^{m_{hi}} Y_{hij} \ ,$$

$$\hat{A}_r = \sum_{h=1}^{L} \sum_{i=1}^{d_h} \pi^{*-1}_{hi} A_{hi} \ ,$$

Y_{hij} is the yield, expanded to a per acre basis, for plot j of the i-th segment in land-use/paper stratum h, m_{hi} is the number of plots observed in segment i of stratum h, and d_h is the number of segments with positive acreage sampled in stratum h at phase 2. Note that d_h is the number of distinct segments. This estimator of the average yield per acre, $\bar{y}_{n,r}$, is approximately equivalent to the currently used estimator, $\bar{y}_n$ of (EQ-1) (3).

An estimator of the variance of $\bar{y}_{n,r}$, under an assumption of probability proportional to size replacement sampling of segments from the JES is

$$\hat{V}(\bar{y}_{n,r}) = \hat{A}_r^{-2} \sum_{h=1}^{L} g_h(g_h - 1)^{-1} \sum_{i=1}^{g_h} (\pi_{hi}^{*-1} u_{hi} - \bar{u}_{h.})^2 , \quad \text{(EQ-6)}$$

where

$$u_{hi} = A_{hi}(\bar{y}_{hi.} - \bar{y}_{n,r}) ,$$

$$\bar{u}_{h.} = g_h^{-1} \sum_{i=1}^{g_h} \pi_{hi}^{*-1} u_{hi} ,$$

and g_h is the total number of segments for which production information is available at the second phase of sampling.

An estimator of the average per segment total production is

$$\hat{\bar{Y}}_{n,r} = \hat{\bar{A}}\, \bar{y}_{n,r} .$$

This estimator is approximately equivalent to $\hat{\bar{Y}}_n$ of (EQ-3). A variance estimator which is based on the Taylor approximation to the unconditional variance of the approximate distribution of $\hat{\bar{Y}}_{n,r}$, is given by

$$\hat{V}(\hat{\bar{Y}}_{n,r}) = \hat{\bar{A}}^2 \hat{V}(\bar{y}_{n,r}) + 2\bar{y}_{n,r} \hat{C}(\bar{Y}_n, \hat{\bar{A}}) - \bar{y}_{n,r}^2 \hat{V}(\hat{\bar{A}}) , \quad \text{(EQ-7)}$$

where

$$\hat{C}(\bar{Y}_n, \hat{\bar{A}}) = \sum_{h=1}^{L} W_h^2 n_h^{-1} \hat{S}_{AYh}^2 ,$$

$$\hat{S}_{AYh}^2 = g_h(g_h - 1)^{-1} \Big(\sum_{\ell=1}^{g_h} \pi_{h\ell}^{*-1} \Big)^{-1} \sum_{i=1}^{g_h} \pi_{hi}^{*-1} (A_{hi} - \bar{A}_h^*)(A_{hi}\bar{y}_{hi.} - \bar{y}_{h..}^*) ,$$

$$\bar{A}^*_h = \left(\sum_{\ell=1}^{g_h} \pi^{*-1}_{h\ell} \right)^{-1} \sum_{i=1}^{g_h} \pi^{*-1}_{hi} A_{hi} \,,$$

$$\bar{y}^*_{h..} = \left(\sum_{\ell=1}^{g_h} \pi^{*-1}_{h\ell} \right)^{-1} \sum_{i=1}^{g_h} \pi^{*-1}_{hi} A_{hi} \bar{y}_{hi.} \,.$$

When $\hat{V}(\hat{\bar{Y}}_{n,r})$ is multiplied by N^2, where N is the total number of segments in the population, it is a stratified double sampling estimator of the variance of the estimated total state production. Unlike estimator $\hat{V}(\hat{\bar{Y}}_n)$ of (EQ-4), this estimator does not assume that the yield and acreage estimators are independent.

A MONTE CARLO COMPARISON OF ESTIMATORS

A Monte Carlo study was performed to illustrate the differences between currently used estimators and the proposed alternative estimators. Cotton acreage data from the 1983 June Enumerative Survey in the San Joaquin Valley of California and the corresponding 1983 objective yield survey data were used as a basis for the study. This section summarizes the results of the simulation study.

Table 1 shows the distribution of cotton among the 28 land-use/paper strata that were observed during the 1983 June Enumerative Survey. The six different land-use strata are defined as follows (2):

Stratum 13 - 50 percent or more cultivated land, primarily general crops with less than or equal to 10 percent fruit or vegetables;

Stratum 17 - 50 percent or more cultivated land, primarily fruit, tree nuts, or grapes mixed with general crops;

Stratum 19 - 50 percent or more cultivated land, primarily vegetables mixed with general crops;

Table 1 -- Cotton acreage estimates from the 1983 June Enumerative Survey in California and cotton acreages in the simulated population.

Land Use/ Paper Stratum	Target Segment Size (Acres)	Number Segments in Stratum	Number Segments Sampled in 1983	Percentage Segments with Cotton		Mean Acres of Cotton in Segments with Cotton	
				1983	Simulated Population	1983	Simulated Population
Stratum 13							
14	640	291	10	60	60	197	200
15	640	291	10	100	100	354	348
16	640	291	10	90	89	167	173
17	640	291	10	90	92	149	148
18	640	291	10	50	53	481	422
19	640	291	10	20	19	249[1]	260
20	640	291	10	90	91	154	155
21	640	291	10	60	61	270	274
22	640	291	10	70	71	205	210
23	640	291	10	80	79	288	279
Stratum 17							
13	320	432	10	30	28	125	122
14	320	432	10	30	31	58	57
15	320	432	10	20	22	86[2]	84
16	320	432	10	10	8	86[2]	89
17	320	432	10	40	38	26	27
18	320	432	10	30	29	144	144
19	320	432	10	30	31	65	67
20	320	432	10	30	30	38	35
21	320	432	10	30	29	133	138
22	320	432	10	50	47	130	131
23	320	432	10	40	40	76	76
Stratum 19							
06	640	362	10	70	73	117	127
07	640	362	10	70	74	192	194
08	640	362	10	80	83	253	246
Stratum 20							
10	640	649	10	30	31	303	306
11	640	649	10	40	41	175	165
Stratum 31							
07	160	1,847	5	20	22	25[3]	25
Stratum 41							
10	2,560	1,044	10	10	10	178	165

[1]Number of segments sampled was less than or equal to 2. Average of all segments in paper strata within land use stratum 13 is shown.

[2]Number of segments sampled was less than or equal to 2. Average of all segments in paper strata within land use stratum 17 is shown.

[3]Number of segments sampled was less than or equal to 2. Approximate acreages for this agri-urban stratum are shown.

Stratum 20 - 15-50 percent cultivated land with extensive cropland and hay;

Stratum 31 - residential mixed with agricultural lands, more than 20 dwellings per square mile;

Stratum 41 - less than 15 percent cultivated land, primarily privately owned rangeland.

A complete area frame population was simulated using mean and variance estimates from the JES. Details of the methods used to simulate the population are given in (3). Table 1 compares the characteristics of the simulated population to the results of the 1983 June Enumerative Survey.

Since estimated yield per acre figures were not readily accessible, an alternative variable which is a major component in the computation of yield estimates was used. This variable is the number of plants per 100 square feet. The estimated overall population mean number of plants per 100 square feet was 79.6 for the 1983 objective yield survey. Table 2 shows the average number of plants per 100 square feet by stratum for the 1983 survey and the simulated population. The average for each stratum is based on all measurements within the land-use/paper stratum which were taken from pairs of plots drawn as part of the probability proportional to estimated size sampling scheme.

Figure 1 shows the observed estimated number of plants per 100 square feet for the plots across all strata of the 1983 June Enumerative Survey. The value of plot two is plotted against the value of plot one for each pair of plots. Duplicate observations which resulted from the imputation of observations when a field within a segment was drawn more than four times are not shown on this graph.

Table 2 -- Average number of plants per 100 square feet from the 1983 objective yield survey for cotton in California and in the simulated population.

Land Use/ Paper Stratum	Average Number of Plants per 100 Square Feet	
	1983 Objective Yield Survey	Simulated Population
Stratum 13		
14	78	76
15	80	80
16	67	68
17	72	73
18	80	80
19	93	93
20	92	91
21	70	69
22	84	84
23	72	71
Stratum 17		
13	118	117
14	96[1]	95
15	96[1]	93
16	96[1]	86
17	96[1]	96
18	139	140
19	96[1]	97
20	96[1]	97
21	89	86
22	79	79
23	84	85
Stratum 19		
06	98	98
07	67	67
08	53	53
Stratum 20		
10	118	118
11	47	47
Stratum 31		
07	80[2]	79
Stratum 41		
10	60	59

[1] Number pairs of plots observed was less than or equal to 2. Plot average for land use stratum 17 is shown.

[2] Number pairs of plots observed was less than or equal to 2. Plot average for all strata is shown.

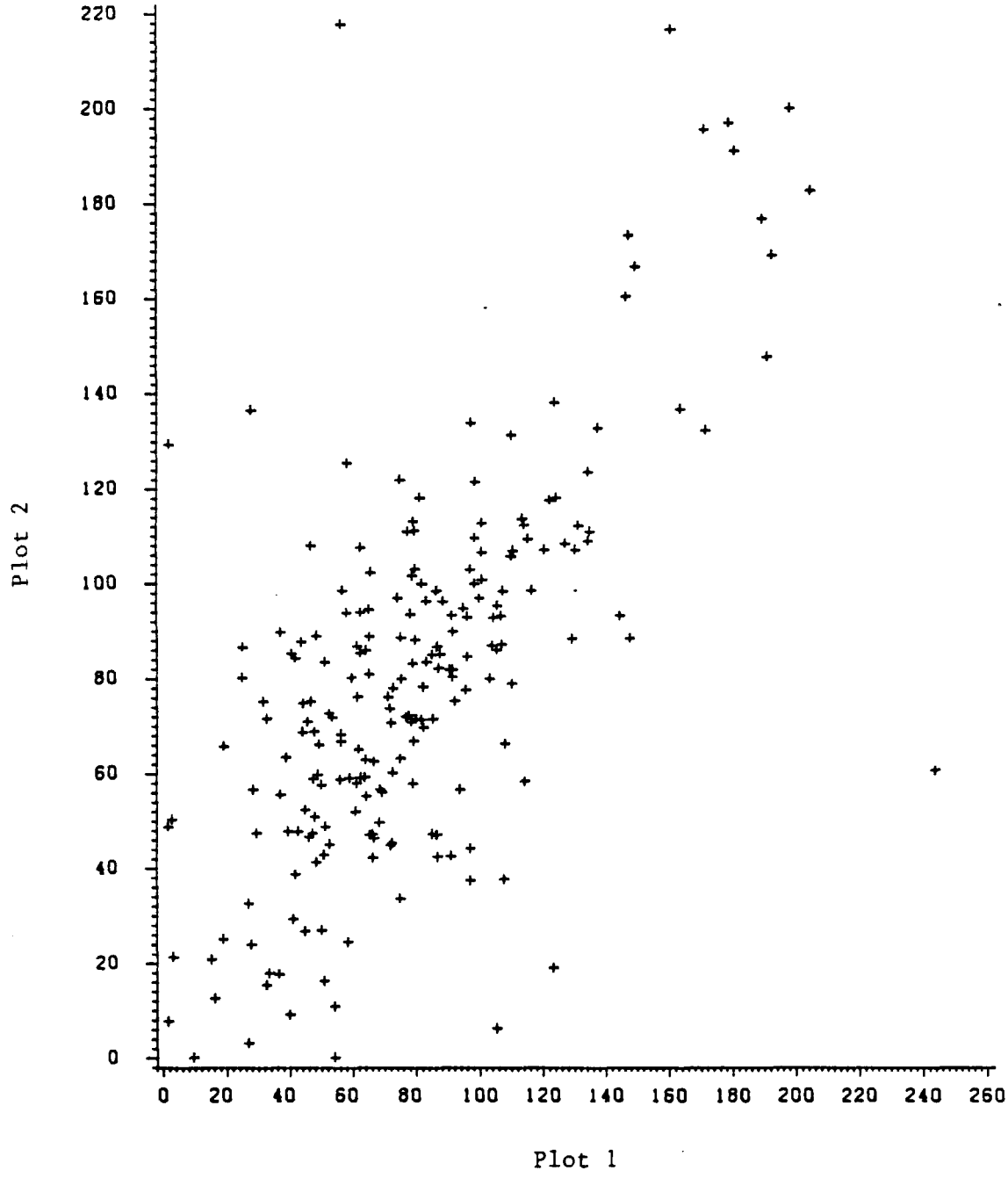

Figure 1 -- Estimated numbers of plants per 100 square feet from plots 1 and 2, 1983 June Enumerative Survey.

An analysis of variance using the 1983 objective yield data (table 3) indicated that 28 percent of the total variance among pairs of plots was due to between segment differences within strata ($s_b^2 = 378.0$) . Variation among pairs of plots within segments ($s_w^2 = 776.6$) accounted for 58 percent of the total variance. If the stratum component is treated as fixed, then 67 percent of the within segment variation is due to variance among pairs of plots.

Table 3 -- Analysis of variance for the 1983 objective yield survey data.

Source	Degrees of Freedom	Sum of Squares	Mean Square	Variance Component	Percent of total
Stratum	26	80,193	3,084.3	187.3	14
Segment within Stratum	85	124,086	1,459.8	378.0	28
Residual	103	79,991	776.6	776.6	58
Total	214	284,270		1,341.9	100

From the simulated cotton population, the June Enumerative Survey was simulated 500 times. A total of 275 segments using stratified random sampling were drawn for each simulated June Enumerative Survey. The number of segments drawn for each stratum was the same as that for the paper strata with cotton in the 1983 June Enumerative Survey (see table 1). Estimates of the mean number of acres per segment in the population were calculated for each of the 500 simulated samples. In each sample the conditional probabilities that the segments in the sample would be selected to receive objective yield plots were calculated. These conditional probabilities were used at the

second stage of sampling in the single start, probability proportional to estimated size, systematic sampling described previously. This systematic sampling scheme was used to select 220 pairs of plots which simulate an objective yield survey sample. Two objective yield survey samples were simulated for each of the 500 June Enumerative Survey samples.

When a segment was selected to receive a set of 2 plots, the yield (number of plants per 100 square feet) observed within a field was simulated. The average number of cotton acres per segment in the simulated population was 9.94, while the average of the 500 sample estimates was 9.93. The actual variance of this stratified estimator is 0.63, while the average estimated variance for the 500 simulated samples was 0.64.

In addition to the previously discussed estimators, random group estimates of the variance of $\bar{y}_n$ and $\hat{\bar{Y}}_n$ were constructed. Properties of the random group estimators are listed in Tables 4 and 5. Two sets of random groups were formed for each objective yield survey sample. One set contained five groups ($g = 5$), and one set contained ten groups ($g = 10$). Random groups were created by dividing the primary sampling units, the segments, within each land use/paper stratum into subsets. The first group in each set of groups was obtained by drawing a simple random sample without replacement of size $q_h = n_h/g$ from each stratum ($h=1,\ldots,28$) of the parent June Enumerative Survey sample. The second random group was obtained in the same fashion by selecting q_h segments from the remaining $n_h - q_h$ segments in each stratum. The remaining random groups were formed in a like manner. One land use/paper stratum, stratum number 3107, had a sample size of $n_h = 5$ segments. Acreage and yield values of the observed five segments were repeated to form the ten observations required for the ten groups.

Random group estimates of yield and production were computed by applying (EQ-1) and (EQ-3) to the pairs of plots in each random group. Variance estimates were calculated by computing the variance of the average yield and production estimates across the g random groups. Details of computations used in the calculation of the random group variance estimates can be found in (3).

Tables 4 and 5 summarize the results of the Monte Carlo study for yield and production estimators. Average values of the estimates and their variance estimates across the 1,000 simulated objective yield survey samples are shown in the tables. Estimated variances of the estimators and their variance estimators were computed and are listed in tables 4 and 5. $\hat{V}_5$ and $\hat{V}_{10}$ refer to the random group $g = 5$ and $g = 10$ estimators respectively. Simulation of two objective yield survey samples for each June Enumerative Survey sample made the estimation of between and within June Enumerative Survey variance components possible.

The currently used estimator (EQ-1), $\bar{y}_n$, and the combined ratio estimator (EQ-5), $\bar{y}_{n,r}$, provided estimates of yield with similar accuracy (see table 4). The variance estimator, $\hat{V}(\bar{y}_n)$, is known to be an underestimate of $V(\bar{y}_n)$ under an assumption of probability proportional to size replacement sampling of segments from the JES. In this population $\hat{V}(\bar{y}_n)$ underestimated the observed variance of $\bar{y}_n$ by 38 percent. The observed variance of $\bar{y}_n$ is 11.57 as compared to an average 7.21 for $\hat{V}(\bar{y}_n)$. This underestimation of the variance was consistent across samples. The estimated variance of $\hat{V}(\bar{y}_n)$ was 0.99, with $\hat{V}(\bar{y}_n)$ ranging from a low of 3.85 to a high of 11.24 in the 1,000 observations. Thus, the maximum observed estimate of variance was less than the true variance.

Table 4 -- Monte Carlo properties of yield per acre estimates and estimated variances.[1]

	Estimator					
	$\bar{y}_n$	$\hat{V}(\bar{y}_n)$	$\hat{V}_5(\bar{y}_n)$	$\hat{V}_{10}(\bar{y}_n)$	$\bar{y}_{n,r}$	$\hat{V}(\bar{y}_{n,r})$
Average	79.74	7.21	12.62	12.39	79.76	12.39
Total Variance	11.57	0.99	74.58	36.86	11.56	12.51
Between JES	7.60	0.48	6.10	4.56	7.64	7.61
Within JES	3.97	0.51	68.48	32.30	3.92	4.90

[1]Two objective yield survey samples were simulated from each of 500 simulated June Enumerative Survey (JES) samples.

Table 5 -- Monte Carlo properties of production estimates and estimated variances.[1]

	Estimator					
	$\hat{\hat{Y}}_n$	$\hat{V}(\hat{\hat{Y}}_n)$	$\hat{V}_5(\hat{\hat{Y}}_n)$	$\hat{V}_{10}(\hat{\hat{Y}}_n)$	$\hat{\hat{Y}}_{n,r}$	$\hat{V}(\hat{\hat{Y}}_{n,r})$
Average	791.89	4,800.78	5,757.80	5,704.23	792.14	5,844.63
Total Variance	5,840.81	1.14×10^6	1.72×10^7	8.41×10^6	5,826.94	3.08×10^6
Between JES	5,448.73	1.08×10^6	7.02×10^5	2.88×10^6	5,440.95	2.76×10^6
Within JES	392.08	6.02×10^4	1.65×10^7	5.53×10^6	385.95	3.23×10^5

[1] Two objective yield survey samples were simulated from each of 500 simulated June Enumerative Survey (JES) samples. The estimates, $\hat{\hat{Y}}_n$ and $\hat{\hat{Y}}_{n,r}$, when multiplied by $N = 92,240$ are estimates of the simulated total cotton production in the San Joaquin Valley.

The estimator (EQ-4), $\hat{V}(\hat{\bar{Y}}_n)$, is an underestimate of the unconditional variance of $\hat{\bar{Y}}_n$. While the observed variance of $\hat{\bar{Y}}_n$ from the Monte Carlo simulations is 5,841, the average of the $\hat{V}(\hat{\bar{Y}}_n)$ is only 4,801. This 18 percent underestimate of the true variance occurs for a number of reasons. As was shown previously, there is a negative bias in $\hat{V}(\bar{y}_n)$ as an estimator of $V(\bar{y}_n)$. Another important factor contributing to the bias is the failure of $\hat{V}(\hat{\bar{Y}}_n)$ to take into account the covariance between $\hat{\bar{A}}$ and $\bar{y}_n$. In this example the bias from omitting the covariance term partially balances the bias associated with $\hat{V}(\bar{y}_n)$.

Use of (EQ-5), $\hat{V}(\bar{y}_{n,r})$, as an estimator of the variance of $\bar{y}_{n,r}$ and of (EQ-6), $\hat{V}(\hat{\bar{Y}}_{n,r})$, as an estimator of $\hat{\bar{Y}}_{n,r}$ provided results which are much more satisfactory. The estimator $\hat{V}(\bar{y}_{n,r})$ was, on the average, a 7 percent overestimate of the observed variance of $\bar{y}_{n,r}$. About one-third of the overestimate (2-4 percent) can be attributed to the use of nonreplacement sampling at the first two stages of sampling. The relative magnitude of this overestimate, after adjusting for the known bias, was small relative to the standard error of the estimated difference. The estimated standard error of the difference was 0.58. Thus, the average value of $\hat{V}(\bar{y}_{n,r})$ is within 1.5 standard errors of the estimated variance of $\bar{y}_{n,r}$. The average estimated variance of $\bar{Y}_{n,r}$ is within 1 percent of the variance observed in the Monte Carlo simulations.

Random group estimators of the variance of $\bar{y}_n$ were more accurate than the currently used variance estimator. The Monte Carlo average of estimators $\hat{V}_5(\bar{y}_n)$ and $\hat{V}_{10}(\bar{y}_n)$ were 9 and 7 percent, respectively, larger than the corresponding Monte Carlo variances. These differences are not significantly different from zero and are comparable to those obtained for the estimator $\hat{V}(\bar{y}_{n,r})$. The variance estimator $\hat{V}(\bar{y}_{n,r})$, however, was a much more stable variance estimator. The coefficient of variation for the estimator $\hat{V}(\bar{y}_{n,r})$ was only

30 percent. It was 75 percent for $\hat{V}_5(\bar{y}_n)$. As expected (11), an increase in the number of random groups resulted in a decrease in the coefficient of variation of the random group variance estimator. The coefficient of variation for $\hat{V}_{10}(\bar{y}_n)$ was 50 percent. Differences among random groupings and yield samples within June Enumerative Surveys accounted for most of the variance in the random groups variance estimators.

The average of the random group variance estimators underestimated the variance of $\hat{\bar{Y}}_n$, but the negative bias for the estimators was less than 2.5 percent and is not significantly different from zero. Results of Monte Carlo simulations for the cotton population also show that $\hat{V}(\hat{\bar{Y}}_{n,r})$ has a lower coefficient of variation than the random group variance estimators.

CONCLUSIONS

The currently used estimator of yield per acre, $\bar{y}_n$, is satisfactory. The estimator of variance currently used by the National Agricultural Statistics Service displayed serious negative bias. The proposed estimators of variance, which are based on the observed between-segment variability, are accurate in samples of the size typically used by the NASS. The cotton population in California was simulated in a Monte Carlo study and provides a good illustration of the magnitude of the variance underestimation problem. The current variance estimator of yield (here plants per one hundred square feet) was found to underestimate the true variance by 38 percent.

RECOMMENDATIONS

1. Currently used estimators of yield and production are adequate for use with the existing survey design. The current yield and production estimators are essentially unbiased.

2. Sampling procedures for the June Enumerative Survey are straightforward and provide unbiased acreage estimates. The allocation of resources between the JES and objective

yield surveys and the potential for improved stratification are not directly addressed here, but there are indications that further work in these areas could result in a more efficient production estimate.

3. The current estimators of the variance of estimated yield and production should be replaced. Both the theoretical and the Monte Carlo results indicate that the alternative procedures developed in the cooperative agreement furnish adequate estimators for the variances of the currently used yield and production estimators if the effect of a systematic sampling is ignored.

4. Random group variance estimators are essentially unbiased estimators of the variance of estimated yield and production but are less stable than $\hat{V}(\bar{y}_{n,r})$ and $\hat{V}(\hat{\bar{Y}}_{n,r})$. Variance estimators $\hat{V}(\bar{y}_{n,r})$ and $V(\hat{\bar{Y}}_{n,r})$ are thus recommended over random group variance estimators.

5. Consideration should be given to replacing systematic sampling at phase two with a selection procedure that permits unbiased estimation of the variance (see (4) for an example of such a procedure). Because selection of segments for yield sampling at phase two is currently computerized and done at the national level, change to a selection procedure with known joint probabilities should be relatively easy to implement.

6. Edit procedures that consider individual plot values should be developed. In working with the data files from the 1983 June Enumerative Survey and the objective yield survey it became evident that further review of field data and additional computerized data editing procedures are warranted.

7. Local nonresponse problems should be identified and steps should be taken to reduce the magnitude of the problem. For example, in the 1983 objective yield survey there was a 31 percent refusal rate for Imperial Valley cotton.

REFERENCES

1. Fecso, Ron. Cluster Analysis as an Aid in Creating Paper Strata. U.S. Dept. of Agr., Stat. Rep. Serv., 1978.

2. Fecso, Ron and Van Johnson. The New California Area Frame: A Statistical Study. U.S. Dept. Agr., Stat. Rep. Serv., Publ. No. SRS-21, Sept. 1981.

3. Francisco, Carol A. and Wayne A. Fuller. "Statistical Properties of Crop Production Estimators," Report on cooperative research with the U.S. Dept. Agr., Stat. Rep. Serv., 1986.

4. Fuller, Wayne A. Sampling with Random Stratum Boundaries. Journal of the Royal Statistical Society, **B32** (1970), 209-226.

5. Goodman, Leo A. On the Exact Variance of Products. Journal of the American Statistical Association, **65**, (1970), 708-713.

6. Geuder, Jeffrey. Paper Stratification in SRS Area Sampling Frames. U.S. Dept. Agr., Stat. Rep. Serv., SF&SR Rept. No. 79, Feb. 1984.

7. Houseman, Earl E. Area Frame Sampling in Agriculture. U.S. Dept. of Agr., Stat. Rept. Serv., Publ. No. SRS-20. 1975.

8. Pratt, William L. The Use of Interpenetrating Sampling in Area Frames. U.S. Dept. Agr., Stat. Rep. Serv., May 1974.

9. U.S. Department of Agriculture. 1983 Wheat Objective Yield Survey: Enumerator's Manual. Stat. Rep. Serv., 1982.

10. U.S. Department of Agriculture. Scope and Methods of the Statistical Reporting Service. Stat. Rep. Serv., Misc. Publ. No. 1308, Sept. 1983.

11. Wolter, Kirk M. Introduction to Variance Estimation. Springer-Verlag, New York, 1985.